Soliton structure of the Drinfel'd–Sokolov–Wilson equation

R. Hirota[a]
Centre de Physique Théorique, Ecole Polytechnique, Plateau de Palaiseau, 91128 Palaiseau Cedex, France
B. Grammaticos
Département de Mathématique, C.N.E.T., 38-40 Rue du Général Leclerc, 92131 Issy les Moulineaux, France
A. Ramani
Centre de Physique Théorique, Ecole Polytechnique, Plateau de Palaiseau, 91128 Palaiseau Cedex, France

(Received 23 July 1985; accepted for publication 12 February 1986)

An integrable equation due to Drinfel'd and Sokolov [Sov. Math. Dokl. **23**, 457 (1981)] and Wilson [Phys. Lett. A **89**, 332 (1982)] (DSW) is studied in detail. It is shown how this system can be obtained as a six-reduction of the Kadomtsev–Petviashvili hierarchy. This equation presents a novel type of solutions called static solitons: they are static solutions that interact with moving solitons without deformations. Examples of such solutions are given, together with a general procedure for their construction. Finally the Painlevé analysis of the DSW equation is performed directly on the bilinear form, which constitutes a new application of the singularity analysis method.

I. INTRODUCTION

In a previous work[1] one of us (Hirota), in collaboration with Satsuma, has proposed a system of coupled KdV equations, describing the interaction of two long waves with different dispersion relations:

$$\begin{aligned} &u_t - \tfrac{1}{2}(u_{xxx} + 6uu_x) = 2b\phi\phi_x, \\ &\phi_t + \phi_{xxx} + 3u\phi_x = 0. \end{aligned} \qquad (1.1)$$

The integrability of (1.1) was subsequently proved by Ref. 2, where it was shown that this equation is a special case of the four-reduced Kadomtsev–Petviashvili (KP) hierarchy, which has been studied in detail by the Kyoto group.[3,4] Moreover, as has been shown by Wilson,[5] Eq. (1.1) can be obtained within a general construction due to Drinfel'd and Sokolov,[6] which involved affine Lie algebras. Wilson has shown in fact how this equation can be related to the affine (Kac–Moody) Lie algebra $C_2^{(1)}$. Starting from this general Drinfel'd–Sokolov construction, Wilson also has identified another interesting equation that is associated to the algebra $D_3^{(2)}$. This equation reads

$$\begin{aligned} &u_t = 3\phi\phi_x, \\ &\phi_t = 2\phi_{xxx} + 2u\phi_x + u_x\phi, \end{aligned} \qquad (1.2)$$

and will be referred to in what follows as the Drinfel'd-Sokolov-Wilson (DSW) equation. This equation possesses an infinite number of conservation laws and, in fact, a Lax representation $L_t = [P,L]$ of the form

$$\begin{aligned} L &= \left(\frac{\partial^3}{\partial x^3} + (u+\phi)\frac{\partial}{\partial x} + \frac{1}{2}(u_x + \phi_x)\right) \\ &\quad \times\left(\frac{\partial^3}{\partial x^3} + (u-\phi)\frac{\partial}{\partial x} + \frac{1}{2}(u_x - \phi_x)\right), \\ P &= -\frac{\partial^3}{\partial x^3} - u\frac{\partial}{\partial x} + \frac{1}{2}(\phi_x - u_x). \end{aligned} \qquad (1.3)$$

Moreover, in a recent work, Jimbo and Miwa[7] have shown that Eq. (1.2) is also a member of the KP hierarchy, thus confirming its integrability.

The object of the present work is to study in detail Eq. (1.2) from the point of view of the soliton structure. As a matter of fact the DSW equation presents very unusual kinds of solutions that we call static solitons. It is quite easy to show that Eq. (1.2) possesses solutions that are static: any time-independent function u (together with $\phi = 0$) is a solution. However these static solutions do not, in general, behave as solitons when they collide with the (also existing) moving solitons of the equation. Only when these static solutions are of a very particular form do they indeed lead to elastic scattering. The aim of this paper is to study these novel (to the extent of our knowledge) static-soliton features.

In Sec. II, we show how one can obtain the DSW equation using the set of equations of the KP and modified KP hierarchies of Jimbo and Miwa, implementing the appropriate reductions. By the same method, we derive another equation associated to the $D_3^{(2)}$ affine Lie algebra, which has already been studied by Ito[8] in a slightly different form and which presents also solutions of the static soliton kind. In Sec. III, we derive the precise form of the static solutions in two different ways. Finally, Sec. IV is devoted to the singularity analysis of the DSW equations. The new feature in this domain is that the Painlevé analysis is implemented directly in the bilinear form, an approach that often can lead to significant simplifications.

II. DERIVATION OF THE DSW EQUATION

According to the Kyoto group the Kadomtsev–Petviashvili (KP) hierarchy plays a fundamental role in the classification of soliton equations. In fact, this hierarchy is associated to the algebra $\mathfrak{gl}(\infty)$. By considering subalgebras of $\mathfrak{gl}(\infty)$ and their representations one can obtain various types of integrable equations. We will not enter any details concerning the general group-theoretical approach, which can be found in great detail in the paper of Jimbo and Miwa.[7]

[a] Permanent address: Hiroshima University, Faculty of Engineering, Higashi-Hiroshima 724, Japan.

We start from the KP equation

$$(u_t + 6uu_x + u_{xxx})_x + u_{yy} = 0, \tag{2.1}$$

which we can rewrite in bilinear form as

$$(D_x D_t + D_x^4 + D_y^2) f \cdot f = 0, \tag{2.2}$$

where $u = 2(\partial^2/\partial x^2)\log f$ and the bilinear operator are defined as usual through

$$D_x^m D_y^n D_t^k f(x,y,t) \cdot g(x,y,t) = \left(\frac{\partial}{\partial x} - \frac{\partial}{\partial x'}\right)^m \left(\frac{\partial}{\partial y} - \frac{\partial}{\partial y'}\right)^n \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial t'}\right)^k \times f(x,y,t) g(x',y',t') \Big|_{\substack{x = x'\cdot \\ y = y' \\ t = t'}}$$

Through a proper renaming ($x \to 1$, $y \to 2$, $t \to 3$) and rescaling of coordinates the KP equation can be written as

$$(D_1^4 - 4D_1D_3 + 3D_2^2) f \cdot f = 0. \tag{2.3}$$

In order to introduce the τ function of Sato and Sato,[3] we start from the two-soliton solution of KP:

$$f = 1 + \exp(\eta_1) + \exp(\eta_2) + a_{12}\exp(\eta_1 + \eta_2), \tag{2.4}$$

with

$$\eta_i = l_i x_1 + m_i x_2 + n_i x_3 + \eta_i^0$$

and

$$a_{12} = -\frac{(l_1 - l_2)^4 - 4(l_1 - l_2)(n_1 - n_2) + 3(m_1 - m_2)^2}{(l_1 + l_2)^4 - 4(l_1 + l_2)(n_1 + n_2) + 3(m_1 + m_2)^2}, \tag{2.5}$$

where the η_i^0 are constants and l_i, m_i, and n_i are related by the dispersion relation

$$l_i^4 - 4l_i n_i + 3m_i^2 = 0. \tag{2.6}$$

We write first the wave numbers l and m and frequency n in terms of new variables p_i and q_i:

$$l_i = p_i - q_i, \quad m_i = p_i^2 - q_i^2, \quad n_i = p_i^3 - q_i^3. \tag{2.7}$$

The dispersion relation is identically satisfied and the phase shift a_{12} reduces to

$$a_{12} = (p_1 - p_2)(q_1 - q_2)/(p_1 - q_2)(q_1 - p_2). \tag{2.8}$$

Next, following the Kyoto group, we introduce our infinite number of coordinates (infinite number of "times") $x_4, x_5, \ldots$ and write η_i as

$$\eta_i = \delta_i + \sum_{n=1}^{\infty} (p_i^n - q_i^n) x_n, \tag{2.9}$$

δ_i being a constant.

From the two-soliton solution f we can construct the two-soliton solution τ involving an infinite number of coordinates, through (2.9), as

$$\tau = 1 + \exp(\eta_1) + \exp(\eta_2) + \frac{(p_1 - p_2)(q_1 - q_2)}{(p_1 - q_2)(q_1 - p_2)} \exp(\eta_1 + \eta_2). \tag{2.10}$$

In a similar way one can introduce the N-soliton τ function. As has been shown,[7] the τ function satisfies simultaneously the whole hierarchy of the KP equation, which in bilinear form can be written as

$$(D_1^4 - 4D_1D_3 + 3D_2^2)\tau \cdot \tau = 0, \tag{2.11}$$

$$((D_1^3 + 2D_3)D_2 - 3D_1D_4)\tau \cdot \tau = 0, \tag{2.12}$$

$$(D_1^6 - 20D_1^3D_3 - 80D_3^2 + 144D_1D_5 - 45D_1^2D_2^2)D_3\tau \cdot \tau = 0, \tag{2.13}$$

$$(D_1^6 + 4D_1D_3 - 32D_3^2 - 9D_1^2D_2^2 + 36D_2D_4)\tau \cdot \tau = 0, \tag{2.14}$$

etc.

The appropriate choice of the constants δ_i allows one to introduce the functions τ_n defined as

$$\tau_n = 1 + (p_1/q_1)^n \exp(\eta_1), \tag{2.15}$$

for the one-soliton solution, and

$$\tau_n = 1 + \left(\frac{p_1}{q_1}\right)^n \exp(\eta_1) + \left(\frac{p_2}{q_2}\right)^n \exp(\eta_2) + \left(\frac{p_1 p_2}{q_1 q_2}\right)^n a_{12} \exp(\eta_1 + \eta_2), \tag{2.16}$$

in the case of two solitons, and analogously for N solitons. It is clear that all the τ_n satisfy the hierarchy defined by (2.11), (2.12), etc. Moreover one can introduce "modified" hierarchies, in which case τ_n and τ_{n+k} will be related (through the k-modified hierarchy). A particularly simple example is given by the modified KP itself, which can be obtained through a Bäcklund transformation on the solution of the KP, leading to the system

$$(D_1^2 + D_2)\tau_0 \cdot \tau_1 = 0, \tag{2.17}$$

$$(D_1^3 - 4D_3 - 3D_1D_2)\tau_0 \cdot \tau_1 = 0. \tag{2.18}$$

Once the hierarchies are established one can obtain soliton equations by appropriate reductions. If one wishes to limit oneself to the case of the $D_3^{(2)}$ subalgebra, then the following restrictions of τ_n must be taken into account:

$$\tau_{n+6\nu}(x) = \tau_{1-n}(\tilde{x}) = \tau_n(x), \tag{2.19}$$

$$\frac{\partial \tau_n(x)}{\partial x_{6\nu}} = 0, \tag{2.20}$$

for $n = 0,1,2,\ldots$ and $\nu = 1,2,\ldots$, where x stands for $(x_1, x_2, \ldots)$ and $\tilde{x}$ is defined as $\tilde{x} = (x_1, -x_2, x_3, -x_4, \ldots)$. The second equation, (2.20), indicates that all equations related to $D_3^{(2)}$ are six-reductions of the KP hierarchies. The first relation leads to the identities

$$\begin{aligned} \tau_0(x) &= \tau_6(x) = \tau_1(\tilde{x}), \\ \tau_{-1}(x) &= \tau_5(x) = \tau_2(\tilde{x}), \\ \tau_{-2}(x) &= \tau_4(x) = \tau_3(\tilde{x}). \end{aligned} \tag{2.21}$$

In order to derive the equations we are interested in we will use the first modified KP hierarchy, acting on τ_0, τ_1. From (2.19) we have

$$\tau_1(x) = \tau_0(\tilde{x}). \tag{2.22}$$

We introduce f and g defined by

$$f = \tau_0|_{x_2 = x_4 = \cdots = 0}, \quad g = \frac{\partial \tau_0}{\partial x_2}\Big|_{x_2 = x_4 = \cdots = 0}. \tag{2.23}$$

From (2.23) we can write

$$\begin{aligned} \tau_0(x) &= f + x_2 g + \mathcal{O}(x_2^2, x_4, \ldots), \\ \tau_1(x) &= \tau_0(\tilde{x}) = f - x_2 g + \mathcal{O}(x_2^2, x_4, \ldots), \end{aligned} \tag{2.24}$$

with f and g depending only on x_{2n+1} and where the higher-order terms will give contributions that disappear at the limit $x_2 = x_4 = \cdots = 0$. We use the following two equations of the first modified KP hierarchy:

$$(D_1^2 + D_2)\tau_0 \cdot \tau_1 = 0, \tag{2.25}$$

$$(D_1^3 D_3 + 2D_3^2 - 3D_1 D_2 D_3 + 6D_6)\tau_0 \cdot \tau_1 = 0, \tag{2.26}$$

and substitute τ_0, τ_1, performing the x_2 derivations explicitly. Moreover the action of D_6 is zero due to (2.20). This leads directly to the equations

$$D_1^2 f \cdot f + 2fg = 0, \tag{2.27}$$

$$(D_1^3 D_3 + 2D_3^2) f \cdot f - 6D_1 D_3 f \cdot g = 0. \tag{2.28}$$

Putting $u = (\partial/\partial x_1)\ln f$ one can write the system (2.27) and (2.28) in the form

$$\frac{\partial^3 u}{\partial x_3^2} + 2\frac{\partial}{\partial x_1}\left(\frac{\partial^3 u}{\partial x_1^2 \partial x_3} + 3\frac{\partial u}{\partial x_1}\frac{\partial u}{\partial x_3}\right) = 0. \tag{2.29}$$

A more familiar form of this equation is the one studied by Ito[8]:

$$(D_3^2 + 2D_1^3 D_3) F \cdot F = 0. \tag{2.30}$$

Putting $u = 2(\partial/\partial x_1)\ln F$, i.e., $f = F^2$, we get the same equation (2.29). In order to obtain the DSW equation we will use the following members of the third modified KP hierarchy acting on τ_1, τ_2, and τ_3 [which are related also by Eqs. (2.21)]:

$$(D_1^4 + 8D_1 D_3 + 6D_1^2 D_2 + 3D_2^2 - 6D_4)\tau_n \cdot \tau_{n+3} = 0, \tag{2.31}$$

$$\begin{aligned}(D_1^6 &- 40D_1^3 D_3 - 96D_1 D_5 + 15D_1^2 D_2^2 - 90D_1^2 D_4 \\ &+ 30D_2^3 + 60D_2 D_4)\tau_n \cdot \tau_{n+3} = 0,\end{aligned} \tag{2.32}$$

$$\begin{aligned}(D_1^6 &- 8D_1^3 D_3 + 16D_3^2 - 9D_1^2 D_2^2 - 18D_1^2 D_4 + 6D_2^3 \\ &+ 36D_2 D_4 + 48D_6)\tau_n \cdot \tau_{n+3} = 0.\end{aligned} \tag{2.33}$$

These equations can be further simplified. First the term involving D_6 can be omitted, as we are performing a six-reduction. Next we use the following symmetry argument: writing any of the operators on the third modified KP as $P(D)$, we have $P(D)(\tau_{-1} \cdot \tau_2) = 0$, $P(D)(\tau_2 \cdot \tau_5) = 0$, and as $\tau_{-1}(x) = \tau_5(x)$ [from (2.21)], we also have $P(D)(\tau_5 \cdot \tau_2) = 0$. Using the symmetry properties of the bilinear D operator we have $P(D)(\tau_5 \cdot \tau_2) = P(-D)(\tau_2 \cdot \tau_5) = 0$. Thus $[P(D) + P(-D)] \times (\tau_2 \cdot \tau_5) = 0$, which means that only the even part of Eqs. (2.31)–(2.33) need be considered. Next one can introduce f and g through $f = \tau_2|_{x_2 = x_4 = \cdots = 0}$, $g = (\partial\tau_2/\partial x_2)|_{x_2 = x_4 = \cdots = 0}$,

which leads to the following expression for τ_2:

$$\begin{aligned}\tau_2(x) &= f + x_2 g + \tfrac{1}{2}x_2^2 h, \\ \tau_5(x) &= \tau_2(\tilde{x}) = f - x_2 g + \tfrac{1}{2}x_2^2 h,\end{aligned} \tag{2.34}$$

where the terms not written explicitly have vanishing contributions at the limit $x = \tilde{x}$. From the KP equations (2.11), we obtain

$$(D_1^4 - 4D_1 D_3) f \cdot f + 6(fh - g^2) = 0, \tag{2.35}$$

while (2.31) gives us

$$(D_1^4 + 8D_1 D_3) f \cdot f + 6(fh + g^2) = 0. \tag{2.36}$$

Subtracting (2.35) from (2.36) we obtain

$$D_1 D_3 f \cdot f + g^2 = 0, \tag{2.37}$$

which is the first equation of the DSW system: in order to obtain the second equation we start by eliminating $D_2 D_4$ between (2.32) and (2.33) after odd terms have been deleted. This gives us

$$\begin{aligned}(D_1^6 &+ 40D_1^3 D_3 + 144D_1 D_5 \\ &+ 40D_3^2 - 45D_1^2 D_2^2)\tau_2 \cdot \tau_5 = 0\end{aligned} \tag{2.38}$$

or, equivalently,

$$\begin{aligned}(D_1^6 &+ 40D_1^3 D_3^2 + 144D_1 D_5 + 40D_3^2) f \cdot f \\ &- 90D_1^2 (f \cdot h + g \cdot g) = 0.\end{aligned} \tag{2.39}$$

Similarly (2.13) applied on $\tau_2 \cdot \tau_2$ gives

$$\begin{aligned}(D_1^6 &- 20D_1^3 D_3 - 80D_3^2 + 144D_1 D_5) f \cdot f \\ &- 90D_1^2 g \cdot g = 0.\end{aligned} \tag{2.40}$$

Subtracting (2.40) from (2.39) leads to

$$(D_1^3 D_3 + 2D_3^2) f \cdot f - 3D_1^2 g \cdot g = 0, \tag{2.41}$$

which is the second equation of the DSW system. In fact putting $v = (\partial/\partial x_1)\log f$ and $\phi = g/f$ in (2.37) and (2.41) results in

$$2\frac{\partial v}{\partial x_3} + \phi^2 = 0, \tag{2.42}$$

$$\phi\frac{\partial \phi}{\partial x_3} + \frac{\partial}{\partial x_1}\left(2\phi\frac{\partial^2 \phi}{\partial x_1^2} - \left(\frac{\partial \phi}{\partial x_1}\right)^2 + 3\phi^2\frac{\partial v}{\partial x_1}\right) = 0, \tag{2.43}$$

which coincide with (1.2) provided $u = 3\,\partial v/\partial x_1$.

III. STATIC AND MOVING SOLITON SOLUTION

By inspecting Eqs. (2.29), (2.42), and (2.43), one can check that any static object $u(x,t) = u(x)$ and $v(x,t) = v(x), \phi(x,t) = 0$, respectively, is a solution. Still among these solutions there must exist some special objects that play the role of soliton solutions. They are, however, harder to define than normal moving solitons. Indeed usually a soliton is a traveling wave with vanishing boundary conditions and giving the velocity determines entirely the soliton. However, the characteristic property of the soliton, i.e., what distinguishes a soliton of an integrable equation from a solitary traveling wave of a nonintegrable equation is that solitons interact by phase-shifting only. For Eqs. (2.29), (2.42), and (2.43), where the zero-velocity case is highly degenerate, boundary conditions are insufficient. However one can still determine a "static" soliton by identifying it to a static solution, which, upon interacting with any number of moving solitons (previously determined in the usual way), is shifted with no further deformation and induces only phase shifts on the moving solitons. Let us illustrate that on Eq. (2.29), which we write in bilinear form as

$$(D_t^2 + 2D_x^3 D_t) F \cdot F = 0.$$

We first look for a moving soliton as $F = 1 + e^{\eta}$ with $\eta = kx + \Omega t$ and readily find the dispersion relation $\Omega(\Omega + 2k^3) = 0$. So a regular moving soliton has

$\Omega = -2k^3$. The simplest choice of a static soliton, centered at the origin, would be

$$F = 1 + e^{\xi} \quad \text{with } \xi = px + q. \tag{3.1}$$

This, of course, is a solution, but so would be any static object. We should rather check whether (3.1) satisfies our definition upon interaction: this will be the case if a two-soliton solution can be found of the form

$$F = 1 + e^{\eta} + e^{\xi} + \alpha e^{\eta + \xi}. \tag{3.2}$$

Indeed "before" ($t \ll 0$) the moving soliton comes near the origin, i.e., $e^{\eta} \gg 1$ near $x \sim 0$, $u = (\partial^2/\partial x^2)\ln F \approx (\partial^2/\partial x^2)(1 + \alpha e^{\xi})$. "After" ($t \gg 0$) its departure ($e^{\eta} \ll 1$ near $x \sim 0$), $u \approx (\partial^2/\partial x^2)(1 + e^{\xi})$. This corresponds to a shift of the peak of the soliton $\Delta x = \ln \alpha / p$.

Expression (3.2) is indeed a solution provided that

$$\alpha = -\frac{\Omega + 2(k-p)^3}{\Omega + 2(k+p)^3} = \frac{3k^2 - 3kp + p^2}{3k^2 + 3kp + p^2}, \tag{3.3}$$

so (3.1) is indeed a good candidate for a one-static-soliton solution. One should further check that not only a single moving soliton but any number of them can cross it without any deformation. Analogously one can define a two-static (and in fact N-static) soliton solution by its interaction with moving solitons. More precisely, an

$$F = 1 + e^{\xi_1} + e^{\xi_2} + \beta e^{\xi_1 + \xi_2} \quad (\xi_i = p_i x + q_i) \tag{3.4}$$

will be called a two-static soliton solution, if, upon interacting with moving solitons, it recovers its form up to shifting the ξ_i's. Suppose that a one-moving plus two-static solution exists and has the form

$$F = 1 + e^{\xi_1} + e^{\xi_2} + \beta e^{\xi_1 + \xi_2} + e^{\eta}(1 + \alpha_1 e^{\xi_1} + \alpha_2 e^{\xi_2} + \gamma e^{\xi_1 + \xi_2}). \tag{3.5}$$

In order for the moving soliton to cross the two-static solution (3.4) with only shifts we must ensure that the form (3.4) be recovered from the term in parentheses by absorbing α_1, α_2 in the definition of the ξ_i's. This gives $\gamma = \alpha_1 \alpha_2 \beta$. By taking the limit $\xi_2 \to -\infty$ (resp. $\xi_1 \to -\infty$) one can convince oneself that for (3.5) to be a solution, α_1 (resp. α_2) must be of the form (3.3) with $p = p_1$ (resp. p_2). We find that (3.5) is indeed a solution provided that

$$\beta = \frac{p_1^4 - 3p_1^3 p_2 + 4p_1^2 p_2^2 - 3p_1 p_2^3 + p_2^4}{p_1^4 + 3p_1^3 p_2 + 4p_1^2 p_2^2 + 3p_1 p_2^3 + p_2^4}. \tag{3.6}$$

Thus (3.4) is a two-static soliton solution for this choice of β.

Let us now turn to the study of the static solitons of the DSW equation. The first step is to find the moving soliton. Following the usual approach for equations with this bilinear structure we look for

$$f = 1 + e^{2\eta}, \quad g = \lambda e^{\eta} \quad \text{with } \eta = kx + \Omega t. \tag{3.7}$$

We get thus $2(2k)(2\Omega) + \lambda^2 = 0$ and $(2\Omega)[(2k)^3 + 2(2\Omega)] = 0$. The second equation leads to the dispersion relation $\Omega(2k^3 + \Omega) = 0$ and we choose the moving ($\Omega \neq 0$) solution $\Omega = -2k^3$. Hence $\lambda^2 = 16k^4$. The simplest choice for a static soliton is

$$f = 1 + e^{2\xi} \quad \text{with } \xi = px + q, \quad g = 0. \tag{3.8}$$

This would have been a soliton if there existed a solution of the form

$$\begin{aligned} f &= 1 + e^{2\eta} + e^{2\xi} + ae^{\eta + \xi} + be^{2\eta + 2\xi}, \\ g &= \lambda e^{\eta}(1 + ce^{2\xi}). \end{aligned} \tag{3.9}$$

By asymptotic considerations, similar to those that followed Eq. (3.2), we should have $b = c^2$. However, one can check that (3.9) is not a solution for any choice of a and c as functions of p and k. Therefore we are led to try a more general form for a static solution

$$f = 1 + Ae^{\xi} + e^{2\xi}, \quad g = 0, \tag{3.10}$$

where A must be independent of p. This will be a soliton solution if there exists a solution of the form

$$\begin{aligned} f &= 1 + Ae^{\xi} + e^{2\xi} + e^{2\eta}(1 + \alpha A e^{\xi} + \beta e^{2\xi}), \\ g &= \lambda e^{\eta}(1 + \gamma A e^{\xi} + \delta e^{2\xi}), \end{aligned} \tag{3.11}$$

where $\alpha,\beta,\gamma,\delta$ are functions of p and k. Asymptotic considerations show that $\beta = \alpha^2$, $\delta = \alpha$. We find that (3.11) is indeed a solution provided

$$A = 4,$$

$$\alpha = (3k^2 - 3kp + p^2)/(3k^2 + 3kp + p^2),$$

and (3.12)

$$\gamma = (6k^2 - p^2)/2(3k^2 + 3kp + p^2).$$

Thus the one-static soliton solution for DSW is

$$f = 1 + 4e^{px} + e^{2px}, \quad g = 0. \tag{3.13}$$

In order to find a two-static soliton solution one must consider as in the case of Eq. (2.30) the interaction of two static with one moving soliton. This is an extremely tedious task. However a different approach allows us to obtain a general form for an N-static plus M-moving soliton solution. Indeed, going back to the KP hierarchy, we know[7] that the general τ-function solution for an arbitrary number N of solitons has the form

$$\tau_0 = \sum_{X \subseteq [0,N]} \left(\prod_{\substack{i \neq j \\ i,j \in X}} a_{ij} \right) \exp\left(\sum_{i \in X} \eta_i \right), \tag{3.14}$$

where the first sum is taken over all subsets X of $[0,N]$ (including the empty set) with η_i given by (2.9) and a_{ij} by (2.8). In order for τ_0 to be a solution of DSW, it must, in addition, satisfy the reduction conditions (2.19) and (2.20). For all i we must have $p_i^6 = q_i^6$. This will ensure both (2.20) and $\tau_{n+6\nu}(x) = \tau_n(x)$. We must still implement $\tau_n(x) = \tau_{1-n}(\tilde{x})$. For this we require N to be even. [Indeed, if $N = 1$, we cannot implement this condition. Write $\tau_0 = 1 + e^{\eta_1(x)}$. Then $\tau_1 = 1 + (p_1/q_1)e^{\eta_1(x)}$ and thus $(p_1/q_1)e^{\eta_1(\tilde{x})} = e^{\eta_1(x)}$ or $p_1/q_1 = e^{2\eta'}$ with

$$\eta' = \sum_n (p_1^{2n} - q_1^{2n}) x_{2n}.$$

This implies $\eta' = 0$ and thus $p_1 = q_1$. But then $\eta = 0$ and τ is a mere constant.]

Let us now look at the $N = 2$ case. Indeed, if we choose $p_2 = -q_1$, $q_2 = -p_1$, then $p_2^{2n} - q_2^{2n} = -(p_1^{2n} - q_1^{2n})$ while $p_2^{2n+1} - q_2^{2n+1} = +(p_1^{2n+1} - q_1^{2n+1})$ and thus $\eta_2(x) - \delta_2 = \eta_1(x) - \delta_1$. Therefore starting from

$$\tau_0 = 1 + e^{\eta_1(x)} + e^{\eta_2(x)} + a_{12}e^{\eta_1(x) + \eta_2(x)},$$

we find

$$\tau_0(x) = 1 + e^{\eta_1(x)} + \lambda e^{\eta_1(\tilde{x})} + a_{12}\lambda e^{\eta_1(\tilde{x}) + \eta_1(x)}$$
$$(\text{with } \lambda = e^{\delta_2 - \delta_1})$$

and

$$\tau_1(x) = 1 + \left(\frac{p_1}{q_1}\right) e^{\eta_1(x)} + \lambda \left(\frac{p_2}{q_2}\right) e^{\eta_1(\tilde{x})} + a_{12}\left(\frac{p_1}{q_1}\right)\left(\frac{p_2}{q_2}\right) e^{\eta_1(x) + \eta_1(\tilde{x})}. \tag{3.15}$$

But $p_2/q_2 = q_1/p_1$. So going from x to $\tilde{x}$, we get

$$\tau_1(\tilde{x}) = 1 + \lambda \left(\frac{q_1}{p_1}\right) e^{\eta_1(x)} + \left(\frac{p_1}{q_1}\right) e^{\eta_1(\tilde{x})} + a_{12}\lambda e^{\eta_1(x) + \eta_1(\tilde{x})}. \tag{3.16}$$

This is identical to $\tau_0(x)$ for $\lambda = p_1/q_1$. In a similar way, in order to have $\tau_{-1}(x) = \tau_2(\tilde{x})$ we can use the same form as (3.15) for τ_{-1} but with $\lambda = (p_1/q_1)^3$. (This form is immediately obtained from

$$\tau_{-1} = 1 + \left(\frac{q_1}{p_1}\right) e^{\eta_1} + \left(\frac{q_2}{p_2}\right) e^{\eta_2} + a_{12}\left(\frac{q_1}{p_1}\right)\left(\frac{q_2}{p_2}\right) e^{\eta_1 + \eta_2}$$

by redefining δ_1 and δ_2.) The expressions of the form (3.15) for τ_0 and τ_{-1} coming from the two-soliton solution of the hierarchy depend now on only one couple (p_1,q_1) with $p_1^6 = q_1^6$. They are in fact one-soliton solutions for Eq. (2.30) and DSW, respectively.

Indeed let us compute τ_i $(i = 0, -1)$ and its derivatives on the line $x_2 = x_4 = \cdots = 0$ as a function of $x' = (x_1,x_3,\ldots,x_{2n+1},\ldots)$. We find

$$f = \tau_i|_{x_2 = x_4 = \cdots = 0} = 1 + (1 + \lambda)e^{\eta(x')} + a_{12}e^{2\eta(x')},$$
$$g = \left.\frac{\partial\tau_i}{\partial x_2}\right|_{x_2 = x_4 = 0} = (p^2 - q^2)(1 - \lambda)e^{\eta(x')}, \tag{3.17}$$

with

$$\eta(x') = \sum_n (p^{2n+1} - q^{2n+1})x_{2n+1}$$

and

$$a_{12} = \frac{(p_1 - p_2)(q_1 - q_2)}{(p_1 - q_2)(q_1 - p_2)} = \frac{(p+q)^2}{4pq},$$

and where $\lambda = p/q$ for τ_0 and $\lambda = (p/q)^3$ for τ_{-1}. Note that for τ_0, since $1 + \lambda = (p+q)/q$, $\lambda a_{12} = (p+q)^2/4q^2$, f is indeed the perfect square of $(1 + [(p+q)/2q]e^{\eta(x')})$ as expected.

For τ_{-1}, we start from $p^6 = q^6$. There are two subcases: (i) $p^3 = q^3$, i.e., $\lambda = 1$; and (ii) $p^3 = -q^3$, i.e., $\lambda = -1$. In subcase (i) we have $g = 0$ and indeed $\Omega = p^3 - q^3 = 0$. This contains the static soliton branch. However, the case $p = q$ leads to $\eta = 0$, i.e, $f = 1$ $(u = 0)$, and is not interesting. The interesting case is $p^2 + pq + q^2 = 0$. There $F = 1 + 2e^{\eta(x')} + a_{12}e^{2\eta(x')}$ with $a_{12} = (p^2 + 2pq + q^2)/4pq = \frac{1}{4}$. Redefining η by $\eta' = \eta - \ln 2$, we recover the form (3.16) with $A = 4$. In subcase (ii) we must also eliminate $p + q = 0$ as it leads to $g = 0$ and $f = 1$ $(\lambda = -1$ and $a_{12} = 0)$. Restricting to $p^2 - qp + q^2 = 0$ we have $f = 1 - a_{12}e^{2\eta(x')}$, $g = 2(p^2 - q^2)e^{\eta(x')}$, with

$$a_{12} = (p^2 + 2pq + q^2)/4pq = \tfrac{3}{4}.$$

Redefining $2\eta' = 2\eta - \ln(\frac{3}{4}) - i\pi$, we find $f = 1 + e^{2\eta'}$, $g = \mu e^{\eta'}$, with $\mu^2 = -\frac{16}{3}(p^2 - q^2)^2$. On the other hand, $k = p - q$, $\Omega = p^3 - q^3$, and we can easily check that $\Omega + 2k^3 = 0$ and $\mu^2 = 16k^4$ as expected. We thus recover the moving-soliton solution of the DSW equation.

Following the same procedure, starting from a four-soliton τ function with

$$p_2 = -q_1, \quad q_2 = -p_1, \quad p_1^6 = q_1^6,$$
$$e^{\eta_2(x)} = (p_1/q_1)^\alpha e^{\eta_1(\tilde{x})},$$
$$p_4 = -q_3, \quad q_4 = -p_3, \quad p_3^6 = q_3^6,$$
$$e^{\eta_4(x)} = (p_3/q_3)^\alpha e^{\eta_3(\tilde{x})},$$

with $\alpha = 1$ for τ_0 and $\alpha = 3$ for τ_{-1}, we can recover the two-soliton solution, with a soliton being static if the corresponding p and q satisfy $p^3 = q^3$ and moving if $p^3 = -q^3$.

One can check that this general form coincides with the two-soliton solutions for either (2.27) and (2.28) or DSW that were solutions observed directly. In a similar way, starting from a $2(N_1 + N_2)$-soliton τ function we can construct a N_1-static plus N_2-moving soliton solution for both equations.

At this point a remark is in order. This construction tells us that an N_1-static plus N_2-moving soliton solution exists for Eqs. (2.27) and (2.28) as a polynomial in exponentials for f and g. For small $N = N_1 + N_2$ we have checked that this polynomial is indeed a perfect square and that F as a solution of (2.30) is a polynomial. In fact, this is a consequence of Eq. (6.7) of Ref. 7. Although Ito's equation (II.30) does not appear, as it stands, in the BKP hierarchy of Ref. 7, it can be found in Ref. 9 and corresponds to a $(\text{BKF})_6$ reduction.

IV. PAINLEVÉ ANALYSIS OF THE DSW EQUATION

In this section, we will study the behavior of the solutions of the DSW equation in the neighborhood of a singularity ("Painlevé analysis") but we will do it in an original way: we will work directly on the bilinear form. We start from

$$D_tD_x f\cdot f + g^2 = 0, \quad D_x^3D_t f\cdot f - pD_x^2 g\cdot g = 0. \tag{4.1}$$

The DSW equation corresponds to $p = 3$ and $\phi = g/f$, $u = (d/dx)\ln f$. In fact we do not look for a *singular* behavior of f and g, which we expect to be entire functions, but rather for zeros of f of arbitrary multiplicity n (which induce a simple pole of residue n for v) at places where g is either nonzero or has a zero of lower multiplicity m leading to poles of multiplicity n (or $n - m$) for ϕ. We will insist that f and g are indeed *entire* functions and that the free integration constants that enter in f and g do not enter at noninteger powers or induce logarithms, just as one does in the usual Painlevé analysis near poles.

Let us thus assume that on a manifold $x - \varphi(t) = 0$, f has a zero of nonzero multiplicity n and g a zero of multiplicity m (possibly zero):

$$\begin{aligned} f &= (x - \varphi(t))^n[a_0(t) + a_1(t)(x - \varphi(t) + \cdots)], \\ g &= (x - \varphi(t))^m[b_0(t) + \cdots]. \end{aligned} \tag{4.2}$$

The first equation leads (at leading order) to

$$-\dot{\varphi}(t)a_0^2(t)(2n(n-1)-2n^2)[x-\varphi(t)]^{2n-2}$$
$$+b_0^2(t)[x-\varphi(t)]^{2m}=0. \tag{4.3}$$

Since we exclude $n=0$, which is a regular point of f, we must have $2m=2n-2$, $m=n-1$, and $b_0^2(t)=2n\dot{\varphi}(t)a_0^2(t)$. The second equation gives

$$-\dot{\varphi}(t)a_0^2(t)[2n(n-1)(n-2)(n-3)$$
$$-8n^2(n-1)(n-2)+6n^2(n-1)^2]-pb_0^2(t)$$
$$\times[2(n-1)(n-2)-2(n^2-1)]=0. \tag{4.4}$$

We eliminate $b_0(t)$ from (4.3) and (4.4), and taking into account the fact that $a_0^2(t)\neq 0$, $\dot{\varphi}(t)\neq 0$ (since φ is arbitrary), and $n\neq 0$, we obtain $(2p+6)(n-1)=0$. The case $p=-3$ is special. In that case n may be arbitrary and f and g need not be entire functions. Still this case is integrable. Indeed, it we write $f=e^{W/2}$ and $g=e^{(\rho+W)/2}$, we obtain from the first equation

$$W_{xt}+e^{\rho}=0, \tag{4.5}$$

hence $\rho=\ln W_{xt}+i\pi$. The second equation yields

$$W_{xxxt}+3W_{xx}W_{xt}+3(W_{xx}+\rho_{xx})e^{\rho}=0, \tag{4.6}$$

which simplifies to $(-e^{\rho})_{xx}+3\rho_{xx}e^{\rho}=0$ or

$$2\rho_{xx}-\rho_x^2=0. \tag{4.7}$$

Integrating (4.7) we obtain $\rho=2A(t)-2\ln[x-x_0(t)]$. Thus

$$\phi=g/f=e^{\rho/2}=e^{A(t)}/[x-x_0(t)], \tag{4.8}$$

which has simple poles and so does $v=(d/dx)\ln f=W_x/2$. The residues of these poles, however, are arbitrary (in fact arbitrary functions of t), which shows that f and g have zeros of arbitrary order and are not the appropriate variables in which to perform the Painlevé analysis. For all other values of p, including $p=3$, which is the case for DSW, we must take $n=1$ and $m=0$ (g is regular).

We must now find the places where the free integration constants enter in the expansion: the "resonances." We start from

$$f=[x-\varphi(t)][a_0(t)+a_r(t)(x-\varphi(t))^r],$$
$$g=b_0(t)+b_r(t)(x-\varphi(t))^r,$$

with $b_0^2(t)=-2\dot{\varphi}(t)a_0^2(t)$. We thus obtain

$$-\dot{\varphi}(t)a_0(t)a_r(t)[2(r+1)r-(r+1)]$$
$$+2b_0(t)b_r(t)=0,$$
$$-\dot{\varphi}(t)a_0(t)a_r(t)[2(r+1)r(r-1)(r-2)$$
$$-8(r+1)r(r-1)]-pb_0(t)b_r(t)[2r(r-1)]=0.$$

This is a system of two equations for the two unknowns $a_r(t)$ and $b_r(t)$, which determines both functions unless the determinant vanishes:

$$\begin{vmatrix}(r+1)(r-2) & 1\\ r(r+1)(r-1)(r-6) & -pr(r-1)\end{vmatrix}=0.$$

Three roots are independent of p, namely -1 [freedom of $\varphi(t)$], 0 [freedom of $a_0(t)$], and $+1$. The fourth root is given by $r=(2p+6)/(p+1)$. For $p=3$ we have $r=3$. In general, r will not be an integer unless $p=(6-l)/(l-2)$ with $r=l$ an integer. The next step will be, for each value of integer $r=l$, to check whether the resonance condition is satisfied or whether a logarithm enters on the expansion.

Note that $l=2$ corresponds to an infinite value of p and must be rejected. For $l=1$ we get $p=-5$; $r=+1$ is a double root of the determinant but the matrix is still of rank 1. The two free constants are not a_1 and b_1 but rather a_1 and a logarithmic term in f with a free coefficient a_1' (which induces a logarithmic term b_1' on g) given by

$$b_1'b_0=-2a_1'a_0\dot{\varphi},\quad b_0b_1=\dot{\varphi}a_0(a_1'-2a_1).$$

The case $r=0$ corresponds to $p=-3$, which we know to be integrable, but the Painlevé property, true in terms of v and ϕ, is violated by f and g. For negative values of r, there can be no more checks, but we get a nongeneric (i.e., with only three free functions of time) expansion. We would not in such a case expect an integrable behavior.

For r a positive integer, larger or equal to 3, a compatibility condition has to be checked. It is indeed satisfied for $p=3$, which is the DSW equation. Thus this equation satisfies the Painlevé criterion as expected from its known integrable character. For $r=4$ ($p=1$) we find that the compatibility condition is *not* satisfied: logarithms enter of that order, the equation does not have the Painlevé property and is presumably not integrable. Beyond these values calculations become increasingly tedious but we do not expect any new Painlevé case to arise.

V. SUMMARY AND CONCLUSION

In this paper, we have examined an integrable equation first proposed by Drinfel'd, Sokolov, and Wilson, which can be obtained as a reduction of the KP hierarchy, as described by Jimbo and Miwa. This equation presents a most interesting feature: it possesses solitons that are static. Namely, among the time-independent solutions of this equation there exists a particular class that behave as solitons: after interacting with moving solitons they are simply shifted and induce just phase shifts on the moving waves. This is the reason why we have dubbed these solutions static solitons. The bilinear formalism was used for the explicit construction of such solutions and a general method of computation of the N-static-soliton solutions was presented based on the reduction of the τ function that Jimbo and Miwa have obtained for the KP hierarchy.

We have also examined the DSW equations from the point of view of the Painlevé analysis. A novel and interesting approach in this domain was the implementation of the singularity analysis directly on the bilinear form of the equations. Finally, although the main bulk of the paper was devoted to the study of the DSW equation, we also have examined an equation that was already studied under a different form, by Ito, and that we have shown to possess static soliton solutions also.

[1] R. Hirota and J. Satsuma, Phys. Lett. A **85**, 407 (1981).
[2] R. Hirota and J. Satsuma, J. Phys. Soc. Jpn. **51**, 3390 (1982).
[3] M. Sato and Y. Sato, RIMS Kôkyûroku **388**, 183 (1980); **414**, 181 (1981).
[4] E. Date, M. Kashiwara, and T. Miwa, J. Phys. Soc. Jpn. **50**, 3806, 3813 (1981).
[5] G. Wilson, Phys. Lett. A **89**, 332 (1982).
[6] V. G. Drinfel'd and V. V. Sokolov, Sov. Math. Dokl. **23**, 457 (1981).
[7] M. Jimbo and T. Miwa, Publ. RIMS, Kyoto Univ. **19**, 943 (1983).
[8] M. Ito, J. Phys. Soc. Jpn. **49**, 771 (1980).
[9] E. Date, M. Jimbo, M. Kashiwara, and T. Miwa, Publ. RIMS, Kyoto Univ. **18**, 1077 (1982).

Generalized Burgers equations and Euler–Painlevé transcendents. I

P. L. Sachdev, K. R. C. Nair, and V. G. Tikekar
Department of Applied Mathematics, Indian Institute of Science, Bangalore-560012, India

(Received 3 September 1985; accepted for publication 30 January 1986)

Initial-value problems for the generalized Burgers equation (GBE) $u_t + u^\beta u_x + \lambda u^\alpha = (\delta/2)u_{xx}$ are discussed for the single hump type of initial data—both continuous and discontinuous. The numerical solution is carried to the self-similar "intermediate asymptotic" regime when the solution is given analytically by the self-similar form. The nonlinear (transformed) ordinary differential equations (ODE's) describing the self-similar form are a generalization of a class discussed by Euler and Painlevé and quoted by Kamke. These ODE's are new, and it is postulated that they characterize GBE's in the same manner as the Painlevé equations categorize the Kortweg–de Vries (KdV) type. A connection problem, for some related ODE's satisfying proper asymptotic conditions at $x = \pm\infty$, is solved. The range of amplitude parameter is found for which the solution of the connection problem exists. The other solutions of the above GBE, which display several interesting features such as peaking, breaking, and a long shelf on the left for negative values of the damping coefficient λ, are also discussed. The results are compared with those holding for the modified KdV equation with damping.

I. INTRODUCTION

Two model equations have pervaded mathematical physics extensively. They are the Burgers equation

$$u_t + uu_x = (\delta/2)u_{xx} \tag{1.1}$$

and the Korteweg–de Vries (KdV) equation

$$u_t + \sigma uu_x = u_{xxx}. \tag{1.2}$$

While the former describes a balance between nonlinear convection and (small) viscous diffusion, the latter represents the effect on nonlinear convection of the (simplest form of) linear dispersion. In spite of the fact that the two equations epitomize quite distinct physical phenomena, their mathematical (and to some extent physical) structures lend themselves to several analogies. For example, while the Hopf–Cole transformation $u = -\delta(\log\phi)_x$ exactly linearizes (1.1) into the heat equation, its "straightforward generalization" $\sigma u = 12(\log F)_{xx}$ further "nonlinearizes" (1.2) to a uniformly second degree partial differential equation (PDE) of order 4. Nevertheless, this transformation helps the analysis of the soliton interaction in a simple manner.[1–3] Recently, Whitham[4] has given a representation of periodic waves as a sum of solitons for (1.2), which is analogous to an earlier one due to Parker[5] for the Burgers equation (1.1) in terms of a sum of shocks. The Burgers equation often motivated the analysis for the equation of the KdV type. The exception is the inverse scattering transform (IST), which reduces the problems for the KdV-type equations to the solution of linear integral equations of Gel'fand–Levitan type. Generalized Burgers equations (GBE's), such as

$$u_t + u^\beta u_x + \lambda u^\alpha = (\delta/2)u_{xx} \tag{1.3}$$

(where α and β are real), that extend the class (1.1) and that occur in many applications (see discussion below), however, do not seem to be amenable to the IST.

There is another strand that runs through the class of KdV-type equations admitting IST. According to Ablowitz *et al.*,[6] all these equations, either directly or through some simple transformations, admit similarity solutions governed by one of the six Painlevé equations. These nonlinear second-order differential equations belong to the class of 50 equations, classified by Painlevé and his contemporaries, whose only movable singularities are poles.[7] The solutions of only six of these equations are not expressible generally in terms of elliptic functions or other special functions—hence the name Painlevé transcendents. These, thus, in some sense, characterize the nonlinear dispersive model equations of the KdV type. It may be pointed out that the KdV type of equations can be interpreted as Hamiltonian systems,[3] while the Burgers type of equations are not Hamiltonian systems. Indeed the Painlevé transcendents reflect the complete integrability of a Hamiltonian system of KdV equations. It will be shown in a subsequent publication[8] that the GBE (1.3) does not enjoy the Painlevé property; this work will also deal with possible Lie–Bäcklund symmetries of (1.3).

It appears that GBE's such as (1.3) and

$$u_t + u^\alpha u_x + j(u/2t) = (\delta/2)u_{xx} \tag{1.4}$$

(where $j = 0,1,2$ for plane, cylindrical, and spherical symmetry) (which we shall discuss in detail in Part II) are expressible through similarity transformation (see Sec. III) to nonlinear ordinary differential equations (ODE's) of the form

$$yy'' + ay'^2 + f(x)yy' + g(x)y^2 + by' + c = 0, \tag{1.5}$$

whose solutions we refer to as Euler–Painlevé transcendents. Here $f(x)$ and $g(x)$ are (sufficiently smooth) arbitrary functions and a, b, and c are real constants. A special case of (1.5) with $b = 0$ and $c = 0$ was considered by Euler and Painlevé[9] and is, in fact, exactly linearizable by the simple transformation

$$y = v^{1/(a+1)}$$

to

$$v'' + fv' + (a+1)gv = 0. \tag{1.6}$$

However, for the GBE's (and indeed even for the Burgers equation), $b \neq 0$, and c may or may not be equal to zero. For

0022-2488/86/061506-17$02.50

the Burgers equations, $c=0$, $b=-2^{3/2}$, $a=-2$, $f(x)=2x$, $g(x)=-2$, and the solution is expressible in terms of a complementary error function and an exponential function. We postulate that the generalized Euler–Painlevé equation (GEPE) (1.5) represent GBE's and generally do not have solutions expressible in terms of known functions. The analogy with the Painlevé equations for the KdV type is obvious; hence the name Euler–Painlevé transcendents for the solutions of (1.5). We hasten to add that these equations seem to be analytically much nicer than the Painlevé transcendents and, in the physically interesting cases, do not exhibit any singularities.

The inviscid form of (1.3), with more general convective and damping terms, namely

$$u_t + g(u)u_x + \lambda h(u) = 0,$$
$$\lambda > 0, \quad g_u(u) > 0, \quad h_u(u) > 0, \quad \text{for } u > 0, \tag{1.7}$$

has been considered by Murray.[10] It includes model equations describing stress wave propagation in a nonlinear Maxwell rod with damping, ion exchange in fixed columns, and a realistic model equation, which has been suggested to explain the Gunn effect in semiconductors. In general, $\lambda > 0$ and the term $\lambda h(u) > 0$ is dissipative, but there are interesting cases for which λ can be negative.[11] Murray has found the asymptotic solution of (1.7) for the case $h(u) = O(u^\alpha)$ under the assumption that $0 < u \ll 1$. Choosing an initial single hump profile (which may be continuous or discontinuous at the front), he arrived at the following asymptotic behavior of the solution, which depends only on α and is independent of the form of $g(u)$ except for the requirement that $g_u(u) > 0$.

(i) If $0 < \alpha < 1$, the solution is unique under certain conditions and decays in a finite time and a finite distance.

(ii) If $\alpha = 1$, it decays in a finite distance but in an infinite time exponentially.

(iii) If $1 < \alpha \leqslant 3$, it decays in an infinite distance and infinite time like $O(t^{-1/(\alpha-1)})$.

(iv) If $\alpha > 3$, it decays like $O(t^{-1/2})$.

Lardner and Arya[12] have studied matched asymptotic solutions of a special case of (1.3), namely

$$u_t + uu_x + \lambda u = (\delta/2)u_{xx}, \tag{1.8}$$

under the constraint that the shock is thin. They have also considered an extended form of (1.8) wherein the coefficient of u_x is $\mu u + \gamma u^2 + \gamma_1 c(t)$; μ, γ, and γ_1 being constants. These model equations arise when considering the motions of continuous medium for which the stress–strain relation contains a large linear term proportional to the strain, a small term that is quadratic (and/or cubic) in the strain and a small dissipative term proportional to the strain rate. The λu term in (1.8) would arise in such a system if the equation of motion includes a small viscous damping term proportional to the velocity.[13]

There are several purposes to this paper.[14] We study an initial value problem for (1.3) for different values of α and β, both when λ is positive and when it is negative. The purpose is to discover for what values of α and β the solutions, for a class of the single hump form of initial conditions, are asymptotic to the (terminal) similarity solution. The initial conditions are taken to be either continuous or discontinuous at the front and vanishing at $\pm\infty$ in a "reasonable" manner. While for the former, the usual implicit finite difference scheme proves quite adequate, for the latter we have to resort to the pseudospectral method (see Sec. III) to tackle in a precise manner the initial shock discontinuity and its embryonic evolution. However, once the smooth Taylor shock has formed, we make a switch to the Douglas–Jones[15] implicit scheme, which now delivers an accurate solution over the long duration of the evolution of the profile in a relatively shorter computational time. It was clearly established by Sachdev and Seebass[16] that the Douglas–Jones implicit predictor–corrector method for solving nonlinear parabolic equations of the Burgers type is quite adequate. In particular, the evolution of a smooth initial N wave was considered. Ames[17] has given numerical solution of the initial boundary value problem for the Burgers equation (1.1) with the conditions $u(0,t) = u(1,t) = 0$, $u(x,0) = \sin \pi x$, $0 < x < 1$, $0 < t \leqslant T$, and has shown that the Douglas–Jones implicit scheme gives excellent agreement with the exact solution. (See also Mitchell and Griffiths,[18] p. 97.) We find that the similarity form emerges if $1 < \alpha \leqslant 3$ and is governed by a special case of GEPE (1.5). This range of α coincides with that of Murray's case (iii), which was identified by using the method of characteristics. We study in detail the GEPE's—their asymptotic behavior, series, and numerical solutions. We discover, in particular, when the single hump type of solutions of (1.3) exist [see Eq. (2.7)], tending to zero as $x \to \pm\infty$. This may be said to constitute a connection problem.[19] This may also be compared with the study of Miles,[20] who has treated Painlevé transcendents in a similar manner. We follow the (numerical) solutions of the initial value problem for the PDE (1.3) in the similarity range of parameters until they emerge as similarity solutions or intermediate asymptotics, as they are often referred to in the Soviet literature. To quote Barenblatt and Zeldovich,[21] "these are solutions which do not merely represent specific examples but describe the intermediate asymptotic behavior of solutions of wider classes of initial, boundary and mixed problems, that is, they describe the behavior of these solutions away from the boundaries of the regions of independent variables or alternatively, in the region where in a sense the solution is no longer dependent on the details of the initial and/or boundary conditions but the system is still far from being in a state of equilibrium." We thus identify the solutions of the GEPE's (with suitable simple transformations) with universal "similarity functions" for (1.3). We also study the non-self-similar cases for $\alpha \leqslant 1$ and $\alpha > 3$ and compare them with the asymptotic solutions of Murray, allowing, of course, for the absence of viscous effects and other assumptions in his study.

We note that a similar study for the modified KdV equation

$$U_t = \mu U U_x + \beta U_{xxx} - \lambda U,$$
$$U(x,0) = f(x), \tag{1.9}$$
$$U(x,t) \to 0 \quad \text{as} \quad |x| \to \infty, \quad \text{for all } t < \infty,$$

where μ and β are positive constants, and λ a positive or negative constant, was carried out by Leibovich and Randall.[22]

The scheme of this paper is as follows. Section II analyzes the self-similar solutions. Section III deals with the numerical solution of (1.3). Section IV connects the numerical solutions of (1.3) with the self-similar ones and demonstrates their self-similar character. Section V discusses the non-self-similar solutions. Finally, Sec. VI contains the conclusions of the present study.

II. ANALYSIS OF SELF-SIMILAR SOLUTION—EULER-PAINLEVÉ TRANSCENDENTS

With single hump type initial conditions (see Sec. III), we expect the solutions to have self-similar form analogous to the one for the Burgers equation (1.1), namely[1,23]

$$u = (\delta/t)^{1/2}\{[(2\pi)^{1/2}/(e^R-1)]\exp(\eta^2) + (\pi/2)^{1/2}\exp(\eta^2)\operatorname{erfc}\eta\}^{-1} \quad (2.1)$$

$$= (\delta/t)^{1/2}[1/H_B(\eta)], \quad (2.2)$$

where

$$\eta = x/(2\delta t)^{1/2}.$$

Whitham has identified this solution as one arising from the singular initial condition $u(x,0) = c_0 + A\delta(x)$. A change of variables $u = u_0 + \bar{u}, x = u_0 t + \bar{x}$ leaves (1.1) invariant and permits writing the initial condition as $\bar{u}(\bar{x},0) = A\delta(\bar{x})$. One may therefore assume the initial condition for (2.1) as $u = A\delta(x)$, where A is the area under the profile. Most self-similar solutions arise directly from singular initial conditions. The solution (2.1) represents a single pulse whose length increases with time, but whose Reynolds number

$$R = \frac{1}{\delta}\int_{-\infty}^{\infty} u\,dx, \quad (2.3)$$

which is the ratio of the area under the profile to viscous diffusion, is constant. This can be easily checked either by using the solution (2.1) or integrating (1.1) with respect to x from $-\infty$ to $+\infty$ and assuming that u and u_x vanish at $x = \pm\infty$. Hopf[24] refers to R as an integral of (1.1). The solution (2.1) motivates our study for (1.3). We seek a solution of the latter in the form

$$u = t^{a_1} f[(2\delta)^{-1/2} t^{b_1} x], \quad (2.4)$$

where a_1 and b_1 are real constants. Substitution of (2.4) into (1.3) shows that, for the similarity solution to exist, $a_1 = 1/(1-\alpha)$, $b_1 = -\frac{1}{2}$ so that (2.4) becomes

$$u = t^{1/(1-\alpha)}f(\eta), \quad \eta = x(2\delta t)^{-1/2}, \quad (2.5)$$

provided

$$\beta = (\alpha-1)/2. \quad (2.6)$$

Then, (1.3) reduces to

$$f'' + 2\eta f' - [4/(1-\alpha)]f - 4(2\delta)^{-1/2}f^{(\alpha-1)/2}f' - 4\lambda f^{\alpha} = 0, \quad (2.7)$$

where a prime denotes differentiation with respect to η. The form (2.5) shows that the solution decays explicitly with time if $\alpha > 1$ and grows if $\alpha < 1$. The form of (2.1) suggests that the denominator function $H_B(\eta)$ may admit generalization so that we transform (2.7) in terms of the "reciprocal" function

$$H = \delta^{1/2} f^{(1-\alpha)/2}. \quad (2.8)$$

(This transformation may be readily guessed so as to remove fractional powers of f.) Therefore, H as a function of η is governed by

$$HH'' - 2(1+\alpha_1)H'^2 + 2\eta HH' - 2H^2 - 2^{3/2}H' - 2\lambda_1 = 0, \quad (2.9)$$

where

$$\alpha_1 = \tfrac{1}{2}(3-\alpha)/(\alpha-1), \quad \lambda_1 = \lambda\delta(1-\alpha).$$

This is a special case of (1.5) with

$$a = -2(1+\alpha_1) = (1+\alpha)/(1-\alpha),$$
$$f(\eta) = 2\eta, \quad g(\eta) = -2,$$
$$b = -2^{3/2}, \quad c = -2\lambda_1.$$

We note that the function $H_B(\eta)$ for the Burgers equation, defined by (2.1), is a special case of (2.9) with $\alpha_1 = 0$, $\lambda_1 = 0$ corresponding to $\alpha = 3, \lambda = 0$ in (1.3). It is governed by

$$HH'' - 2H'^2 + 2\eta HH' - 2H^2 - 2^{3/2}H' = 0, \quad (2.10)$$

with the solution

$$H_B = [(2\pi)^{1/2}/(e^R-1)]\exp(\eta^2) + (\pi/2)^{1/2}\exp(\eta^2)\operatorname{erfc}\eta, \quad (2.11)$$

where the erfc has the expansion

$$\operatorname{erfc} z = 1 - \operatorname{erf} z = 1 - \frac{2z}{\pi^{1/2}}e^{-z^2}\sum_{k=0}^{\infty}\frac{(2z^2)^k}{1.3\cdots(2k+1)}, \quad |z| < \infty. \quad (2.12)$$

Even (2.10) is more general than the Euler–Painlevé equation

$$yy'' + ay'^2 + f(x)yy' + g(x)y^2 = 0, \quad (2.13)$$

since it has the additional term $-2^{2/3}H'$. Comparing (2.10) with (2.9), we find that the form (2.9) for the GBE (1.3) has different numerical coefficients and an additional constant term $2\lambda_1$. But these "simple" changes for nonlinear DE's make a drastic difference to the solution. Indeed, it does not seem possible to express the solution of (2.9) in terms of known functions analogous to (2.11). The series expansion for the solution (2.11) suggests that we may seek a similar one for (2.9) for the case of decaying solutions with $\alpha > 1$, namely

$$H = \sum_{n=0}^{\infty} a_n\eta^n. \quad (2.14)$$

Substitution of (2.14) into (2.9) gives

$$
\begin{aligned}
a_2 &= (1/a_0)\{(a_0^2 + 2^{1/2}a_1 + a_1^2) + \lambda_1 + \alpha_1 a_1^2\},\\
a_3 &= (1/3a_0)\{(a_0 a_1 + 2^{3/2}a_2 + 3a_1 a_2) + 4\alpha_1 a_1 a_2\},\\
a_{k+2} &= \frac{2a_k}{(k+1)(k+2)} + \frac{2a_{k+1}}{(k+2)a_0}(2^{1/2} + \alpha_1 a_1 + a_1) + \frac{2}{(k+1)(k+2)a_0}\\
&\quad \times \sum_{i=1}^{k} \Big\{ -\frac{(k+1-i)(k+2-i)}{2} a_i a_{k+2-i} + (1+\alpha_1)(i+1)(k+1-i)a_{i+1}a_{k+1-i}\\
&\quad + a_i a_{k-i} - (k+1-i)a_{i-1}a_{k+1-i}\Big\},\\
k &= 1,2,3,\ldots . \qquad (2.15)
\end{aligned}
$$

Thus, we have a two parameter a_0, a_1 family of series solutions. The convergence of this series by direct computation seems difficult to establish. For the Burgers equation (1.1), for which $\alpha = 3$, $\lambda = 0$, the function H_B given by (2.11) follows from (2.14) if $a_1 = -2^{1/2}$. The free parameter a_0 gives a single parameter family of solutions and corresponds to the (constant) value of the Reynolds number, which fixes a definite (single hump) profile. For the GEPE (2.9), it does not seem possible to fix *a priori* the range of the parameters a_0 and a_1 such that the series (2.14) converges over $-\infty < \eta < \infty$. We shall first find the asymptotic solution of the corresponding f equation (2.7) for large η and then numerically integrate it from $\eta \to \infty$ to $\eta \to -\infty$ and isolate the family of solutions f that vanish at $-\infty$. It is easily checked that the linearized form of (2.7), namely

$$f'' + 2\eta f' - [4/(1-\alpha)]f = 0, \qquad (2.16)$$

has the solution

$$
\begin{aligned}
f &= A\exp(-\eta^2)H_\nu(\eta)\\
&\sim A\exp(-\eta^2)(2\eta)^{2\alpha_1},\\
&\quad \text{as } \eta\uparrow +\infty, \qquad (2.17)
\end{aligned}
$$

and

$$f \sim O(\eta^{-2\alpha_1 - 1}), \quad \text{as} \quad \eta\downarrow -\infty, \qquad (2.18)$$

where

$$\nu = 2\alpha_1 = (3-\alpha)/(\alpha-1),$$

H_ν is the Hermite function of order ν, and A is the amplitude parameter. Thus, the linear solution decays exponentially as $\eta \to +\infty$, and algebraically as $\eta \to -\infty$, provided $2\alpha_1 + 1 > 0$, that is $\alpha > 1$.

We now pose the boundary value or connection problem for (2.7), namely

$$
\begin{aligned}
&f'' + 2\eta f' - [4/(1-\alpha)]f\\
&\quad - 4(2\delta)^{-1/2} f^{(\alpha-1)/2} f' - 4\lambda f^\alpha = 0, \qquad (2.19)\\
&f \sim A\exp(-\eta^2)H_\nu(\eta) \sim A\exp(-\eta^2)(2\eta)^{2\alpha_1}\\
&\quad (\eta\uparrow\infty), \qquad (2.20\text{a})\\
&f \to 0 \quad (\eta\downarrow -\infty), \qquad (2.20\text{b})
\end{aligned}
$$

and

$$|f| < \infty, \quad -\infty < \eta < \infty. \qquad (2.21)$$

Before solving (2.19)–(2.21) we note two special exact solutions of (2.19). The first is the constant solution

$$f = [\lambda(\alpha-1)]^{1/(1-\alpha)} = f_m, \qquad (2.22)$$

say. It is easy to check that f_m is also the maximum value of f that the maxima of the single hump solution can attain. This follows from (2.19) if we note that, at the maximum, $f' = 0$, $f'' < 0$, etc.

The second exact solution is

$$f = \begin{cases} (A_+\eta)^{2/(1-\alpha)}, & \eta > 0,\\ (-A_-\eta)^{2/(1-\alpha)}, & \eta < 0, \end{cases} \qquad (2.23)$$

where

$$
\begin{aligned}
A_+ &= (2/\delta)^{1/2}((\alpha-1)/(\alpha+1))\\
&\quad \times[((1+\lambda\delta(1+\alpha))^{1/2} + 1], \qquad (2.24)\\
A_- &= (2/\delta)^{1/2}((\alpha-1)/(\alpha+1))\\
&\quad \times[((1+\lambda\delta(1+\alpha))^{1/2} - 1]. \qquad (2.25)
\end{aligned}
$$

This solution is singular tending to infinity as $\eta \to 0$ for $\alpha > 1$. The corresponding solution for (2.9) is

$$
H(\eta) = \begin{cases} A_+\delta^{1/2}\eta, & \eta > 0, \qquad (2.26)\\ -A_-\delta^{1/2}\eta, & \eta < 0. \qquad (2.27) \end{cases}
$$

The solution (2.26) can be embedded in the two parameter

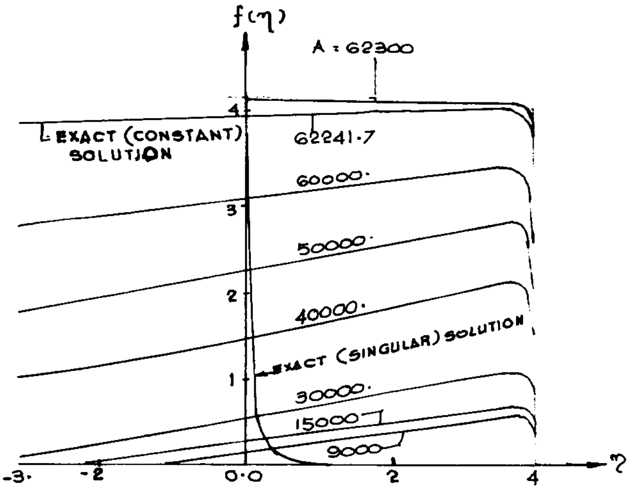

FIG. 1. Solution of Eq. (2.19) for various values of A and for $\alpha = 1.5$, $\lambda = 1$. Constant solution (2.22) and singular solution (2.23) are also shown.

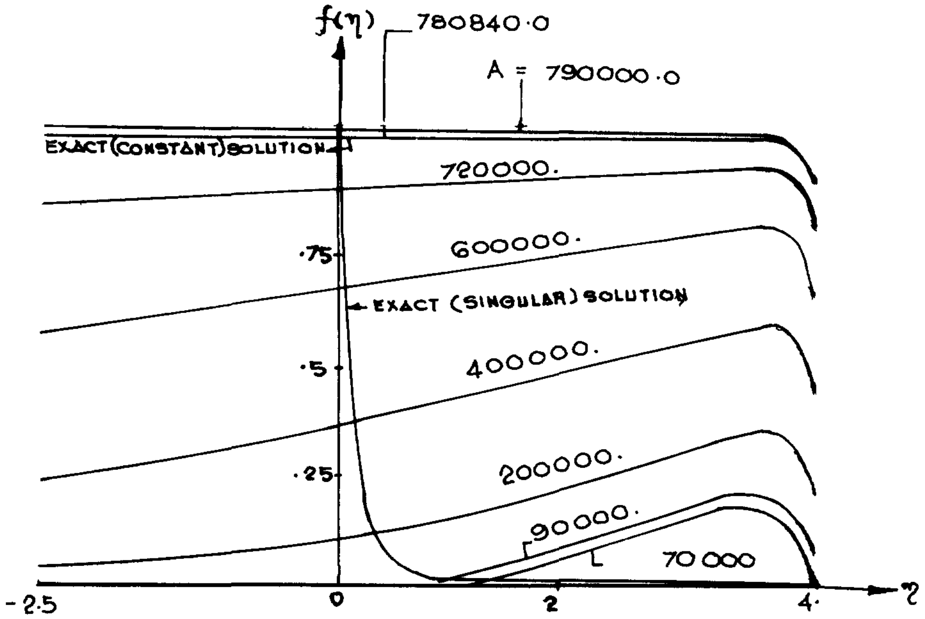

FIG. 2. Same as in Fig. 1 for $\alpha = 2$.

family of solutions

$$H = b_0\eta + \sum_{i=0}^{\infty} a_i \eta^{-i}, \qquad (2.28)$$

where

$$a_1 = -(1/2b_0)\left[(1+\alpha_1)b_0^2 + 2^{1/2}b_0 + \lambda_1\right],$$

$$a_2 = -a_0a_1/b_0, \qquad (2.29)$$

$$a_3 = (1/4b_0)\left[4a_0a_2 + (2a_1 - 2^{1/2} - 3b_1 - 2a_1b_0)a_1\right].$$

This singular solution may be compared with that of the Thomas–Fermi equation.[7,25]

We integrated (2.19) numerically from $\eta \sim 4$, choosing a certain value of A and initial conditions (2.20a) and proceeded towards negative η until the solution became essentially zero. Figures (1–4) show the solution for a set of values of $\alpha = 1.5,2,2.5,3$ in the similarity range of α, and the corresponding value of $\beta = (\alpha - 1)/2 = 0.25$, 0.5, 0.75, and 1, respectively. For each such pair (α,β), there is a value of $A = A_{\max}$ for which the solution does not decrease to zero as $\eta \rightarrow -\infty$ but, instead, approaches the (exact) constant solution (2.22) asymptotically. For $A > A_{\max}$, the integral curves grow monotonically to infinity as $\eta \rightarrow -\infty$. (Compare again with the solution of the Thomas–Fermi equation, Bender and Orzag,[25] Fig. 4.10.) Table I, gives the values of $\text{A}_{\max}$ for various (α,β) pairs. Figures (1–4) also include the singular solutions (2.23)–(2.25). From these

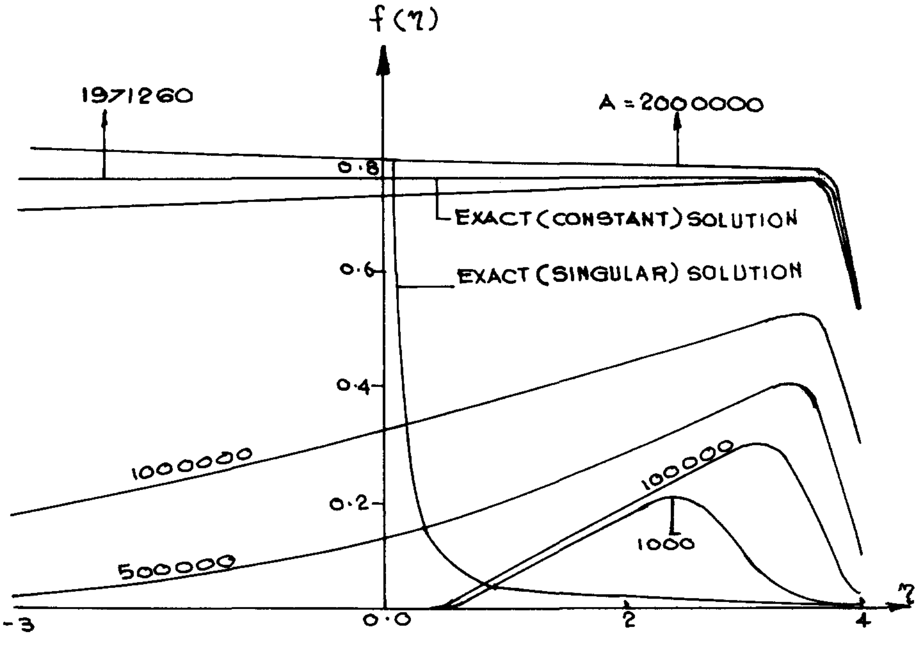

FIG. 3. Same as in Fig. 1 for $\alpha = 2.5$.

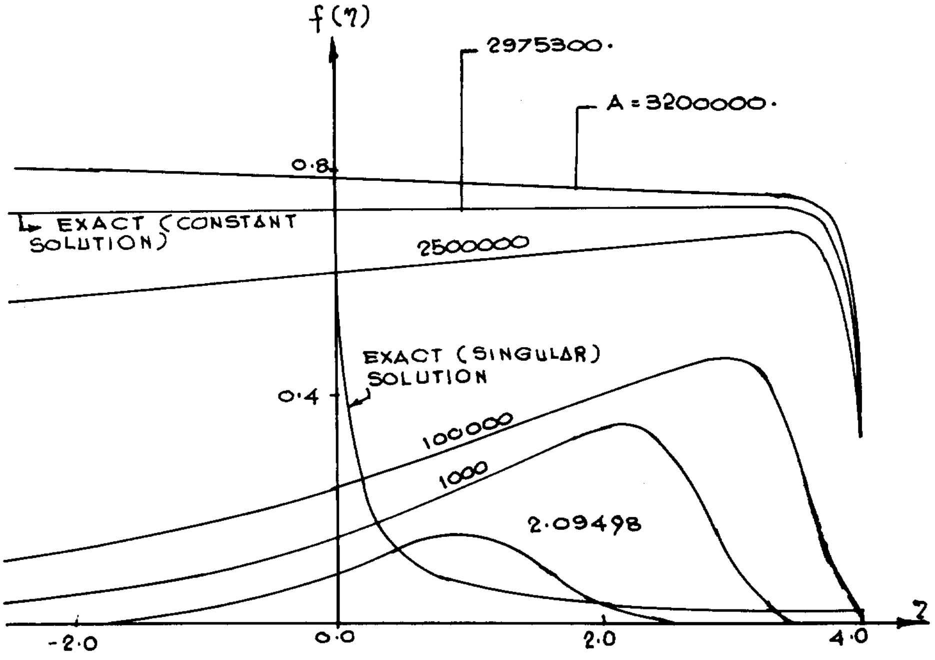

FIG. 4. Same as in Fig. 1 for $\alpha = 3$.

values of A lying between 0 and $A_{\max}$ and the numerical solution, we calculated the first two coefficients a_0 and a_1 of the series (2.14). They are the values of H and its derivative at $\eta = 0$. The function H is related to f by (2.8). Table II contains the ranges of a_0 and a_1 while Fig. 5 shows a_1 vs a_0, for various (α, β) pairs. With a_0 and a_1 thus determined, the series (2.14) was summed up and compared with the (exact) numerical solution. The series converged up to some value of η and then its convergence slowed down. However, analytic continuation of the series at a couple of η points yielded an accurate solution in a large finite range of η. The agreement of this analytic solution with the numerical one was found to be excellent, the discrepancy being $O(10^{-7})$ (see Table III).

It is clear from the asymptotic form (2.17) (and has been numerically checked by us) that the solution of the connection problem for (2.19) exists for all $\alpha > 1$. However, the similarity solution (2.5) of (1.3) is significant only in the range $1 < \alpha \leqslant 3$, since as we shall see in Sec. III, the solutions to initial value problem for (1.3) with suitable vanishing initial conditions at infinity approach the self-similar form asymptotically only in this range of α. The reason for this, as we shall discuss later, is the physically unrealistic decay predicted by the similarity form (2.5) for $\alpha > 3$. Nevertheless, Eq. (2.7) has the single hump solutions vanishing at $\eta = \pm \infty$ for positive λ and $\alpha > 1$. The left end limit of the range $1 < \alpha < 3$, namely, $\alpha = 1$, gives $\beta = 0$ according to (2.6) and the solution (2.5) becomes invalid but the PDE (1.3), in this case, is linear. We shall discuss its exact solu-

TABLE I. Critical values of the amplitude parameter A and $f_{\max}$ for different choices of α and β in the similarity range for $\lambda = 1$. [See Eqs. (2.20) and (2.22).]

			$f_{\max}$	
α	β	$A_{\max}$	Numerical	Exact
1.5	0.25	62 241.75	4.0	4.0
2.0	0.5	780 840.6	1.0	1.0
2.5	0.75	1971 256.0	0.763 143	0.763 143
3.0	1.0	2975 300.0	0.707 128	0.707 107

TABLE II. Coefficients a_0 and a_1 in the series (2.14) for the permissible (similarity) range of the amplitude parameter A corresponding to different values of α and β and for $\lambda = 1$, $\delta = 0.01$.

No.	A	a_0	a_1
(i) $\alpha = 3$, $\beta = 1$			
1	0.0	0.0	0.0
2	2.095	1.218 2	− 1.195 3
3	3.25	1.423 3	− 1.330 5
4	1 000.0	0.700 23	− 0.365 71
5	100 000.0	0.445 29	− 0.130 77
6	2500 000.0	0.167 09	− 0.005 55
7	2953 000.0	0.141 414	0.0
(ii) $\alpha = 2.5$, $\beta = 0.75$			
1	0.0	0.0	0.0
2	500 000.0	0.490 37	− 0.170 83
3	1000 000.0	0.241 86	− 0.031 08
4	1800 000.0	0.134 62	− 0.002 2
5	1900 000.0	0.127 28	− 0.000 85
6	1971 256.0	0.122 48	0.0
(iii) $\alpha = 2$, $\beta = 0.5$			
1	200 000.0	0.324 06	− 0.070 34
2	400 000.0	0.169 60	− 0.013 36
3	600 000.0	0.122 76	− 0.003 59
4	720 000.0	0.106 51	− 0.000 95
5	780 841.0	0.1	0.0

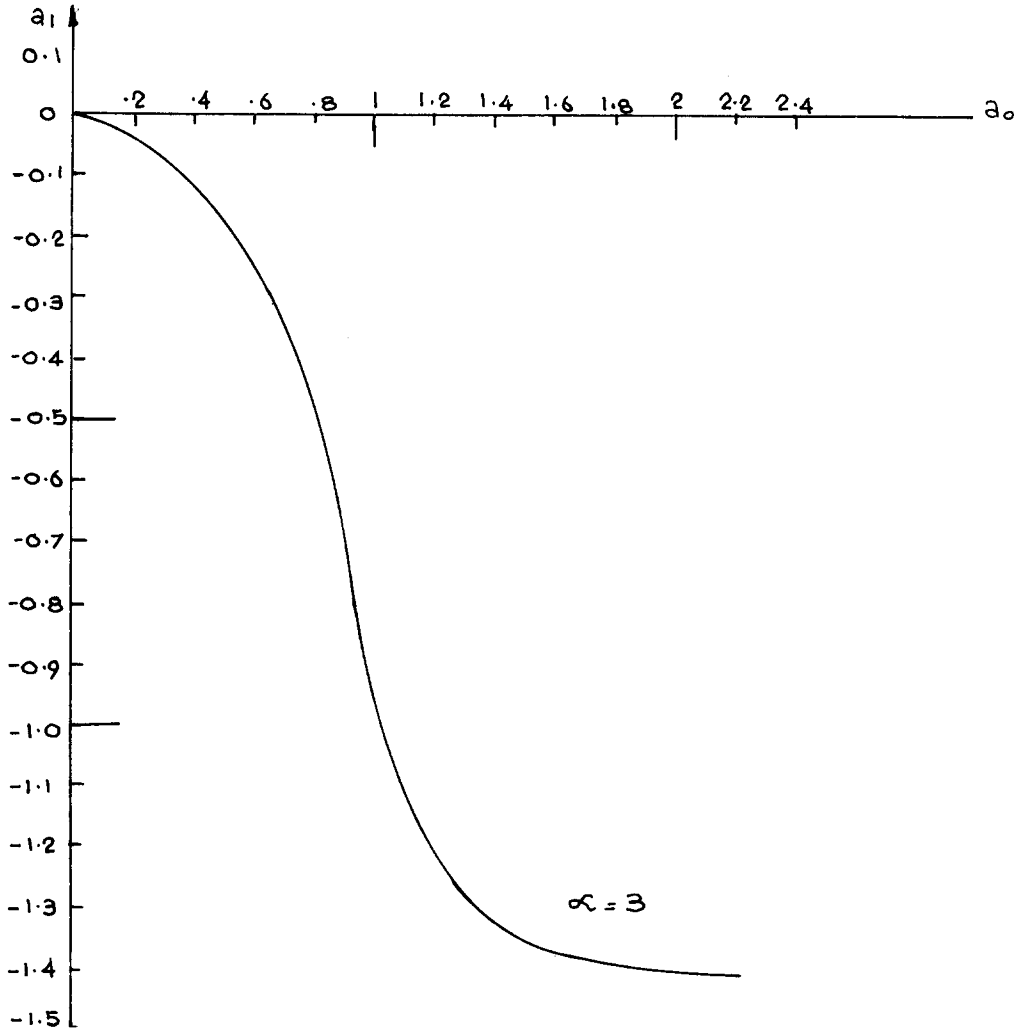

FIG. 5. a_1 versus a_0 for $0 \leqslant A \leqslant A_{\max}$ [see Eq. (2.14)].

TABLE III. Comparison of series solution (2.14) and numerical solution of (2.19) for $\alpha = 3, \beta = 1, \lambda = 1$, and $\delta = 0.01$. The a_i in (2.19) from $i = 0$ to $i = 13$ are 1.2182, − 1.1957, 0.987 36, − 0.603 57, 0.346 00, − 0.141 34, 0.054 71, − 0.007 13, − 0.002 59, 0.005 37, − 0.002 97, 0.001 55, − 0.000 44, 0.00 07.

	Series solution		Numerical solution
η	$H(\eta)$	$f(\eta)$	$f(\eta)$
− 3.0	36.775 90	0.002 7192	0.002 7192
− 2.5	29.151 79	0.003 4303	0.003 4303
− 2.0	19.825 09	0.005 0441	0.005 0441
− 1.5	10.269 01	0.009 7380	0.009 7380
− 1.0	4.541 242	0.022 0204	0.022 0204
− 0.5	2.165 305	0.046 1829	0.046 1829
0.0	1.218 223	0.082 0868	0.082 0867
0.5	0.809 7551	0.123 4941	0.123 4941
1.0	0.659 5272	0.151 6237	0.151 6237
1.5	0.854 7518	0.116 9930	0.116 9928
2.0	2.926 066	0.034 1755	0.034 1754
2.5	24.990 6	0.004 0015	0.004 0015
3.0	387.044	0.000 2583	0.000 2584
3.5	9 993.162	0.000 0100	0.000 0100
4.0	366 869.9	0.000 0003	0.000 0003

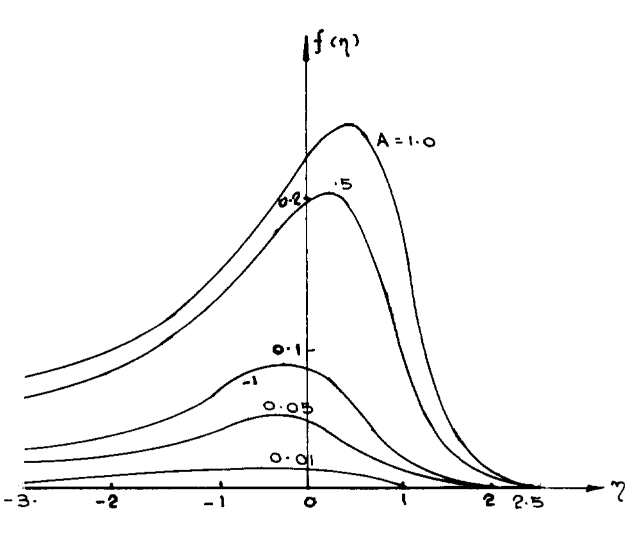

FIG. 6. Solution of Eq. (2.19) for various values of A, and for $\alpha = 4$ and $\lambda = -1$.

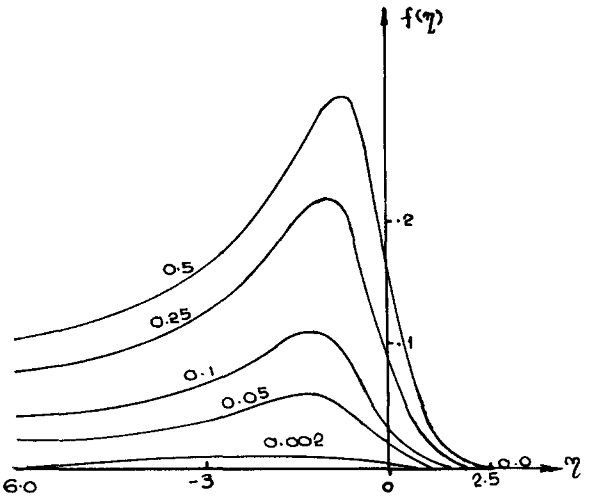

FIG. 7. Same as in Fig. 6 for $\alpha = 4, \lambda = -5$.

tion in Sec. III. This solution displays, in contrast, an exponential decay.

When λ is negative, the asymptotic form (2.5) is still valid for $\alpha > 1$, but the constant solution (2.22) ceases to exist. This suggests that, in this case, there is probably no upper limit $A_{\max}$ to the amplitude that the linear solution can possess. Our numerical study of (2.19)–(2.21) for $\lambda < 0$ and $\alpha > 1$ confirms this conclusion (see Figs. 6–8). We shall see in Sec. IV that these solutions do not constitute intermediate asymptotics. We arrive at the conclusion that the solutions to the problem (2.19)–(2.21) exist for $\alpha > 1$ and all λ, but these solutions are intermediate asymptotics only for $1 < \alpha \leqslant 3$ and $\lambda > 0$.

We have carried out the numerical study for (2.19) rather than for GEPE (2.9) because the asymptotic solutions for the former decay as $|\eta| \to \infty$, while for the latter they would grow to infinity [see (2.8)]. Nevertheless, Eq. (2.9) seems to be analytically more important and involves, unlike (2.7), only integral powers of H and its derivatives.

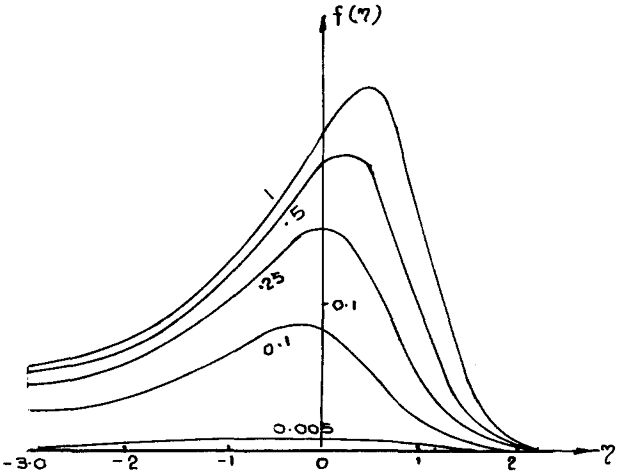

FIG. 8. Same as in Fig. 6 for $\alpha = 5, \lambda = -5$.

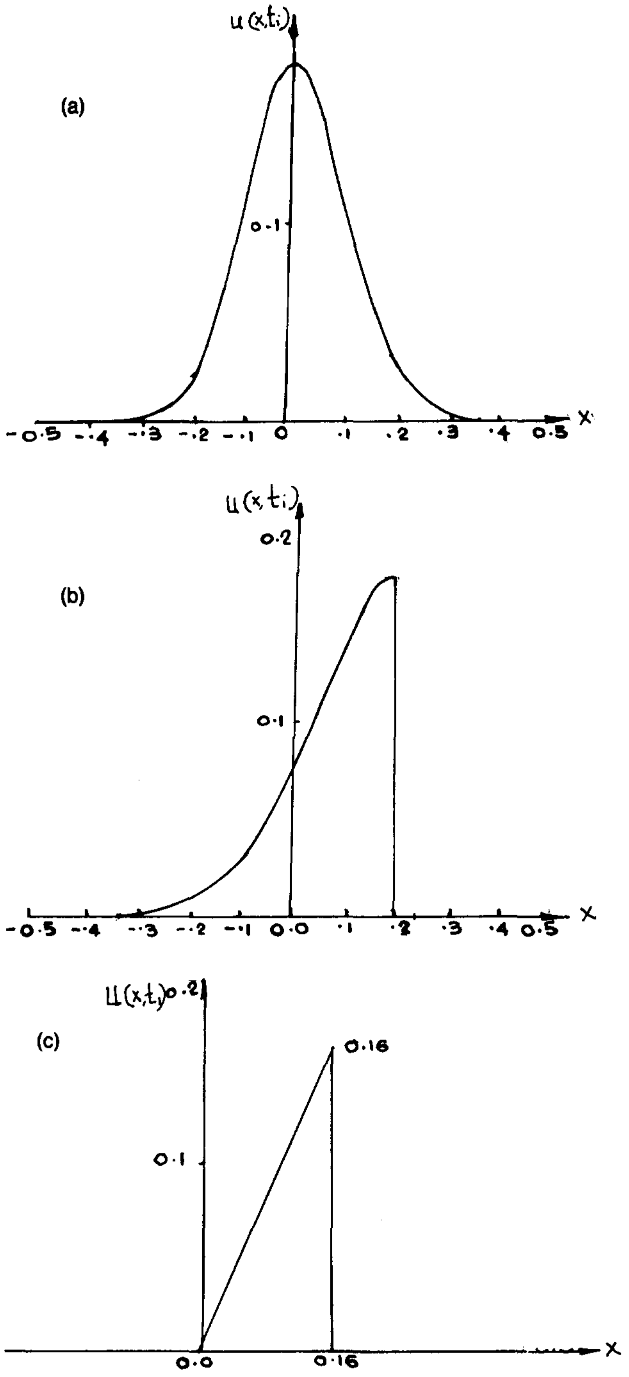

FIG. 9. Initial profiles for solving Eq. (1.3).

III. PSEUDOSPECTRAL AND IMPLICIT DIFFERENCE SCHEME FOR SOLVING EQ. (1.3)

We solve (1.3) subject to the initial conditions

$$u(x,t_i) = \begin{cases} 0, & x < x_0, \\ f(x), & x_0 \leqslant x \leqslant x_1, \\ 0, & x > x_1, \end{cases} \qquad (3.1)$$

where the function $f(x)$ has the typical forms shown in Fig. 9. Unlike in the study of Murray, we do not restrict ourselves to the case of $0 < f(x) < 1$, nor the assumption $|u| \ll 1$, which he imposes to get some asymptotic results. As mentioned in the Introduction, the Douglas–Jones implicit predictor–corrector method for the nonlinear parabolic equation is quite adequate to describe the evolution of initial smooth single hump profiles. The difference analog of Eq. (1.3) is

$$\begin{aligned} &u_{i+1,j+1/2} - 2\left(1+\frac{2(\Delta x)^2}{\delta(\Delta t)}\right)u_{i,j+1/2} + u_{i-1,j+1/2} \\ &\quad = \frac{2(\Delta x)^2}{\delta}\left(\lambda u^{\alpha}_{i,j} - \frac{2}{\Delta t}u_{i,j}\right) \\ &\quad\quad + \frac{\Delta x}{\delta}u^{\beta}_{i,j}(u_{i+1,j} - u_{i-1,j}) \quad \text{(predictor)}, \qquad (3.2) \end{aligned}$$

and

$$\begin{aligned} &\left(1-\frac{\Delta x}{\delta}u^{\beta}_{i,j+1/2}\right)u_{i+1,j+1} - 2\left(1+\frac{2(\Delta x)^2}{\delta(\Delta t)}\right)u_{i,j+1} \\ &\quad + \left(1+\frac{\Delta x}{\delta}u^{\beta}_{i,j+1/2}\right)u_{i-1,j+1} \\ &\quad = \left(\frac{\Delta x}{\delta}u^{\beta}_{i,j+1/2} - 1\right)u_{i+1,j} + 2\left(1-\frac{2(\Delta x)^2}{\delta(\Delta t)}\right)u_{i,j} \\ &\quad\quad - \left(1+\frac{\Delta x}{\delta}u^{\beta}_{i,j+1/2}\right)u_{i-1,j} \\ &\quad\quad + \frac{4\lambda(\Delta x)^2}{\delta}u^{\alpha}_{i,j+1/2} \quad \text{(corrector)}. \qquad (3.3) \end{aligned}$$

Here, $u_{i,j} = u(i\Delta x, j\Delta t)$ and Δx and Δt are spatial and time mesh sizes, respectively. This difference scheme has a truncation error $O(\Delta x^2 + \Delta t^2)$. Douglas and Jones have demonstrated the convergence of the difference scheme (3.2) and (3.3) for (1.3). However, this scheme is not adequate to solve (1.3) with a sharp discontinuous initial profile; that is, to discover the evolution of a shock wave through its "embryonic shock" region. The reason is that the accuracy of the solution of (1.3) with an initial discontinuous profile by implicit scheme (3.2) and (3.3) is severely affected. So we resort to a numerical scheme referred to as pseudospectral.[26–28] The essence of the pseudospectral method is that the spatial derivatives u_x, u_{xx} of the distribution $u(x,t)$ are computed very accurately by finite Fourier transformation. The finite Fourier transform of $u(x,t)$ is defined as

$$\bar{u}(k_j,t) = \frac{1}{K}\sum_{l=0}^{K-1} u(l\Delta x,t)\exp(-ik_j l\Delta x) \qquad (3.4)$$

over the interval $(0,2\pi)$ of x. Here, $\Delta x = 2\pi/K$, K denoting the number of mesh points and the k_j are the wave numbers varying between 0 and $K-1$. The inverse finite Fourier transform is

$$u(l\Delta x,t) = \sum_{|k_j|<K/2} \bar{u}(k_j,t)\exp(ik_j l\Delta x). \qquad (3.5)$$

The spatial derivatives at the mesh points are

$$u_x(l\Delta x,t) = \sum_{|k_j|<K/2} ik_j\bar{u}(k_j,t)\exp(ik_j l\Delta x), \qquad (3.6)$$

$$u_{xx}(l\Delta x,t) = \sum_{|k_j|<K/2} (ik_j)^2\bar{u}(k_j,t)\exp(ik_j l\Delta x). \qquad (3.7)$$

The solution $u(x,t+\Delta t)$ at the next time level is obtained from the truncated Taylor series

$$\begin{aligned} u(x,t+\Delta t) &= u(x,t) + \Delta t u_t \\ &\quad + (\Delta t^2/2!)u_{tt} + (\Delta t^3/3!)u_{ttt}, \qquad (3.8) \end{aligned}$$

TABLE IV. Evolution of the initial discontinuous profile under Burgers equation to the self-similar form as evidenced by the convergence of the area A_0 under the profile to a fixed constant value. Here, $\beta = 1$, $\lambda = 0$, $\delta = 0.001$.

Time	A_0
1.0	0.013 067
1.01	0.013 078
1.02	0.013 151
1.03	0.013 196
1.04	0.013 209
1.05	0.013 211
1.06	0.013 211
1.07	0.013 211
1.08	0.013 210
1.09	0.013 210
1.10	0.013 209
1.11	0.013 209
1.15	0.013 209
1.20	0.013 209
1.25	0.013 209
1.3	0.013 209

wherein the time derivatives u_t, u_{tt}, etc. are substituted from Eq. (1.3) in terms of the spatial derivatives as

$$\begin{aligned} u_t &= -u^{\beta}u_x - \lambda u^{\alpha} + (\delta/2)u_{xx}, \\ u_{tt} &= -\beta u^{\beta-1}u_t u_x - u^{\beta}u_{xt} \\ &\quad -\lambda\alpha u^{\alpha-1}u_t + (\delta/2)u_{xxt}, \\ u_{ttt} &= -\beta(\beta-1)u^{\beta-2}u_t^2 u_x - 2\beta u^{\beta-1}u_{xt}u_t \\ &\quad -\beta u^{\beta-1}u_x u_{tt} - u^{\beta}u_{xtt} \\ &\quad -\alpha(\alpha-1)\lambda u^{\alpha-2}u_t^2 - \alpha\lambda u^{\alpha-1}u_{tt} + (\delta/2)u_{xxtt}. \end{aligned} \qquad (3.9)$$

Gazdag[27] has given a stability analysis of a pseudospectral scheme for the inviscid Burgers equation; the amplitude and phase of the error in the Fourier components remain bounded. In our computations we used four terms in the Taylor series (3.8) so that the truncation error is $O(\Delta t^4)$. The do-

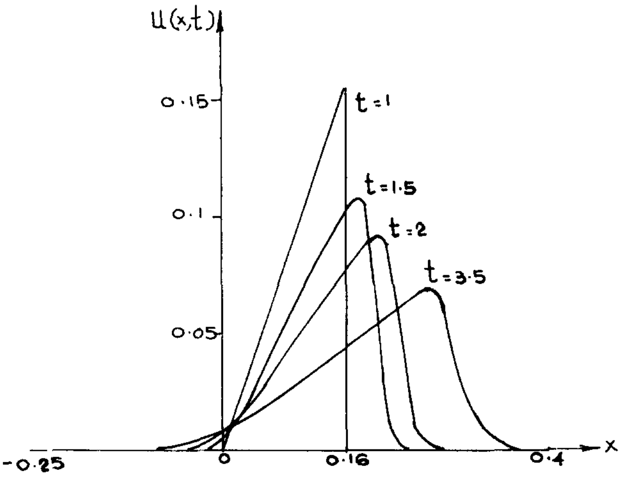

FIG. 10. Solution of Eq. (1.3) with initial discontinuous profile.

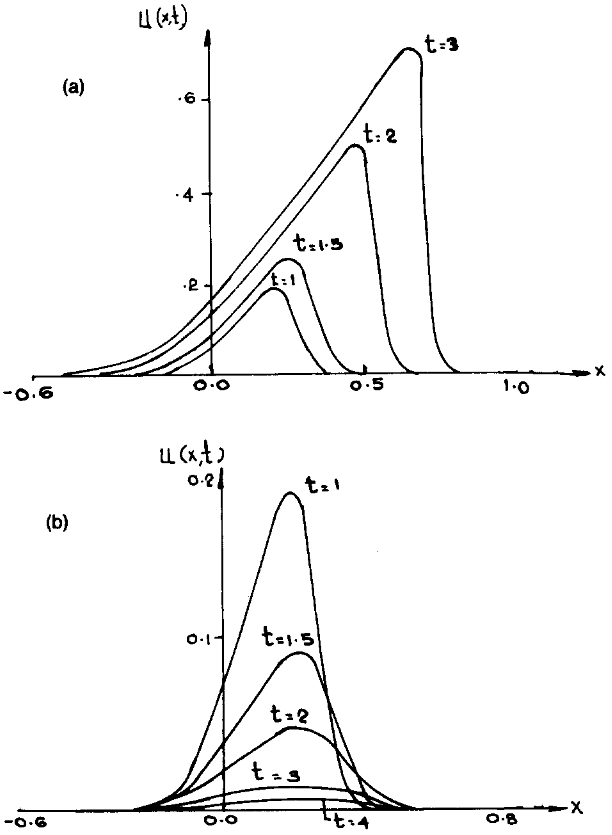

FIG. 11. Solution of Eq. (1.3) for $\alpha = \beta = 1$: (a) $\lambda = -1$, (b) $\lambda = 1$.

main $(0,2\pi)$ was divided into 128 mesh points in which the initial discontinuous nonzero profile occupied about 64 points in the middle of the domain so as to allow it to grow due to diffusion as it evolves. We chose the mesh sizes $\Delta x = 0.005$ and $\Delta t = 0.01$, and the initial profile $f(x) = x$, $x_0 = 0, x_1 = 0.2, t_i = 1$ [see Eq. (3.1)]. As the computation commenced, a tail of $O(10^{-3})$ on either side of the nonzero part of the profile was noticed. Being spurious, it was artificially cut off. The tail in the subsequent calculations was much smaller and, in fact, vanished after a few steps. For $\alpha = \beta = 1$, Eq. (1.3) reduces to

$$u_t + uu_x + \lambda u = (\delta/2)u_{xx}. \tag{1.8}$$

Integrating (1.8) with respect to x we get

$$A = A_0 e^{-\lambda t}, \tag{3.10}$$

where A_0 is a constant of integration, $A = \int_{-\infty}^{\infty} u\,dx$ and u, $u_x, u_{xx} \to 0$ as $x \to \pm\infty$. At each time level t_n, we calculated $A_0 = A_n e^{\lambda t_n}$. The embryonic shock for the solution of Eq. (1.8) settled down to a smooth Taylor structure when A_0 converged to a definite finite value (see Table IV for the convergence of A_0). This indicated the evolution of the discon-

TABLE V. Solution of Eq. (1.3) with initial discontinuous profile by pseudospectral and implicit predictor–corrector schemes at $t = 1.75$ (an implicit scheme was used when the profile became smooth by a pseudospectral scheme). Here, $\alpha = \lambda = 0, \beta = 1, \delta = 0.001, x_1 = 0.16$.

	$u(x,1.75)$	
x	Pseudospectral	Implicit
−0.08	0.000 02	0.000 02
−0.04	0.001 02	0.001 00
0.0	0.008 72	0.008 70
0.04	0.026 42	0.026 47
0.08	0.048 46	0.048 57
0.12	0.071 19	0.071 35
0.16	0.093 64	0.093 96
0.20	0.084 80	0.084 81
0.24	0.001 83	0.001 76
0.26	0.000 09	0.000 10

TABLE VI. Comparison of numerical (pseudospectral and implicit finite difference) solutions and the exact analytic solution for Burgers equation, with smooth initial data at $t_i = 1$. Here, $\lambda = 0, \beta = 1, \delta = 0.001$.

	$u(x,2)$		
x	Implicit	Pseudospectral	Exact
−0.10	0.000 742	0.000 742	0.000 742
−0.08	0.001 869	0.001 870	0.001 870
−0.06	0.003 983	0.003 985	0.003 985
−0.04	0.007 341	0.007 342	0.007 342
−0.02	0.012 000	0.012 000	0.012 000
0.00	0.017 843	0.017 841	0.017 841
0.02	0.024 660	0.024 657	0.024 657
0.04	0.032 232	0.032 227	0.032 227
0.06	0.040 367	0.040 362	0.040 362
0.08	0.048 918	0.048 913	0.048 913
0.10	0.057 774	0.057 767	0.057 767
0.12	0.066 840	0.066 826	0.066 826
0.14	0.075 986	0.075 934	0.075 934
0.16	0.084 648	0.084 423	0.084 423
0.18	0.088 732	0.088 036	0.088 034
0.20	0.065 779	0.066 041	0.066 041
0.22	0.017 660	0.018 405	0.018 405
0.24	0.002 151	0.002 161	0.002 161
0.26	0.000 190	0.000 180	0.000 180
0.28	0.000 015	0.000 012	0.000 012
0.30	0.000 001	0.000 001	0.000 001

TABLE VII. Comparison of the exact and numerical solutions for the special (linear) PDE with $\beta = 0, \alpha = 1$ [see eqs. (3.11) and (3.12)].

	$u(x,2)$		$u(x,4)$	
x	Numerical	Exact	Numerical	Exact
−0.11	0.000 51	0.000 57	0.0	0.0
0.37	0.005 29	0.005 46	0.0	0.0
0.85	0.028 86	0.028 97	0.000 03	0.000 03
1.33	0.086 56	0.086 33	0.000 16	0.000 17
1.81	0.144 87	0.144 63	0.000 71	0.000 72
2.29	0.135 98	0.136 21	0.002 31	0.002 33
2.77	0.071 86	0.072 11	0.005 63	0.005 62
3.25	0.021 43	0.021 46	0.010 23	0.010 19
3.73	0.003 57	0.003 59	0.013 87	0.013 83
4.21	0.000 32	0.000 34	0.014 07	0.014 09
4.69	0.000 01	0.000 02	0.010 70	0.010 75
5.17	0.0	0.0	0.006 12	0.006 15
5.65	0.0	0.0	0.002 64	0.002 64
6.13	0.0	0.0	0.000 86	0.000 85

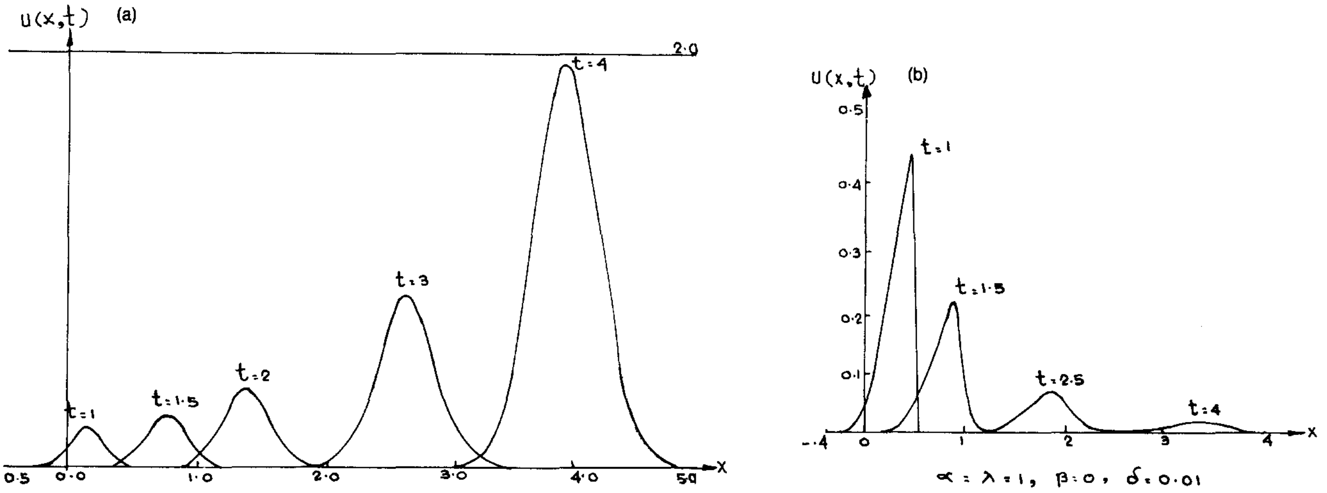

FIG. 12. Solution of Eq. (3.11) for (a) $\lambda = -1$, (b) $\lambda = 1$.

tinuous profile into a smooth Taylor structure (see Fig. 10). At this stage we switched over to implicit scheme (3.2) and (3.3) and continued the computation. The latter is less expensive in terms of computer time and is sufficiently accurate [see Table V, and Fig. 11 for solution of Eq. (1.8) with a smooth initial profile]. We shall show the accuracy of the implicit finite difference scheme (3.2) and (3.3) with the help of exact solutions of two special cases of the PDE (1.3). One is the Burgers equation ($\beta = 1, \lambda = 0$), whose exact solution is given by Eq. (2.1). The other is the linear partial differential equation

$$u_t + u_x + \lambda u = (\delta/2) u_{xx} \tag{3.11}$$

($\alpha = 1, \beta = 0$). It has an exact single hump solution

$$u = \frac{A}{(2\delta t)^{1/2}} \exp\left\{ -\left[\frac{(x-t)^2}{2\delta t} + \lambda t \right] \right\}, \tag{3.12}$$

which decays exponentially with time (and distance). This is a product solution, quite different from (2.5). The numerical solution (by implicit predictor–corrector scheme) of (1.3) with initial smooth profile and the exact solutions (2.1) and (3.12) are presented in Tables VI and VII. The agreement is very good, ensuring the adequacy of the implicit predictor–corrector scheme for solving (1.3) with smooth initial profiles (see Fig. 12).

IV. TRANSITION OF SOLUTION OF INITIAL VALUE PROBLEMS TO SELF-SIMILAR FORM OR INTERMEDIATE ASYMPTOTICS

We have already detailed in Sec. III the numerical scheme, implicit finite difference for smooth initial profiles and pseudospectral for the discontinuous initial profiles. We give here the results of the computations, when the initial profiles evolve into (fully nonlinear) self-similar solutions discussed in Sec. II. We find that for $1 < \alpha \leqslant 3$ and $\lambda > 0$, the initial profile, continuous or discontinuous at the front, soon evolves into a self-similar form discussed earlier. Figure 13 shows a typical evolution of the profile to its self-similar form for the case $\alpha = 3, \beta = 1$, both when the initial amplitude $u_{\max}$ is less than 1 and when it is greater than 1; only it is required to vanish at $x = \pm\infty$ in a reasonably smooth way. The self-similar regime was identified by matching the maximum of the numerical solutions of (1.3) and (2.19) and ensuring that the difference between the two solutions in the entire interval $-\infty < x < \infty$ is less than 5×10^{-3}. This required a proper choice of the amplitude parameter A [see Eq. (2.17)]. Table VIII shows the times t_s at which the self-similarity comes about for different pairs (α, β). These terminal solutions are fully nonlinear and hold for all $t > t_s$. Their decay law is given by $u_{\max} = O(t^{1/(1-\alpha)})$, which is the same as found by Murray for (1.7) for the range $1 < \alpha \leqslant 3$, subject to the condition $|u| \ll 1$ and $g_u \neq 0, u \geqslant 0$. We note here, however, that the condition $g_u(u) \neq 0$ does not play any role in our case.

V. NON-SELF-SIMILAR SOLUTION

For $\alpha > 3$, the numerical solution of (1.3) does not obey the asymptotic decay law $u_{\max} = O(t^{1/(1-\alpha)})$; instead $u_{\max}$ decays like $O(t^{-1/2})$ in agreement with Murray[10] (see Table IX). This is plausible, since in the present case the final (old age) regime of the wave is essentially linear, nonlinear convection and damping playing no significant role. The single hump in this case has the form $u = ct^{-1/2} \exp(-\eta^2)$. The self-similar decay law $O(t^{1/(1-\alpha)})$, on the other hand, predicts a rate slower than $t^{-1/2}$ for $\alpha > 3$. Thus, for $\alpha > 3$, even though the self-similar form of (1.3) exists and satisfies boundary conditions at $x = \pm\infty$, it is physically unrealistic. Another interesting feature that emerges from our numerical solution of Eq. (2.19) for $\alpha > 3$ and $\lambda < 0$ is the appearance of a shelf on the left end tail of the self-similar profile (see Fig. 14). The solution decays in an extremely slow manner characteristic of the shelf. Equation (1.3) shares this feature with the modified KdV[22] equation (1.9) when $\mu > 0$. It must be pointed out, however, that these self-similar solutions for $\lambda < 0, \alpha > 3$ are not intermediate asymptotics and the solution of (1.3) does not manifest any shelves (see Fig. 15).

Now we turn to the case $0 < \alpha \leqslant 1$, for which the self-similar form does not exist. The solution depends on the sign of λ. If $\lambda > 0$, the initial profile shrinks, decays, and becomes

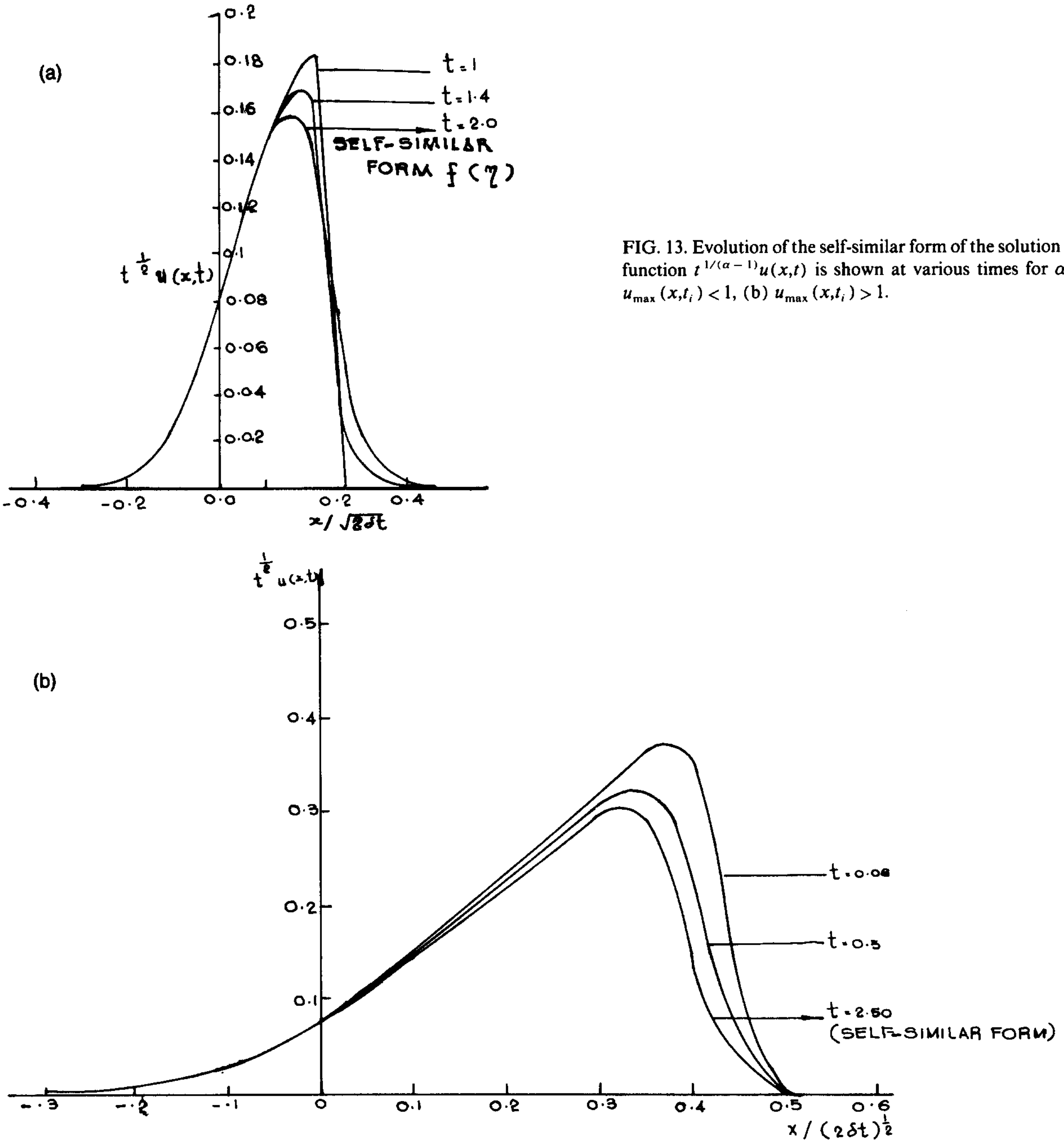

FIG. 13. Evolution of the self-similar form of the solution of Eq. (1.3). The function $t^{1/(\alpha-1)}u(x,t)$ is shown at various times for $\alpha = 3$, $\lambda = 1$: (a) $u_{\max}(x,t_i) < 1$, (b) $u_{\max}(x,t_i) > 1$.

extinct in a finite distance and a finite time in agreement with Murray's[10] analysis. The case of negative damping, $\lambda < 0$, unfolds several fascinating features. The nature of solution again depends crucially on the parameter α. The special val-

TABLE VIII. Approximate time t_s when the self-similar regime for GBE (1.3) sets in for different α and β. The initial time is $t_i = 1$.

α	β	t_s
1.5	0.25	61
2.0	0.50	16
2.5	0.75	4
3.0	1.00	3

TABLE IX. The (large time) asymptotic behavior of the solution of Eq. (1.3) for $\alpha > 3$: comparison of numerical $u_{\max}$ and $u_{\max} = Ct^{-1/2}$.

	$u_{\max}$			$u_{\max}$	
t	Numerical	Analytic	t	Numerical	Analytic
$\alpha = 5, \beta = 2, \lambda = 1, \delta = 0.01$			$\alpha = 4, \beta = 1.5, \lambda = 1, \delta = 0.01$		
5.0	0.0867	0.0855	5.0	0.0856	0.0846
10.0	0.0621	0.0604	10.0	0.0613	0.0598
15.0	0.0510	0.0493	15.0	0.0503	0.0488
20.0	0.0443	0.0427	20.0	0.0437	0.0423
$\alpha = 5, \beta = 2, \lambda = -1, \delta = 0.01$			$\alpha = 4, \beta = 1.5, \lambda = -1, \delta = 0.01$		
3.5	0.1028	0.1022	3.5	0.1023	0.1019
4.0	0.0965	0.0956	4.0	0.0960	0.0953
5.0	0.0869	0.0855	5.0	0.0865	0.0852
6.0	0.0796	0.0781	6.0	0.0792	0.0772

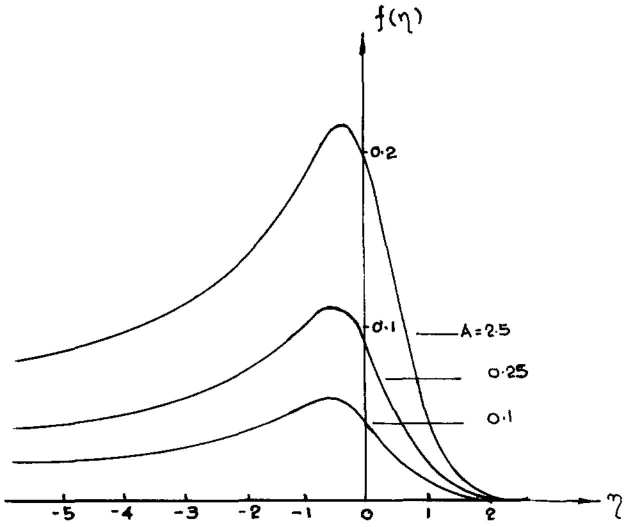

FIG. 14. Solution of Eq. (2.19) for $\alpha = 5, \lambda = -1$. Shelf appears on the left.

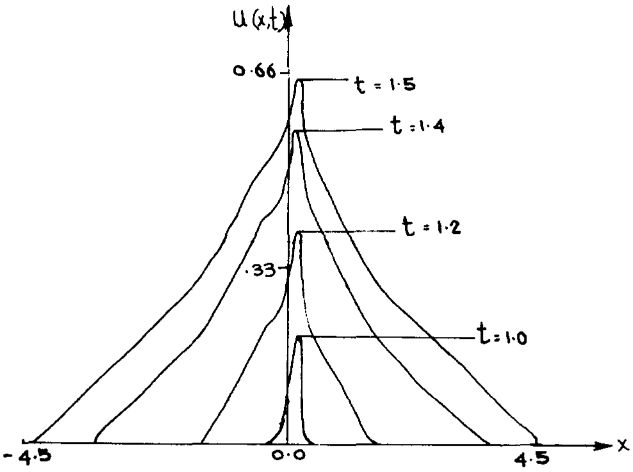

FIG. 16. Solution of Eq. (1.3) for $\alpha = 0, \beta = 3, \lambda = -1$.

ues $\alpha = 1$ and $\alpha = 2$ seem to demarcate distinct behavior of the solution. We assumed λ to be -1 in all cases. For $0 < \alpha < 1$, the solution grows to peak somewhere in the middle in a short time; it shows some small persisting wiggles when $\beta > 1$ (see Figs. 16–19). When $1 \leqslant \alpha < 2$, the solution grows and breaks at the front in a short time (see Figs. 11a, 12a, and 20–22). For the case $\alpha = 2$, the solution first decays (implying the dominance of nonlinear convection in the early stages) and then grows to break at the front (see Figs. 23 and 24). For $\alpha > 2$, the negative damping is too small and the solution continuously decays with time (see Figs. 15, 25–27).

It is of some interest to compare the special case (1.8), Lardner and Arya,[12] with the corresponding modified KdV equation (1.9). Both have the same form of convective and damping terms. Leibovich and Randall have numerically studied the initial value problem for (1.9). They treated a whole class of initial conditions that give rise to a variety of solitons, differing in number and amplifying or decaying depending on whether $\lambda < 0$ or $\lambda > 0$, respectively. They dis-

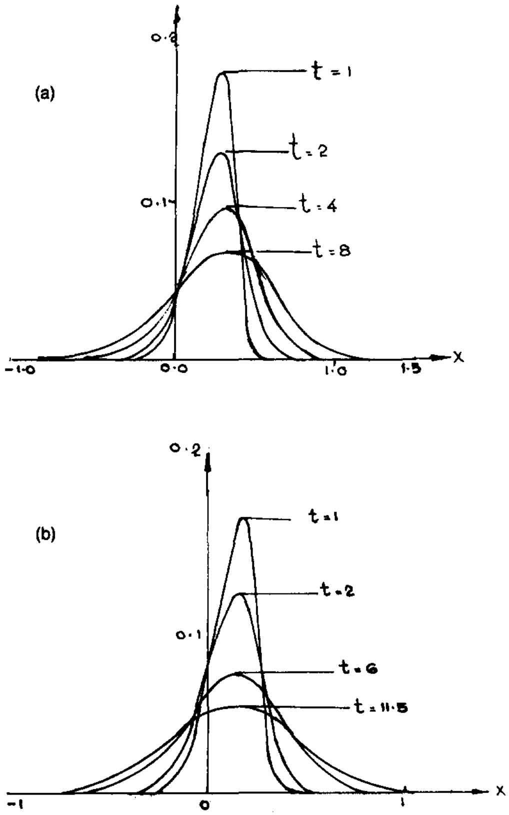

FIG. 15. Solution of Eq. (1.3) for $\lambda < 0$: (a) $\alpha = 4, \beta = 1.5$, (b) $\alpha = 5, \beta = 2$.

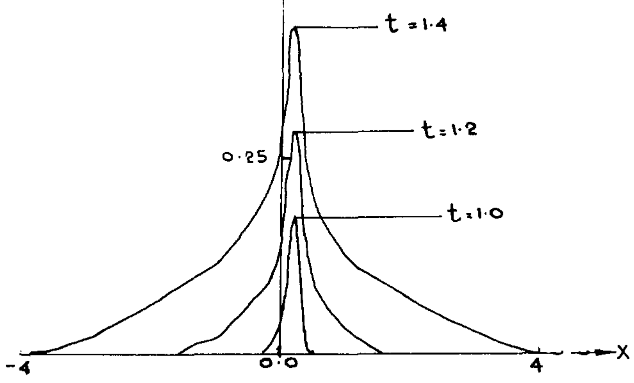

FIG. 17. Solution of Eq. (1.3) for $\alpha = 0.25, \beta = 3, \lambda = -1$.

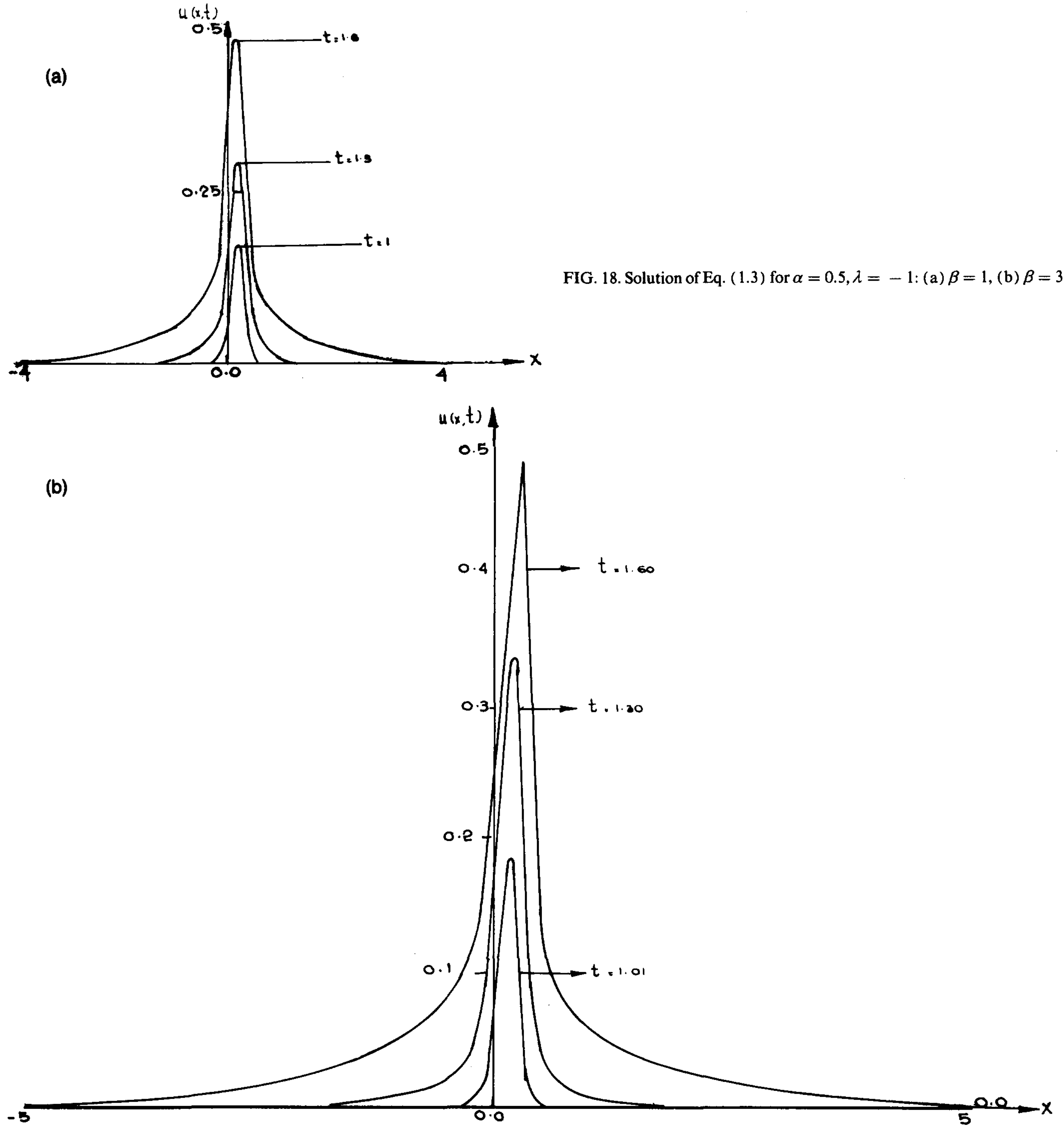

FIG. 18. Solution of Eq. (1.3) for $\alpha = 0.5, \lambda = -1$: (a) $\beta = 1$, (b) $\beta = 3$.

covered three integrals for the entities

$$\begin{aligned} A_r &= \int_{-\infty}^{\infty} u(x,t)dx, \\ E &= \frac{1}{2}\int_{-\infty}^{\infty} u^2(x,t)dx, \qquad (5.1) \\ \bar{x} &= \frac{1}{A_r}\int_{-\infty}^{\infty} xu(x,t)dx, \end{aligned}$$

which represent area under the wave, or its momentum, its energy, and its center of gravity, respectively. It is easily checked by direct integration of (1.9) and integration after multiplication by u and x, respectively, that A_r, E, and $\bar{x}$ satisfy the following relations:

$$\begin{aligned} A_r &= A_0 e^{-\lambda t}, \\ E &= E_0 e^{-2\lambda t}, \qquad (5.2) \\ \bar{x} - \bar{x}_0 &= (\mu E_0/\lambda A_0)(e^{-\lambda t_0} - e^{-\lambda t}). \end{aligned}$$

The subscript $_0$ refers to the value of the relevant quantity at

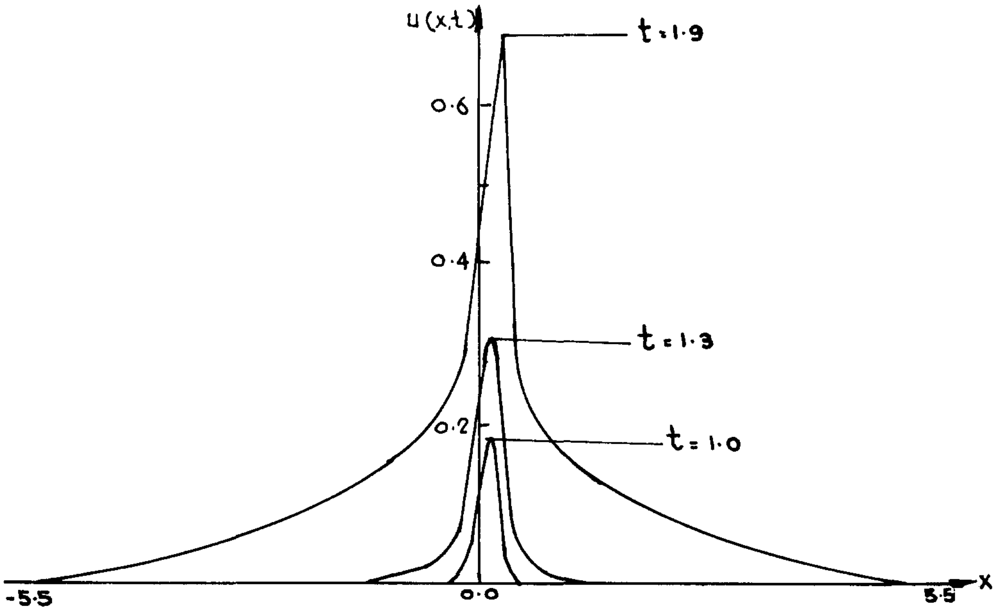

FIG. 19. Solution of Eq. (1.3) for $\alpha = \beta = 0.5, \lambda = -1$.

$t = t_0$. It is straightfoward to check that the integrals (5.2) with $\mu = -1$ exist for the modified Burgers equation (1.8) as well. The main features in the solitary wave study of Leibovich and Randall is the appearance of a trailing shelf and amplification or decay of the wave depending on the sign of λ. They also found a terminal similarity solution for (1.9) for each soliton in isolation. While this solution confirmed the major features (dominant soliton plus shelf), it was not a uniformly valid solution since it failed to satisfy the boundary condition at $x = \infty$. Equation (1.8) does not possess a self-similar solution. It shows amplification or decay of the initial profile depending on whether $\lambda < 0$ or $\lambda > 0$ (see Fig. 11). These numerical solutions satisfy the relations (5.2) (see Table X).

VI. CONCLUSIONS

We have studied the initial value problem for GBE (1.3) with the single hump type of initial conditions with a view to confirm that the self-similar form (2.5) indeed constitutes an intermediate asymptotic. This turns out to be the case for $1 < \alpha \leqslant 3$, $\lambda > 0$, in agreement with Murray's case (iii). It is remarkable that the self-similar form (2.5) of (1.3) yields the same asymptotic decay law for the wave as the characteristic method does for the inviscid form of (1.3). The structures of the two waves will be quite different. Indeed, the sharp shock of Murray[10] remains sharp for all time and there is neither spreading nor decay due to diffusion (which is absent in his model). We believe that Eq. (2.9) for the "reciprocal" function H, which is a special case of (1.5), is new. Equation (1.5), we postulate, should have an importance for the Burgers equation similar to the Painlevé equations for the KdV-type equations. As we mentioned earlier, Eq. (1.4) has the solution

$$\begin{aligned} u &= t^{-1/2\alpha} g(\eta) \\ &= (t/\delta)^{-1/2\alpha} G^{-1/\alpha}(\eta), \end{aligned} \tag{6.1}$$

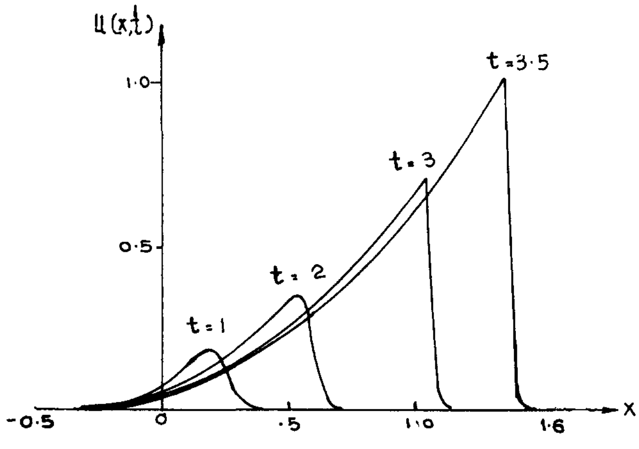

FIG. 20. Solution of Eq. (1.3) for $\alpha = 1, \beta = 0.5, \lambda = -1$.

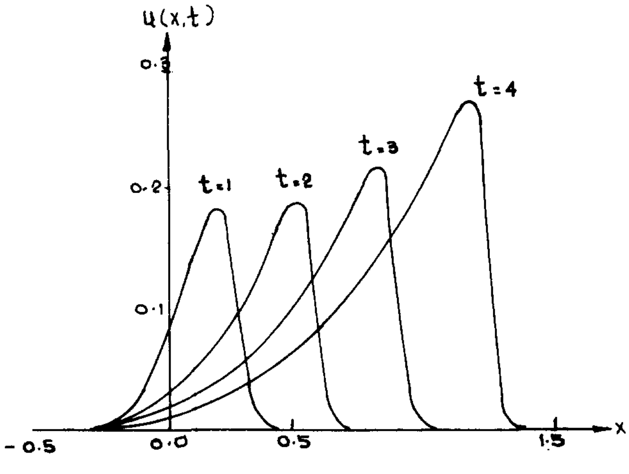

FIG. 21. Solution of Eq. (1.3) for $\alpha = 1.5, \beta = 0.5, \lambda = -1$.

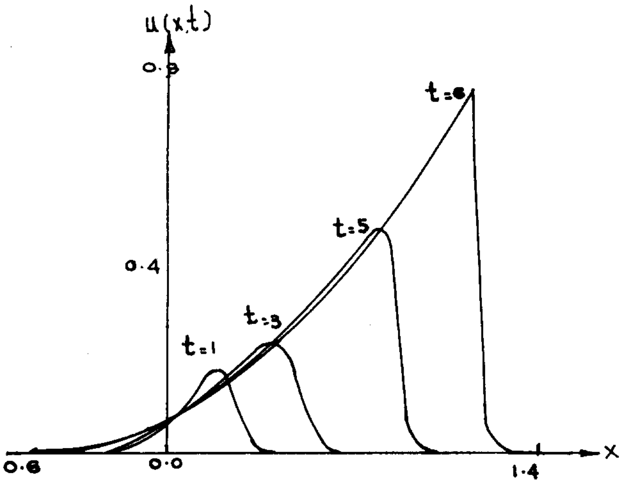

FIG. 22. Solution of Eq. (1.3) for $\alpha = 1.5, \beta = 1, \lambda = -1$.

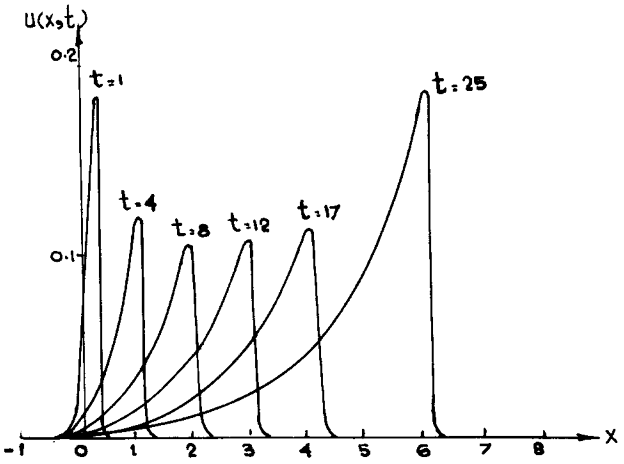

FIG. 24. Solution of Eq. (1.3) for $\alpha = 2, \beta = 0.5, \lambda = -1$.

where G is governed by

$$GG'' - ((\alpha + 1)/\alpha)G'^2 + 2\eta GG' - 2(1 - \alpha j)G^2 - 2^{3/2}G' = 0 \tag{6.2}$$

and

$$\eta = x(2\delta t)^{-1/2}.$$

This equation is again a special case of (1.5), and differs from (2.10) for the Burgers equation in merely having different numerical coefficients. Indeed, the Euler–Painlevé equation (2.13) is rather special and has only four terms on the left; even the GEPE for the Burgers equation has five terms. We found it more convenient to solve the connection problem (2.19)–(2.21) for (2.7) and then draw conclusions for (2.9). Equation (2.9) has been analyzed by us only to some extent; further analysis of this equation or (2.7) may

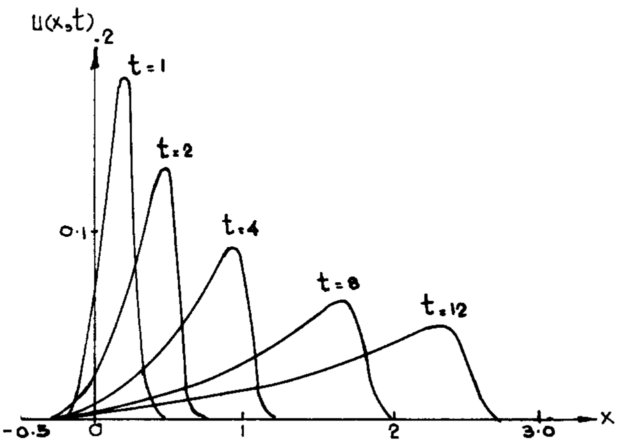

FIG. 25. Solution of Eq. (1.3) for $\alpha = 2.5, \beta = 0.5, \lambda = -1$.

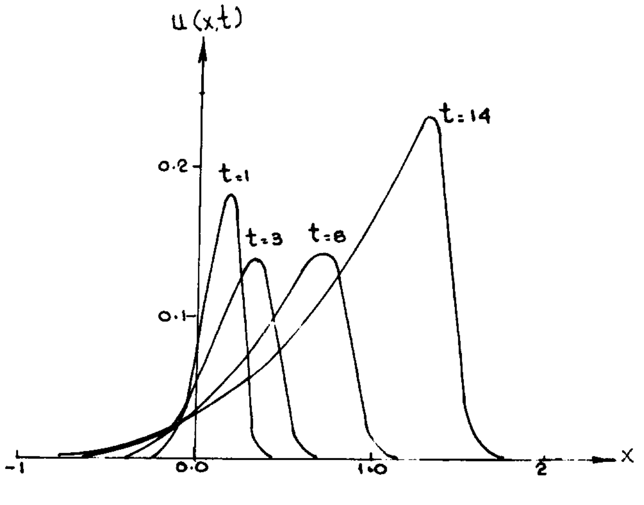

FIG. 23. Solution of Eq. (1.3) for $\alpha = 2, \beta = 1, \lambda = -1$.

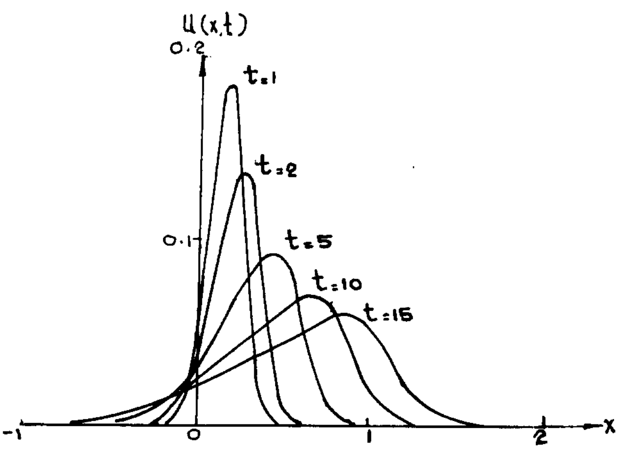

FIG. 26. Solution of Eq. (1.3) for $\alpha = 2.5, \beta = 1, \lambda = -1$.

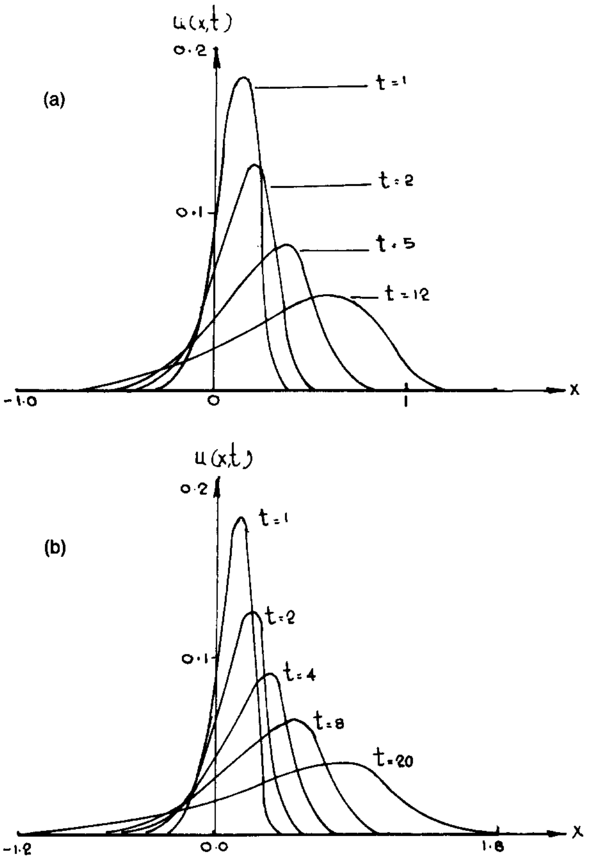

FIG. 27. Solution of Eq. (1.3) for $\beta = 1, \lambda = -1$: (a) $\alpha = 3$. (b) $\alpha = 4$.

TABLE X. Numerical and analytic values of the area A_r, energy E, and center of gravity $\bar{x}$ of the profile at various times for Eq. (1.8) ($t_0 = 1.6$, $A_0 = 0.1359$, $E_0 = 0.01897$, $\bar{x}_0 = 0.155\,37$) [see Eq. (5.2)].

	A_r		E		$\bar{x}$	
t	Numerical	Analytic	Numerical	Analytic	Numerical	Analytic
1.8	0.022 48	0.022 47	0.000 49	0.000 52	0.160 95	0.160 48
2.0	0.018 39	0.018 39	0.000 31	0.000 34	0.164 19	0.164 66
2.2	0.015 06	0.015 06	0.000 20	0.000 23	0.167 20	0.168 09
2.4	0.012 33	0.012 33	0.000 13	0.000 16	0.169 57	0.170 89
2.6	0.010 09	0.010 09	0.000 08	0.000 10	0.171 43	0.173 19
2.8	0.008 26	0.008 26	0.000 05	0.000 07	0.172 90	0.175 07
3.0	0.006 77	0.006 77	0.000 04	0.000 05	0.174 07	0.176 61
3.2	0.005 54	0.005 54	0.000 02	0.000 03	0.175 00	0.177 87
3.4	0.004 54	0.004 54	0.000 02	0.000 02	0.175 73	0.178 90

be taken up at some subsequent time. However, as far as the numerical solution of the initial value problem for (1.3) is concerned, our study is fairly complete. Some of the important conclusions are the following. Equation (1.3) has no self-similar solution except when $1 < \alpha \leqslant 3$, $\lambda > 0$. The solutions for $0 < \alpha < 1$, $\lambda > 0$ decay in a finite time and a finite distance; when $\lambda < 0$, they spread and spike in the middle in a relatively short time. For $\alpha > 3$, no self-similar solution of (2.5) forms an intermediate asymptotic. The wave profiles in this case, with $\lambda < 0$, display a long shelf at the left tail—a feature that has been noticed for the solitary wave evolving under the modified KdV equation with a damping term (see Fig. 14). As a nonlinear ordinary differential equation, Eq. (2.19) with (2.20) as the asymptotic conditions does have solutions for all $\alpha > 1$ and all λ, positive or negative. In Part II, we shall give a detailed study of the initial value problem for (1.4) and its self-similar solutions governed by (6.2) to fortify our claim regarding the importance of Eq. (1.5).

[1] G. B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974).
[2] R. Hirota, Phys. Rev. Lett. **27**, 1192 (1971).
[3] *Solitons*, edited by R. K. Bullough and P. J. Caudrey (Springer, New York, 1980).
[4] G. B. Whitham, IMA J. Appl. Math. **32**, 353 (1984).
[5] D. F. Parker, Proc.R. Soc. London Ser. A **369**, 409 (1980).
[6] M. J.Ablowitz, A. Ramani, and H. Segur, J. Math. Phys. **21**, 715 (1980).
[7] E. Hille, *Lectures on Ordinary Differential Equations* (Addison-Wesley, Reading, MA, 1969).
[8] K. M. Tamizhmani, M. Lakshmanan, and P. L. Sachdev (to be published).
[9] E. Kamke, *Differential gleichungen*: *Losungsmethoden und Losungen* (Akademische Verlagagesellschaft, Leipzig, 1943).
[10] J. D. Murray, SIAM J. Appl. Math. **19**, 135 (1970).
[11] J. D. Murray J. Fluid Mech. **44**, 315 (1970).
[12] R. W. Lardner and J. C. Arya, Acta Mech. **37**, 179 (1980).
[13] D. G. Crighton, Annu. Rev. Fluid Mech. **11**, 11 (1979).
[14] P. L. Sachdev, Proc.Natl. Acad. Sci. India (Professor P. L. Bhatnagar Commemoration Volume) **1979**, 73.
[15] J. Douglas and B. F. Jones, J. Soc. Ind. Appl. Math. **11**, 195 (1963).
[16] P. L. Sachdev and A. R. Seebass, J. Fluid Mech. **58**, 197 (1973).
[17] W. F. Ames, *Numerical Methods for Partial Differential Equations* (Academic, New York, 1977), 2nd ed.
[18] A. R. Mitchell and D. F. Griffiths, *The Finite Difference Method in Partial Differential Equations* (Wiley, New York, 1980).
[19] S. P. Hastings and J. B. McLeod, Arch. Ration. Mech. Anal. **73**, 31 (1980).
[20] J. W. Miles, Proc. R. Soc. London, Ser. A **361**, 277 (1978).
[21] G. I. Barenblatt and Y. B. Zeldovich, Annu. Rev. Fluid Mech. **4**, 285 (1972).
[22] S. Leibovich and J. D. Randall, Phys. Fluids **22**, 2289 (1979).
[23] M. J. Lighthill, in *Surveys in Mechanics*, edited by G. K. Batchelor and R. M. Davis (Cambridge U.P., Cambridge, England, 1956), pp. 250–351.
[24] E. Hopf, Commun. Pure Appl. Math. **3**, 201 (1950).
[25] O. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers* (McGraw-Hill, New York, 1978).
[26] B. Fornberg and G. B. Whitham, Philos. Trans. R. Soc. London **289**, 373 (1978).
[27] J. Gazdag, J. Comput. Phys. **13**, 100 (1973).
[28] P. L.Sachdev, V. G.Tikekar, and K. R. C. Nair (to appear in J. Fluid Mech.).

The four-dimensional conformal Kepler problem reduces to the three-dimensional Kepler problem with a centrifugal potential and Dirac's monopole field. Classical theory

Toshihiro Iwai and Yoshio Uwano
Department of Applied Mathematics and Physics, Kyoto University, Kyoto, Kyoto 606, Japan

(Received 27 August 1985; accepted for publication 12 February 1986)

The four-dimensional conformal Kepler problem is reduced by an S^1 action, when the associated momentum mapping takes nonzero fixed values. The reduced Hamiltonian system proves to be the three-dimensional Kepler problem along with a centrifugal potential and Dirac's monopole field. The negative-energy surface turns out to be diffeomorphic to $S^3 \times S^2$, on which the symmetry group SO(4) acts. Constants of motion of the reduced system are also obtained, which include the total angular momentum vector and a Runge–Lenz-like vector. The Kepler problem is thus generalized so as to admit the same symmetry group.

I. INTRODUCTION

Reduction of Hamiltonian systems has been investigated for years. Marsden and Weinstein gave a unified framework for reduction of symplectic manifolds.[1] When a Lie group acts symplectically on a symplectic manifold (M,ω) one can get a lower-dimensional symplectic manifold, called a reduced phase space, by using the momentum mapping for the Lie group action. A Hamiltonian system (M,ω,H), whose Hamiltonian H is invariant under the group action, can be reduced to a Hamiltonian system on the reduced phase space. If the original Hamiltonian system admits a symmetry group that is commutative with the group used for the reduction, the reduced Hamiltonian system admits the same symmetry group.

Iwai[2] defined a "conformal" Kepler problem to associate the four-dimensional harmonic oscillator to the three-dimensional Kepler problem. The conformal Kepler problem, which is closely related to the harmonic oscillator, was proved to be reduced by an S^1 action to the ordinary Kepler problem. Owing to this reduction, the Kepler problem becomes feasible to analyze globally. The reduction was carried out in the case where the momentum mapping of the S^1 action is set to take a fixed value zero.

A question now arises as to what reduced system comes out if the momentum mapping is set to take nonzero fixed values. The purpose of this paper is to answer the question. The materials of the present article is outlined as follows.

Section II treats a reduction of the symplectic manifold $(T^*\dot{\mathbf{R}}^4,d\theta)$. Here $\dot{\mathbf{R}}^4$, denoting $\mathbf{R}^4-\{0\}$, is endowed with a conformally flat metric defined in the Cartesian coordinates (x_j), $j=1,\dots,4$, by

$$ds_c^2 = 4r\sum_{j=1}^{4} dx_j^2, \quad \text{with } r=\sum_{j=1}^{4} x_j^2,$$

and $d\theta$ is the standard symplectic form on the cotangent bundle $T^*\dot{\mathbf{R}}^4$, which is expressed in the canonical coordinates (x_j,p_j) on $T^*\dot{\mathbf{R}}^4 \simeq \dot{\mathbf{R}}^4\times\mathbf{R}^4$ as $\Sigma dp_j\wedge dx_j$. An action of $\mathrm{U}(1)\simeq S^1$ is defined on $T^*\dot{\mathbf{R}}^4$. Since the momentum mapping Ψ: $T^*\dot{\mathbf{R}}^4\to \mathrm{u}(1)^*\simeq\mathbf{R}$ for this action is manifestly Ad*-equivariant,[1] the Ψ can be used for reduction of $(T^*\dot{\mathbf{R}}^4,d\theta)$. The reduced phase space $\Psi^{-1}(\mu)/U(1)$, $\mu\in\mathbf{R}$, will be shown to be diffeomorphic to $T^*\dot{\mathbf{R}}^3$, the cotangent bundle of $\dot{\mathbf{R}}^3=\mathbf{R}^3-\{0\}$, by using the Hopf mapping π: $\dot{\mathbf{R}}^4\to\dot{\mathbf{R}}^3$. This reduction is known also as the Kustaanheimo–Stiefel transformation,[3] and compactly reviewed in Ref. 4. However, the reduced phase space is not symplectomorphic to the cotangent bundle $T^*\dot{\mathbf{R}}^3$ equipped with the standard symplectic form. Let $d\theta'$ denote the standard symplectic form on $T^*\dot{\mathbf{R}}^3$, and Ω_μ a two-form describing Dirac's monopole field on $\dot{\mathbf{R}}^3$. Then, given a symplectic form σ_μ defined as $d\theta'+\Omega_\mu$, the $T^*\dot{\mathbf{R}}^3$ becomes symplectomorphic with the reduced phase space. Thus the reduced phase space is identified with $(T^*\dot{\mathbf{R}}^3,\sigma_\mu)$. It is to be noted here that through the reduction the Euclidean metric is induced on $\dot{\mathbf{R}}^3$ from the metric ds_c^2 on $\dot{\mathbf{R}}^4$. It should be also noted that π: $\dot{\mathbf{R}}^4\to\dot{\mathbf{R}}^3$ is a principal U(1) bundle, so that the reduction to be performed in this section gives an example of Kummer's work[5] about the reduction of cotangent bundles of principal fiber bundles.

Though this section is an application of Kummer's work, the technique used is of great importance in treating the symmetry in the following sections.

Section III is concerned with reduction of the conformal Kepler problem by the U(1) action defined in Sec. II. The conformal Kepler problem is a Hamiltonian system on the symplectic manifold $(T^*\dot{\mathbf{R}}^4,d\theta)$ endowed with the Hamiltonian

$$H=\frac{1}{2}\left(\frac{1}{4r}\sum_{j=1}^{4}p_j^2\right)-\frac{k}{r}.$$

It is to be noted here that the first term in the right-hand side is the kinetic energy with respect to ds_c^2. Since the H is invariant under the U(1) action, it can be reduced to a certain Hamiltonian H_μ on $T^*\dot{\mathbf{R}}^3$. As is pointed out above, the Euclidean metric is induced on the $\dot{\mathbf{R}}^3$. Let $(\tilde{x}_j,\tilde{p}_j)$ be the Cartesian coordinates on $T^*\dot{\mathbf{R}}^3\simeq\dot{\mathbf{R}}^3\times\mathbf{R}^3$. Then the reduced Hamiltonian H_μ proves to have the form

$$H_\mu=\frac{1}{2}\sum_{j=1}^{3}\tilde{p}_j^2+\frac{\mu^2}{2r^2}-\frac{k}{r}, \quad r^2=\sum_{j=1}^{3}\tilde{x}_j^2.$$

The reduced system $(T^*\dot{\mathbf{R}}^3,\sigma_\mu,\mathrm{H}_\mu)$ coincides with the Hamiltonian system that MacIntosh and Cisneros[6] treated. The equation of motion for $(T^*\dot{\mathbf{R}}^3,\sigma_\mu,\mathrm{H}_\mu)$ indeed describes

0022-2488/86/061523-07$02.50

the Kepler motion in the presence of a centrifugal potential and Dirac's monopole field. This system can be regarded as a variation of the ordinary Kepler problem, because it will be shown in Secs. V and VI to admit the same symmetry group as the ordinary Kepler problem does. In view of this, the reduced system will be referred to as the MIC–Kepler problem, where MIC is short for McIntosh and Cisneros. It is worth mentioning also that the reduction in this article is closely related with the reduction that Satzer[7] carried out for the planar three-body problem. In this article, the metric ds_c^2 is utilized to obtain the standard kinetic energy term in H_μ. In the Kummer's work,[5] a change of time parameter was carried out to obtain the same Hamiltonian.

Section IV shows that the regularized energy-momentum manifold $S^3\times S^3$ of the conformal Kepler problem is reduced to the negative-energy surface $S^3\times S^2$ of the MIC–Kepler problem.

In Sec. V, a symmetry group of the MIC–Kepler problem is studied. Combined with the result in Ref. 8, the reduction in this paper shows that the global symmetry group SO(4) acts on every negative-energy surface of the MIC–Kepler problem.

Section VI deals with first integrals for the MIC–Kepler problem, which form clearly the Lie algebra of SO(4) under the Poisson bracket. They can be obtained from the U(1)-invariant first integrals of the conformal Kepler problem. It is shown that the first integrals for the MIC–Kepler problem consist of the vector constants obtained in Ref. 6; one is the total angular momentum, and the other a Runge–Lenz-like vector.

A quantum version of this article will appear in the next paper.[9]

II. REDUCTION OF A PHASE SPACE ($T^*\dot{\mathbb{R}}^4, d\theta$)

Let (x_j), $j=1,2,3,4$, be the Cartesian coordinates of $\mathbb{R}^4$, and $\dot{\mathbb{R}}^4 = \mathbb{R}^4 - \{0\}$. We define a conformally flat metric ds_c^2 on $\dot{\mathbb{R}}^4$ by

$$ds_c^2 = 4r \sum_{j=1}^{4} dx_j^2, \quad \text{with } r = \sum_{j=1}^{4} x_j^2. \tag{2.1}$$

Consider the cotangent bundle $T^*\dot{\mathbb{R}}^4$ of $\dot{\mathbb{R}}^4$. Let (v_j) be any tangent vector at $x\in\dot{\mathbb{R}}^4$. Then a cotangent vector $p=(p_j)$ at x is assigned by $p_j = 4rv_j$, $j=1,2,3,4$, on account of (2.1). Since $T^*\dot{\mathbb{R}}^4$ is identified with $\dot{\mathbb{R}}^4\times\mathbb{R}^4$, every point of $T^*\dot{\mathbb{R}}^4$ can be expressed as a pair of column vectors (x,p). The canonical symplectic form on $T^*\dot{\mathbb{R}}^4$ is given by $d\theta$ with

$$\theta = \sum_{j=1}^{4} p_j\, dx_j. \tag{2.2}$$

We define a symplectic U(1) action Φ_t on $T^*\dot{\mathbb{R}}^4$ by

$$\Phi_t(x,p) = (T(t)x, T(t)p) \quad (t\in\mathbb{R}), \tag{2.3}$$

where[10]

$$T(t) = \begin{pmatrix} R(t) & \\ & R(t) \end{pmatrix},$$
$$\text{with } R(t) = \begin{pmatrix} \cos(t/2) & -\sin(t/2) \\ \sin(t/2) & \cos(t/2) \end{pmatrix}. \tag{2.4}$$

A simple calculation shows that the U(1) action leaves θ invariant, so that the group $\mathrm{U}(1)\simeq S^1$ acts symplectically on $(T^*\dot{\mathbb{R}}^4, d\theta)$.

In order to reduce $(T^*\dot{\mathbb{R}}^4, d\theta)$ by the U(1) action, we look for the momentum mapping of the U(1) action. According to Abraham and Marsden,[11] the momentum mapping Ψ of $T^*\dot{\mathbb{R}}^4$ to $\mathrm{u}(1)^*$, the dual space to the Lie algebra u(1) of U(1), is obtained by

$$\Psi(x,p)\cdot\xi = \theta_{(x,p)}(\xi_M), \tag{2.5}$$

where ξ is an element of u(1), and ξ_M, $M=T^*\dot{\mathbb{R}}^4$, is the infinitesimal generator of Φ_t. The ξ_M is easy to get from (2.3) and (2.4). Since $\mathrm{u}(1)^*\simeq\mathbb{R}$, the $\Psi(x,p)$ is viewed as real valued, and obtained from (2.5) in the form

$$\Psi(x,p) = \tfrac{1}{2}(-x_2p_1 + x_1p_2 - x_4p_3 + x_3p_4). \tag{2.6}$$

It is easy to see that Ψ is invariant under the U(1) action. This fact also means that Ψ is Ad*-equivariant, because U(1) is Abelian.

Consider the level manifold for a fixed real number μ; $\Psi^{-1}(\mu) := \{(x,p)\in T^*\dot{\mathbb{R}}^4;\ \Psi(x,p)=\mu\}$. As $\mu\neq 0$ is a regular value of Ψ, $\Psi^{-1}(\mu)$ is indeed a submanifold of $T^*\dot{\mathbb{R}}^4$.

We show the following.

Lemma 2.1: $\Psi^{-1}(\mu)$ is diffeomorphic to $\dot{\mathbb{R}}^4\times\mathbb{R}^3$.

Proof: Let $\langle\ ,\ \rangle$ denote the standard inner product in $\mathbb{R}^4$. We define a basis $\{s_j(x)\}_{j=0,1,2,3}$, orthogonal with respect to $\langle\ ,\ \rangle$, in each cotangent space $T_x^*\dot{\mathbb{R}}^4$ by

$$\begin{aligned} s_0(x) &= 2(-x_2, x_1, -x_4, x_3)^T, \\ s_1(x) &= 2(x_3, x_4, x_1, x_2)^T, \\ s_2(x) &= 2(-x_4, x_3, x_2, -x_1)^T, \\ s_3(x) &= 2(x_1, x_2, -x_3, -x_4)^T, \end{aligned} \tag{2.7}$$

where the superscript T indicates the transpose. Then, the equation $\Psi(x,p)=\mu$ is written as

$$\langle s_0(x), p\rangle = 4\mu. \tag{2.8}$$

This equation implies that p can be expressed for each $x\in\dot{\mathbb{R}}^4$ in the form

$$p = \frac{\mu}{r} s_0(x) + \sum_{j=1}^{3} a_j s_j(x), \quad a_j\in\mathbb{R}, \quad j=1,2,3. \tag{2.9}$$

Thus we can define a diffeomorphism of $\Psi^{-1}(\mu)$ to $\dot{\mathbb{R}}^4\times\mathbb{R}^3$ by $(x,p)\in\Psi^{-1}(\mu)\to(x,a)$, where $a=(a_j)\in\mathbb{R}^3$. This completes the proof.

Now, since U(1) leaves Ψ invariant, it acts on $\Psi^{-1}(\mu)$. We are looking into how the action is expressed. Let (x,p) and (x',p') be in $\Psi^{-1}(\mu)$, and assume that $(x',p') = \Phi_t(x,p)$. If p is expressed as (2.9), p' is then written in the form

$$p' = \frac{\mu}{r} T(t) s_0(x) + \sum_{j=1}^{3} a_j T(t) s_j(x). \tag{2.10}$$

One can easily verify that

$$r = \sum_{j=1}^{4} x_j^2 = \sum_{j=1}^{4} x_j'^2$$

and

$$T(t) s_j(x) = s_j(x'), \quad j=0,1,2,3.$$

Hence Eq. (2.10) goes over into

$$p' = \frac{\mu}{r} s_0(x') + \sum_{j=1}^{3} a_j s_j(x'). \tag{2.11}$$

Taking the expressions (2.9) and (2.11) into account, we have the following.

Lemma 2.2: The U(1) action on $\Psi^{-1}(\mu) \simeq \dot{\mathbf{R}}^4 \times \mathbf{R}^3$ is given by $(x,a) \to (T(t)x, a)$, where a is defined by (2.9).

Obviously, the U(1) action is free and proper, so that the quotient space $\Psi^{-1}(\mu)/\mathrm{U}(1)$ becomes a manifold.

We are now to show that $\Psi^{-1}(\mu)/\mathrm{U}(1)$ is diffeomorphic to $T^*\dot{\mathbf{R}}^3$, the cotangent bundle of $\dot{\mathbf{R}}^3$. To this end, we define the mapping π: $\dot{\mathbf{R}}^4 \to \dot{\mathbf{R}}^3$ by

$$\begin{aligned} \tilde{x}_1 &= 2(x_1 x_3 + x_2 x_4), \\ \tilde{x}_2 &= 2(-x_1 x_4 + x_2 x_3), \\ \tilde{x}_3 &= x_1^2 + x_2^2 - x_3^2 - x_4^2, \end{aligned} \tag{2.12}$$

which is an extension of the Hopf mapping $S^3 \to S^2$. It is then clear that π: $\dot{\mathbf{R}}^4 \to \dot{\mathbf{R}}^3$ is a principal U(1) bundle, where the group action is the same as we have defined in (2.3). This and Lemma 2.2 are put together to show that $\Psi^{-1}(\mu)/\mathrm{U}(1)$ is diffeomorphic to $\dot{\mathbf{R}}^3 \times \mathbf{R}^3$, and hence to $T^*\dot{\mathbf{R}}^3$.

We wish to make further investigation into the diffeomorphism of $\Psi^{-1}(\mu)/\mathrm{U}(1)$ to $T^*\dot{\mathbf{R}}^3$. Let $d\pi_x^*$: $T^*_{\pi(x)}\dot{\mathbf{R}}^3 \to T_x^*\dot{\mathbf{R}}^4$ be the dual mapping to the tangent mapping $d\pi_x$: $T_x\dot{\mathbf{R}}^4 \to T_{\pi(x)}\dot{\mathbf{R}}^3$. Since for any tangent vector $v = (v_j)$ at $x \in \dot{\mathbf{R}}^4$, $d\pi_x(v)$ is given by

$$d\pi_x(v) = 2\begin{pmatrix} x_3 & x_4 & x_1 & x_2 \\ -x_4 & x_3 & x_2 & -x_1 \\ x_1 & x_2 & -x_3 & -x_4 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}, \tag{2.13}$$

for any cotangent vector $\tilde{p} = (\tilde{p}_j)$ at $\pi(x) \in \dot{\mathbf{R}}^3$, $d\pi_x^*(\tilde{p})$ is given by definition as

$$d\pi_x^*(\tilde{p}) = 2\begin{pmatrix} x_3 & -x_4 & x_1 \\ x_4 & x_3 & x_2 \\ x_1 & x_2 & -x_3 \\ x_2 & -x_1 & -x_4 \end{pmatrix}\begin{pmatrix} \tilde{p}_1 \\ \tilde{p}_2 \\ \tilde{p}_3 \end{pmatrix}. \tag{2.14}$$

Put another way, $d\pi_x^*$ is of the form

$$d\pi_x^*(\tilde{p}) = \sum_{j=1}^{3} \tilde{p}_j s_j(x). \tag{2.15}$$

It follows from (2.15) and (2.9) that the range of $d\pi_x^*$ is the subspace of $T_x^*\dot{\mathbf{R}}^4$ given by $\Psi^{-1}(0) \cap T_x^*\dot{\mathbf{R}}^4$, and therefore $d\pi_x^*$ has the inverse when restricted to $\Psi^{-1}(0) \cap T_x^*\dot{\mathbf{R}}^4$. Further, we see that $\Psi^{-1}(0)$ consists of all the elements $(x, d\pi_x^*(\tilde{p}))$ with $x \in \dot{\mathbf{R}}^4$ and $\tilde{p} \in T^*_{\pi(x)}\dot{\mathbf{R}}^3$. Accordingly, from Lemma 2.2, the quotient space $\Psi^{-1}(0)/\mathrm{U}(1)$ is diffeomorphic to $T^*\dot{\mathbf{R}}^3$. We can then denote the natural projection $\Psi^{-1}(0) \to \Psi^{-1}(0)/\mathrm{U}(1)$ by

$$(d\pi^*)^{-1}: (x,p) \to (\pi(x), (d\pi_x^*)^{-1}(p)) = (\tilde{x}, \tilde{p}). \tag{2.16}$$

It is now an easy matter to obtain the diffeomorphism of $\Psi^{-1}(\mu)/\mathrm{U}(1)$ to $T^*\dot{\mathbf{R}}^3$. In effect, a diffeomorphism ν_μ: $\Psi^{-1}(\mu) \to \Psi^{-1}(0)$ can be defined by

$$\nu_\mu(x,p) = (x, p - (\mu/r)s_0(x)), \tag{2.17}$$

and thereby the composition

$$\pi_\mu(x,p) = (d\pi^*)^{-1} \circ \nu_\mu(x,p) \tag{2.18}$$

provides a principal U(1) bundle $\Psi^{-1}(\mu) \to T^*\dot{\mathbf{R}}^3$, as is known from the following lemma.

Lemma 2.3: Let (x,p) and (x',p') be points of $\Psi^{-1}(\mu)$. Then the relation

$$\pi_\mu(x,p) = \pi_\mu(x',p') \tag{2.19}$$

holds if and only if both (x,p) and (x',p') are on a U(1) orbit.

Proof: For (x,p), $(x',p') \in \Psi^{-1}(\mu)$, the vectors p and p' are expressed as

$$p = \frac{\mu}{r} s_0(x) + \sum_{j=1}^{3} a_j s_j(x),$$

$$p' = \frac{\mu}{r} s_0(x') + \sum_{j=1}^{3} a_j' s_j(x'),$$

respectively. Then from (2.15)–(2.18), we have

$$\pi_\mu(x,p) = (\pi(x), a), \quad \pi_\mu(x',p') = (\pi(x'), a'),$$

where $a = (a_j)$ and $a' = (a_j')$ are column vectors. Therefore, if the relation (2.19) holds, one has $\pi(x) = \pi(x')$ and $a = a'$, and hence, for a certain t,

$$x' = T(t)x, \quad p' = \frac{\mu}{r} s_0(x') + \sum_{j=1}^{3} a_j s_j(x').$$

Thus from (2.11) one has $\Phi_t(x,p) = (x',p')$. The converse is easy to check. This completes the proof.

Thus we have proved the following lemma.

Lemma 2.4: $T^*\dot{\mathbf{R}}^3$ is diffeomorphic with $\Psi^{-1}(\mu)/\mathrm{U}(1)$.

We now proceed to a symplectic form σ_μ induced on the reduced phase space $T^*\dot{\mathbf{R}}^3$. According to Marsden and Weinstein,[1] σ_μ is determined by the relation

$$\pi_\mu^* \sigma_\mu = i_\mu^* \, d\theta, \tag{2.20}$$

where i_μ: $\Psi^{-1}(\mu) \to T^*\dot{\mathbf{R}}^4$ is the inclusion and the superscript asterisk in (2.20) indicates the pullback. Writing out $i_\mu^* \, d\theta$ by the help of $p = \mu s_0(x)/r + \Sigma \tilde{p}_j s_j(x)$, and collecting those terms that amount to $\pi_\mu^*(d\tilde{x}_j)$, we eventually obtain

$$\sigma_\mu = d\theta' + \Omega_\mu, \tag{2.21a}$$

where

$$d\theta' = \sum_{j=1}^{3} d\tilde{p}_j \wedge d\tilde{x}_j, \tag{2.21b}$$

$$\begin{aligned} \Omega_\mu = {} & -\mu r^{-3}(\tilde{x}_1 \, d\tilde{x}_2 \wedge d\tilde{x}_3 + \tilde{x}_2 \, d\tilde{x}_3 \wedge d\tilde{x}_1 \\ & + \tilde{x}_3 \, d\tilde{x}_1 \wedge d\tilde{x}_2), \end{aligned} \tag{2.21c}$$

and r is written as

$$r^2 = \sum_{j=1}^{3} \tilde{x}_j^2.$$

We note that $d\theta'$ is the canonical symplectic form of $T^*\dot{\mathbf{R}}^3$ and Ω_μ, viewed as a form on $T^*\dot{\mathbf{R}}^3$, is Dirac's monopole field of strength $-\mu$, which is turned off when the angular momentum Ψ is zero.

Thus we have shown the following theorem.

Theorem 2.5: The reduced phase space of $(T^*\dot{\mathbf{R}}^4, d\theta)$ is symplectomorphic to $(T^*\dot{\mathbb{R}}^3,\sigma_\mu)$, where σ_μ is given by (2.21).

In conclusion, we show that a metric ds_0^2 induced on $\dot{\mathbb{R}}^3$ from ds_c^2 is the Euclidean one. The metric ds_0^2 is indeed defined because ds_c^2 is invariant under the U(1) action. Let $(ds_c^2)^\#_x$ and $(ds_0^2)^\#_{\pi(x)}$ denote the inner products on the cotangent spaces $T^*_x\dot{\mathbb{R}}^4$ and $T^*_{\pi(x)}\dot{\mathbb{R}}^3$, respectively. Then one has the defining relation, for $\tilde{p},\tilde{p}'\in T^*_{\pi(x)}\dot{\mathbb{R}}^3$,

$$(ds_0^2)^\#_{\pi(x)}(\tilde{p},\tilde{p}') = (ds_c^2)^\#_x(d\pi_x^*(\tilde{p}),d\pi_x^*(\tilde{p}')). \quad (2.22)$$

We notice here that the right-hand side of (2.22) is independent of a choice of x' such that $\pi(x)=\pi(x')$ because of the invariance of ds_c^2 under the U(1) action.

To show that ds_0^2 is the Euclidean metric, we have only to point out that the basis $\{s_j(x)\}$ introduced in (2.7) is an orthonormal system with respect to $(ds_c^2)^\#$;

$$(ds_c^2)^\#_x(s_j(x),s_k(x)) = \delta_{jk}, \quad j,k=0,1,2,3. \quad (2.23)$$

Then from (2.15), (2.22), and (2.23) it follows that

$$(ds_0^2)^\#_{\pi(x)}(\tilde{p},\tilde{p}') = \sum_{j=1}^{3}\tilde{p}_j\tilde{p}_j', \quad (2.24)$$

which proves our assertion.

III. REDUCTION OF THE CONFORMAL KEPLER PROBLEM

The conformal Kepler problem is a triple $(T^*\dot{\mathbb{R}}^4,d\theta,H)$, where the Hamiltonian H is defined by

$$H=\frac{1}{2}\left(\frac{1}{4r}\sum_{j=1}^{4}p_j^2\right)-\frac{k}{r} \quad (k=\text{a positive const.}). \quad (3.1)$$

By using the notation $(ds_c^2)^\#$ introduced in the last section, the kinetic energy is expressed in the form

$$\tfrac{1}{2}(ds_c^2)^\#(p,p). \quad (3.2)$$

Furthermore, the distance between the origin of $\mathbb{R}^4$ and a point $x\in\dot{\mathbb{R}}^4$ is proportional to r, so that the potential term $-k/r$ is of Kepler type. This is why we call the Hamiltonian system the conformal Kepler problem.

It is easy to see that H is invariant under the U(1) action Φ_t: that is, $H=H\circ\Phi_t$. Consequently, H can be reduced by the U(1) action to a Hamiltonian defined on the reduced phase space $(T^*\dot{\mathbb{R}}^3,\sigma_\mu)$; the reduced Hamiltonian H_μ is determined[11] by

$$H\circ i_\mu = H_\mu\circ\pi_\mu. \quad (3.3)$$

Substituting $\mu s_0(x)/r+\Sigma\tilde{p}_j s_j(x)$ for p in (3.2), and using (2.23), we find from (3.2) the reduced kinetic energy and therefore the reduced Hamiltonian H_μ in the form

$$H_\mu=\frac{1}{2}\sum_{j=1}^{3}\tilde{p}_j^2+\frac{\mu^2}{2r^2}-\frac{k}{r}. \quad (3.4)$$

We note again that $r^2=\Sigma\tilde{x}_j^2$. The reduced system $(T^*\dot{\mathbb{R}}^3,\sigma_\mu,H_\mu)$ can be interpreted as follows. Recalling that the Euclidean metric ds_0^2 is induced on $\dot{\mathbb{R}}^3$, we see that the H_μ is a Hamiltonian for the usual Kepler problem plus the centrifugal potential $\mu^2/2r^2$. If $\mu=0$, H_μ becomes the Hamiltonian for the Kepler problem. As we have seen in Sec. II that the symplectic form σ_μ contains Dirac's monopole field, we now understand that the reduced system describes a Kepler motion on which both a centrifugal force and a magnetic force due to Dirac's field are put on. To see this in detail, we consider the Hamiltonian flow of the reduced system. The Hamiltonian vector field X_μ for H_μ is determined by

$$-dH_\mu = X_\mu \lrcorner\, \sigma_\mu, \quad (3.5)$$

where $\lrcorner$ indicates the interior product. Hence the equation of motion

$$\frac{d}{dt}(\tilde{x},\tilde{p}) = X_\mu(\tilde{x},\tilde{p})$$

can be put, after a calculation, into the form

$$\frac{d^2\tilde{x}}{dt^2}=\frac{d\tilde{x}}{dt}\times\left(-\frac{\mu}{r^3}\tilde{x}\right)-\text{grad}\left(\frac{\mu^2}{2r^2}-\frac{k}{r}\right), \quad (3.6)$$

where × denotes the vector product operation.

Theorem 3.1: The reduced Hamiltonian system $(T^*\dot{\mathbb{R}}^3,\sigma_\mu,H_\mu)$ of the conformal Kepler problem describes motions of a charged particle in the presence of Dirac's monopole field $B_\mu=-\mu\tilde{x}/r^3$ of strength $-\mu$ and a Newtonian potential $-k/r$ plus a centrifugal potential $\mu^2/2r^2$. This Hamiltonian system will be referred to as the MIC–Kepler problem.

Remarks: The velocity of light, the particle charge and the mass of the particle are all set at unity in Theorem 3.1. The reduced system $(T^*\dot{\mathbb{R}}^3,\sigma_\mu,H_\mu)$ is the very one that MacIntosh and Cisneros[6] treated. However, they described the system by using a vector potential. The fact that the vector potential for Dirac's field B_μ can be defined only locally prevented them from treating the system globally. Further, they adopted the Euclidean metric in reducing the kinetic energy on $T^*\dot{\mathbb{R}}^4$, and consequently came to an excessive factor in the angular momentum term of the kinetic energy on $T^*\dot{\mathbb{R}}^3$.

The result in this section is deeply concerned with the Kummer–Satzer work.[5,7] They started with a Hamiltonian having the kinetic energy term related to the standard flat metric in $\mathbb{R}^4$, and performed the reduction of the Hamiltonian followed by a change of the time parameter to get a reduced Hamiltonian similar to ours. However, because of the use of the metric ds_c^2, our reduction does not need the excessive procedure of the parameter change.

In the following sections, we study the symmetry aspect of the reduced Hamiltonian system. As a result, the system is regarded as a generalization of the Kepler problem.

IV. NEGATIVE-ENERGY SURFACES FOR THE REDUCED SYSTEM

In this section, we consider reduction of energy-momentum manifolds to obtain energy surfaces for the reduced system. We denote by $M_{\lambda,\mu}$ the energy-momentum manifold that is defined as the intersection of the energy surface $H=-\lambda^2/8$, λ being a positive constant, and the level submanifold $\Psi^{-1}(\mu)$. We remark that in Ref. 2 the regularized energy surface, denoted by $\overline{H}=-\lambda^2/8$, was considered in order to treat the flows going out of the domain $T^*\dot{\mathbb{R}}^4$. However, in the present case we do not have to treat the regularized energy surface, because the flows do not go out of $T^*\dot{\mathbb{R}}^4$ on account of $\mu\neq 0$.

As in Ref. 2, we consider a Hamiltonian system $(T^*\dot{\mathbb{R}}^4, d\theta, K)$, where K is given by

$$K=\frac{1}{2}\sum_{j=1}^{4}p_j^2+\frac{\lambda^2}{2}\sum_{j=1}^{4}x_j^2. \tag{4.1}$$

We treat (x,p) as if they were Cartesian coordinates in $\mathbb{R}^4\times\mathbb{R}^4\supset T^*\dot{\mathbb{R}}^4$. Then the Hamiltonian system (4.1) is thought of as the harmonic oscillator. From (3.1) and (4.1) we obtain the relation between H and K,

$$4r(H+\lambda^2/8)=K-4k. \tag{4.2}$$

It follows from this that the energy surface $\overline{H}=-\lambda^2/8$ coincides with the energy surface $K=4k$. Thus the energy momentum manifold $M_{\lambda,\mu}$ is given by

$$M_{\lambda,\mu}=\{(x,p)\in T^*\dot{\mathbb{R}}^4;\ K(x,p)=4k \text{ and } \Psi(x,p)=\mu\}. \tag{4.3}$$

It is of great use to introduce the complex variables w_j, $j=1,2,3,4$, by

$$\begin{aligned} w_1&=(\lambda x_1-p_2)+i(\lambda x_2+p_1),\\ w_2&=(\lambda x_3-p_4)+i(\lambda x_4+p_3),\\ w_3&=(\lambda x_1+p_2)+i(\lambda x_2-p_1),\\ w_4&=(\lambda x_3+p_4)+i(\lambda x_4-p_3). \end{aligned} \tag{4.4}$$

Hence $\mathbb{R}^4\times\mathbb{R}^4$ is identified with $\mathbb{C}^4$. With (4.3) and (4.4), a simple calculation shows that $M_{\lambda,\mu}$ is defined by

$$|w_1|^2+|w_2|^2=4(2k-\lambda\mu), \tag{4.5a}$$

$$|w_3|^2+|w_4|^2=4(2k+\lambda\mu). \tag{4.5b}$$

In order that $M_{\lambda,\mu}$ exists, λ and μ must satisfy

$$\lambda\,|\mu|\leqslant 2k. \tag{4.6}$$

In case of $\lambda\,|\mu|<2k$, the conditions (4.5) imply that $M_{\lambda,\mu}$ is diffeomorphic to $S^3\times S^3$. On the other hand, if $\lambda\,|\mu|=2k$, the conditions (4.5) define S^3, since either of (4.5a) and (4.5b) defines a single point $\{0\}$.

Theorem 4.1: Under the condition $\lambda\,|\mu|\leqslant 2k$, the energy-momentum manifold $M_{\lambda,\mu}$ is diffeomorphic to either of the following:

(a) $S^3\times S^3$ $(\lambda\,|\mu|<2k)$,

(b) S^3 $(\lambda\,|\mu|=2k)$.

Now, we proceed to the reduction of $M_{\lambda,\mu}$. Let $(\tilde{x},\tilde{p})=\pi_\mu(x,p)$ for $(x,p)\in M_{\lambda,\mu}$. Then from (3.3), one has $H_\mu(\tilde{x},\tilde{p})=H\circ i_\mu(x,p)=-\lambda^2/8$. Owing to the fact that π_μ is a projection from $\Psi^{-1}(\mu)$ to $T^*\dot{\mathbb{R}}^3$, $M_{\lambda,\mu}$ is mapped onto the energy surface, $H_\mu=-\lambda^2/8$, of the reduced Hamiltonian system $(T^*\dot{\mathbb{R}}^3,\sigma_\mu,H_\mu)$.

We will examine the topology of the energy surface $H_\mu=-\lambda^2/8$ in what follows. To do so, we point out that the U(1) action on $\mathbb{C}^4$, the w space, takes the matrix form

$$\begin{pmatrix} e^{it/2}I_2 & \\ & e^{-it/2}I_2 \end{pmatrix}, \tag{4.7}$$

where I_2 is the 2×2 unit matrix. This expression together with (4.5) shows that $M_{\lambda,\mu}/\mathrm{U}(1)$ is diffeomorphic to $S^3\times S^3/\mathrm{U}(1)\simeq S^3\times S^2$ (see Ref. 8), if $\lambda\,|\mu|<2k$.

In the case of $\lambda\,|\mu|=2k$, $M_{\lambda,\mu}$ degenerates to S^3, so that we have $S^3/\mathrm{U}(1)\simeq S^2$. Thus we are led to the following theorem.

Theorem 4.2: By the U(1) action, the energy-momentum manifold $M_{\lambda,\mu}$ is reduced to the energy surface $H_\mu=-\lambda^2/8$ of the reduced Hamiltonian system, which is diffeomorphic to either of the following:

(a) $S^3\times S^2$ $(\lambda\,|\mu|<2k)$,

(b) S^2 $(\lambda\,|\mu|=2k)$.

In conclusion we mention what is happening in the case of $\lambda\,|\mu|=2k$. The energy value $-\lambda^2/8$ is then equal to $-k^2/2\mu^2$. We note that $-k^2/2\mu^2$ is the minimum value of energy because the potential $U_\mu=\mu^2/2r^2-k/r$ of the reduced system has the minimum value $-k^2/2\mu^2$. Therefore, the S^2 in the above theorem consists of all the points of equilibrium.

V. THE SYMMETRY GROUP OF THE REDUCED HAMILTONIAN SYSTEM

This section shows that the symmetry group SO(4) of the reduced system is derived from the symmetry subgroup of the conformal Kepler problem through the reduction.

It is well known that SU(4) acts on the energy surface $K=4k$ of the harmonic oscillator $(T^*\dot{\mathbb{R}}^4,d\theta,K)$. Since the regularized energy surface, $\overline{H}=-\lambda^2/8$, coincides with $K=4k$, as was pointed in Ref. 2, SU(4) also acts on the regularized energy surface $\overline{H}=-\lambda^2/8$.

We now look for subgroups of SU(4) that leave $M_{\lambda,\mu}$ invariant. We start with the case of $\lambda\,|\mu|<2k$. Let A be a 4×4 matrix leaving $M_{\lambda,\mu}$ invariant. Then it must leave invariant the conditions (4.5a) and (4.5b), so that A is expressed in the form

$$A=\begin{pmatrix} B & \\ & C \end{pmatrix},\quad B,C\in\mathrm{U}(2). \tag{5.1}$$

If A is a matrix of SU(4), B and C are subject to $(\det B)\times(\det C)=1$. Thus we have a subgroup $S(\mathrm{U}(2)\times\mathrm{U}(2))$ of SU(4), which acts on $M_{\lambda,\mu}$. We examine $\mathrm{S}(\mathrm{U}(2)\times\mathrm{U}(2))$ in detail. Let $\det B=e^{it}$ $(0\leqslant t\leqslant 2\pi)$. Then A is decomposed to either

$$\begin{pmatrix} e^{it/2}I_2 & \\ & e^{-it/2}I_2 \end{pmatrix}\begin{pmatrix} B' & \\ & C' \end{pmatrix} \tag{5.2a}$$

or

$$\begin{pmatrix} -e^{it/2}I_2 & \\ & -e^{-it/2}I_2 \end{pmatrix}\begin{pmatrix} -B' & \\ & -C' \end{pmatrix}, \tag{5.2b}$$

where B' and C' are the elements of SU(2) that satisfy $B=e^{it/2}B'$ and $C=e^{-it/2}C'$, respectively. Expressions (5.2a) and (5.2b) show that $\mathrm{U}(1)\times\mathrm{SU}(2)\times\mathrm{SU}(2)$ is a double covering group of $S(\mathrm{U}(2)\times\mathrm{U}(2))$, where the factor U(1) in the former has the parameter t ranging over $0\leqslant t\leqslant 4\pi$. It is easy to see that $\mathrm{U}(1)\times\mathrm{SU}(2)\times\mathrm{SU}(2)$ also acts on $M_{\lambda,\mu}$ through (5.2a).

In what follows, we treat $\mathrm{U}(1)\times\mathrm{SU}(2)\times\mathrm{SU}(2)$ instead of $S(\mathrm{U}(2)\times\mathrm{U}(2))$. The U(1) of $\mathrm{U}(1)\times\mathrm{SU}(2)\times\mathrm{SU}(2)$ gives the action (2.3) or (4.7). We now look at $\mathrm{SU}(2)\times\mathrm{SU}(2)$ acting on $M_{\lambda,\mu}$. Let the action of $A\in\mathrm{SU}(2)\times\mathrm{SU}(2)$ be denoted by Φ_A. Since U(1) and SU(2)

$\times$SU(2) commute, one can well define the reduced action $\tilde{\Phi}_A$ on the energy surface $H_\mu = -\lambda^2/8$ by

$$\tilde{\Phi}_A \circ \pi_\mu (w) = \pi_\mu \circ \Phi_A (w), \tag{5.3}$$

where $w \in M_{\lambda,\mu} \subset \mathbb{C}^4$. The action of A on the energy surface $H_\mu = -\lambda^2/8$ is not effective. In fact, for $A = (I_2, I_2)$ and $A = (-I_2, -I_2)$, the definition (5.3) reads $\tilde{\Phi}_A \circ \pi_\mu (w) = \pi_\mu (w)$ for any $w \in M_{\lambda,\mu}$. Hence, we see that SU(2) $\times$SU(2)/$\mathbb{Z}_2 \simeq$SO(4) acts on the energy surface $H_\mu = -\lambda^2/8$ effectively, where $\mathbb{Z}_2 = \{(I_2, I_2), (-I_2, -I_2)\}$.

In the case of $\lambda |\mu| = 2k$, the symmetry subgroup should reduce to SU(2), either of the factors of SU(2)$\times$SU(2), according to the degeneracy of $S^3 \times S^3$ to S^3. The same discussion as above shows that SU(2)/$\mathbb{Z}_2$ $\simeq$SO(3) acts on the energy surface $H_\mu = -k^2/2\mu^2$, where $\mathbb{Z}_2 = \{I_2, -I_2\}$.

Thus we obtain the following.

Theorem 5.1: The symmetry group acting on the energy surface $H_\mu = -\lambda^2/8$ effectively is either of the following:

(a) SO(4) $(\lambda |\mu| < 2k)$,

(b) SO(3) $(\lambda |\mu| = 2k)$.

VI. CONSTANTS OF MOTION FOR THE REDUCED HAMILTONIAN SYSTEM

In Sec. V, we have shown that a symmetry subgroup of the conformal Kepler problem is reduced to the symmetry group of the reduced system. On a similar idea of reduction, we can obtain constants of motion for the reduced system. Let F be a function invariant under the U(1) action. Then one can define a function F_μ on the reduced phase space $(T^*\dot{\mathbb{R}}^3, \sigma_\mu)$ through $F_\mu \circ \pi_\mu = F \circ i_\mu$. Further, the Hamiltonian vector field X_{F_μ} is related to X_F by $\pi_{\mu *} X_F(x,p) = X_{F_\mu}(\pi_\mu(x,p))$, so that the flows of them are in the relation

$$\pi_\mu \circ \exp tX_F = (\exp tX_{F_\mu}) \circ \pi_\mu . \tag{6.1}$$

By (6.1) with F_μ replaced by H_μ, we can prove that F_μ is a constant of motion if F is so;

$$\begin{aligned} F_\mu((\exp tX_{H_\mu})(\pi_\mu(x,p))) &= F_\mu(\pi_\mu((\exp tX_H)(x,p))) \\ &= F(i_\mu((\exp tX_H)(x,p))) \\ &= F(i_\mu(x,p)) \\ &= F_\mu(\pi_\mu(x,p)). \end{aligned} \tag{6.2}$$

Thus we have shown the following lemma.

Lemma 6.1: The U(1)-invariant constants of motion for the conformal Kepler problem are reduced to the constants of motion for the MIC–Kepler problem $(T^*\dot{\mathbb{R}}^3, \sigma_\mu, \mathrm{H}_\mu)$.

The next thing we have to do is then to find U(1)-invariant constants of motion for the conformal Kepler problem. We here recall the relation (4,2), from which we can obtain[2]

$$4rX_H = X_K \quad \text{on } H = -\lambda^2/8 \text{ or } K = 4k. \tag{6.3}$$

This shows that the flows of X_H and X_K coincide, within a change of parameters, on the energy surface $H = -\lambda^2/8$ or $K = 4k$. Accordingly, constants of motion for the harmonic oscillator may be viewed as constants of motion for the conformal Kepler problem when restricted on the energy surface $H = -\lambda^2/8$. Thus we are seeking for U(1)-invariant constants of motion for the harmonic oscillator, for a while.

Constants of motion for the harmonic oscillator are expressed in the form[8]

$$F = \frac{1}{2i\lambda} \sum_{j,k=1}^{4} C_{kj} z_j \bar{z}_k , \tag{6.4}$$

where $C = (C_{jk})$ are anti-Hermitian matrices of $\operatorname{tr} C = 0$, and

$$z_j = \lambda x_j + i p_j . \tag{6.5}$$

Let Q be a constant of motion with coefficient matrix D. Then the Poisson bracket of F and Q is given by

$$\{F, Q\} = \frac{1}{2i\lambda} \sum_{j,k=1}^{4} [C, D]_{kj} z_j \bar{z}_k , \tag{6.6}$$

where $[C, D]$ denotes the commutator of matrices C and D.

We notice here that the momentum mapping Ψ we have dealt with is also given in the form (6.4) with C substituted by

$$N = \frac{1}{2} \begin{pmatrix} N_2 & \\ & N_2 \end{pmatrix} \quad \text{with } N_2 = \begin{pmatrix} & -1 \\ 1 & \end{pmatrix}. \tag{6.7}$$

Thus we know from (6.6) that U(1)-invariant constants of motion are those that have coefficient matrices commuting with N. Such a matrix C that commutes with N can be expressed as a sum of an antisymmetric real matrix A and a symmetric pure imaginary matrix B; $C = A + B$. We take basis A_j's and B_j's, $j = 1,2,3$, such that A and B are expressed, respectively, in the form

$$A = \sum_{j=1}^{3} a_j A_j = \begin{pmatrix} 0 & a_3 & a_2 & a_1 \\ -a_3 & 0 & -a_1 & a_2 \\ -a_2 & a_1 & 0 & a_3 \\ -a_1 & -a_2 & -a_3 & 0 \end{pmatrix}, \tag{6.8a}$$

and

$$B = \sum_{j=1}^{3} b_j B_j = \frac{i\lambda}{4} \begin{pmatrix} b_3 & 0 & b_1 & -b_2 \\ 0 & b_3 & b_2 & b_1 \\ b_1 & b_2 & -b_3 & 0 \\ -b_2 & b_1 & 0 & -b_3 \end{pmatrix}, \tag{6.8b}$$

where $a_j, b_j \in \mathbb{R}, j = 1,2,3$.

Hence the constants of motion corresponding to A_j's and B_j's are expressed, by writing out (6.4) for respective matrices, in the form

$$\begin{aligned} J_1 &= \tfrac{1}{2}(x_1 p_4 - x_4 p_1 + x_3 p_2 - x_2 p_3), \\ J_2 &= \tfrac{1}{2}(x_1 p_3 - x_3 p_1 + x_2 p_4 - x_4 p_2), \\ J_3 &= \tfrac{1}{2}(x_1 p_2 - x_2 p_1 + x_4 p_3 - x_3 p_4), \\ Q_1 &= \tfrac{1}{4}(p_1 p_3 + p_2 p_4) + (\lambda^2/4)(x_1 x_3 + x_2 x_4), \\ Q_2 &= \tfrac{1}{4}(p_2 p_3 - p_1 p_4) + (\lambda^2/4)(x_2 x_3 - x_1 x_4), \\ Q_3 &= \tfrac{1}{8}(p_1^2 + p_2^2 - p_3^2 - p_4^2) + (\lambda^2/8) \\ &\quad \times (x_1^2 + x_2^2 - x_3^2 - x_4^2). \end{aligned} \tag{6.9}$$

These are viewed as constants of motion for the conformal Kepler problem when restricted on the energy surface $H = -\lambda^2/8$. We now show that these functions can be made into constants of motion for the conformal Kepler problem. To this end, we proceed to investigate these constants of motion in detail. One can see that the J_j's are themselves constants of motion for the conformal Kepler problem, because calculation gives $\{J_j,H\} = 0$. We turn to the Q_j's. Calculation shows that the Poisson bracket of Q_j and H is put into the form

$$\{Q_j,H\} = -(1/r)(H+\lambda^2/8)\{Q_j,r\}, \quad j = 1,2,3. \tag{6.10}$$

Since $\{Q_j,r\} \neq 0$, Eq. (6.10) means that Q_j is constant along the flows of H, if and only if the flows are on $H = -\lambda^2/8$. We can make Q_j's into constants of motion, without restriction on the special energy surface, by substituting $-8H$ for λ^2 in the Q_j's. In fact if we denote by the $\widetilde{Q}_j$'s the functions made in such a manner, we see that the Poisson bracket of $\widetilde{Q}_j$ and H vanishes. Further, the Poisson brackets among them are calculated to give

$$\begin{aligned}&\{J_h,J_j\} = \epsilon_{hjk}J_k, \quad \{J_h,\widetilde{Q}_j\} = \epsilon_{hjk}\widetilde{Q}_k,\\ &\{\widetilde{Q}_h,\widetilde{Q}_j\} = \epsilon_{hjk}J_k(-2H).\end{aligned} \tag{6.11}$$

Thus we have the following.

Lemma 6.2: Functions J_j and $\widetilde{Q}_j$, $j = 1,2,3$, defined by (6.9) with λ^2 replaced by $-8H$ are U(1)-invariant constants of motion for the conformal Kepler problem, and subject to the commutation relations (6.11).

Now, from Lemmas 6.1 and 6.2 we can obtain constants of motion for the reduced system by the recipe

$$[J_j]_\mu \circ \pi_\mu = J_j \circ i_\mu, \quad [\widetilde{Q}_j]_\mu \circ \pi_\mu = \widetilde{Q}_j \circ i_\mu. \tag{6.12}$$

After using the same method as applied in obtaining the reduced Hamiltonian H_μ, we can obtain

$$\begin{aligned}&[J_1]_\mu = \tilde{x}_2\tilde{p}_3 - \tilde{x}_3\tilde{p}_2 + \mu\tilde{x}_1/r,\\ &[J_2]_\mu = \tilde{x}_3\tilde{p}_1 - \tilde{x}_1\tilde{p}_3 + \mu\tilde{x}_2/r,\\ &[J_3]_\mu = \tilde{x}_1\tilde{p}_2 - \tilde{x}_2\tilde{p}_1 + \mu\tilde{x}_3/r,\\ &[\widetilde{Q}_1]_\mu = ([J_2]_\mu\tilde{p}_3 - [J_3]_\mu\tilde{p}_2) + k\tilde{x}_1/r,\\ &[\widetilde{Q}_2]_\mu = ([J_3]_\mu\tilde{p}_1 - [J_1]_\mu\tilde{p}_3) + k\tilde{x}_2/r,\\ &[\widetilde{Q}_3]_\mu = ([J_1]_\mu\tilde{p}_2 - [J_2]_\mu\tilde{p}_1) + k\tilde{x}_3/r.\end{aligned} \tag{6.13}$$

Commutation relations among these constants of motion are the same as (6.11) because of the following lemma.

Lemma 6.3: Let F and Q be U(1)-invariant functions on $T^*\dot{\mathbf{R}}^4$, and F_μ and Q_μ the reduced functions on $T^*\dot{\mathbf{R}}^3$ defined by (3.3) with H replaced by F and Q, respectively. Then

$$\{F,Q\}\circ i_\mu = \{F_\mu,Q_\mu\}\circ\pi_\mu, \tag{6.14}$$

where the Poisson bracket in the right-hand side of (6.14) is defined not through the canonical symplectic form $d\theta'$ but through the reduced symplectic form σ_μ.

Proof: We note that the Hamiltonian vector fields for F and F_μ are related by $\pi_{\mu *}X_F(x,p) = X_{F_\mu}(\pi_\mu(x,p))$. The same relation holds for the Hamiltonian vector fields for Q and Q_μ. Then from the definition of the Poisson bracket and of the reduced symplectic form it follows that

$$\begin{aligned}\{F,Q\}\circ i_\mu(x,p) &= d\theta(X_Q,X_F)\circ i_\mu(x,p)\\ &= (i_\mu^* d\theta)(X_Q,X_F)(i_\mu(x,p))\\ &= (\pi_\mu^*\sigma_\mu)(X_Q,X_F)(x,p)\\ &= \sigma_\mu(\pi_{\mu *}X_Q,\pi_{\mu *}X_F)(\pi_\mu(x,p))\\ &= \sigma_\mu(X_{Q_\mu},X_{F_\mu})(\pi_\mu(x,p))\\ &= \{F_\mu,Q_\mu\}\circ\pi_\mu(x,p).\end{aligned} \tag{6.15}$$

This ends the proof.

We mention that, using the equation of motion (3.6), MacIntosh and Cisneros[6] derived two vector constants of motion; for $\tilde{x}\in\dot{\mathbf{R}}^3$ and $\tilde{p}\in\mathbf{R}^3$,

$$D_\mu = \tilde{x}\times\tilde{p} + \mu\tilde{x}/r, \quad R_\mu = D_\mu\times\tilde{p} + k\tilde{x}/r. \tag{6.16}$$

Our constants of motion (6.13) coincide with theirs. The D_μ and R_μ are the total angular momentum and the Runge–Lenz-like vector. If $\mu = 0$, these constants of motion become the well-known ones in the Kepler problem.

Theorem 6.4: The U(1)-invariant constants of motion given in Lemma 6.2 are reduced to constants of motion, given by (6.13), for the MIC–Kepler problem $(T^*\dot{\mathbf{R}}^3,\sigma_\mu,H_\mu)$. These are subject to the Poisson bracket relation

$$\begin{aligned}&\{[J_h]_\mu,[J_j]_\mu\} = \epsilon_{hjk}[J_k]_\mu,\\ &\{[J_h]_\mu,[\widetilde{Q}_j]_\mu\} = \epsilon_{hjk}[\widetilde{Q}_k]_\mu,\\ &\{[\widetilde{Q}_h]_\mu,[\widetilde{Q}_j]_\mu\} = \epsilon_{hjk}[J_k]_\mu(-2H_\mu).\end{aligned} \tag{6.17}$$

[1] J. E. Marsden and A. Weinstein, Rep. Math. Phys. **5**, 121 (1974).

[2] T. Iwai, J. Math. Phys. **22**, 1633 (1981).

[3] E. L. Stiefel and G. Scheifele, *Linear and Regular Celestial Mechanics* (Springer, Berlin, 1971), Chap. 2.

[4] M. Kummer, Commun. Math. Phys. **84**, 133 (1982).

[5] M. Kummer, Indiana Univ. Math. J. **30**, 281 (1981).

[6] H. MacIntosh and A. Cisneros, J. Math. Phys. **11**, 896 (1970).

[7] W. J. Satzer, Jr., Indiana Univ. Math. J. **26**, 951 (1977).

[8] T. Iwai, J. Math. Phys. **22**, 1628 (1981).

[9] T. Iwai and Y. Uwano, "The conformal Kepler problem is reduced to the three-dimensional Kepler problem with a centrifugal potential and Dirac's monopole field. Quantum theory," to be submitted to J. Math. Phys.

[10] Missing entries are all zero.

[11] R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin/Cummings, Reading, MA, 1978), 2nd ed., Chap. 4.

On an expression for the average resolvent using Grassmann integration

Nazakat Ullah
Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400005, India

(Received 2 December 1985; accepted for publication 12 February 1986)

The integral representation of the inverse of a determinant and the Grassmann representation of a determinant are used to derive an expression for the average resolvent for a Gaussian orthogonal ensemble. The expression is compared with the one obtained using Lagrangian formalism.

I. INTRODUCTION

It has been shown[1] recently that Grassmann integration[2] provides a powerful tool in calculating various ensemble averages that are needed, e.g., in the study of the probability density function of single eigenvalue and many other problems. In these studies one makes use of a generating function[3] involving a Lagrangian that has both ordinary and Grassmann variables. The purpose of the present work is to describe a different formalism based on the integral representation of a determinant and its inverse to derive an expression for the average resolvent. We shall show that the final expression that one obtains this way has an explicit dependence on the dimension of the matrix. In working out the ensemble averages we shall consider only the Gaussian orthogonal ensemble[4] (GOE).

We describe the formulation in Sec. II. Concluding remarks will be presented in Sec. III.

II. FORMULATION

The ensemble average resolvent $g(z)$ is defined by

$$g(z) = (1/N)\langle \mathrm{Tr}(z-H)^{-1}\rangle , \tag{1}$$

where H is a real symmetric $N \times N$ matrix. The $\langle\ \rangle$ sign denotes the ensemble average using the following distribution of the matrix elements of H:

$$P(\{H_{ij}\}) = 2^{(1/4)N(N-1)}\pi^{-(1/4)N(N+1)}\exp(-\mathrm{Tr}\,H^2) . \tag{2}$$

It is easy to see that $g(z)$ can be written as

$$g(z) = \frac{1}{N}\frac{(\partial/\partial\xi)\det(\xi-H)}{\det(z-H)}\bigg|_{\xi=z} . \tag{3}$$

Since the distribution of the matrix elements H_{ij} is Gaussian, we use the following integral representation for $[\det(z-H)]^{-1/2}$:

$$[\det(z-H)]^{-1/2} = \pi^{-N/2}i^{-N/2}\int \exp\Big(-\sum_{m,n} x_m x_n [(z_i - iz_r)\delta_{mn} + iH_{mn}]\Big)\prod_m dx_m , \tag{4}$$

where, for the convergence of the integral, we have assumed $z_i > 0$.

For the $\det(\xi-H)$ we use its Grassmann representation given by[2]

$$\det(\xi-H) = \int \exp\Big(-\sum_{m,n} a_m^*(\xi\delta_{mn} - H_{mn})a_n\Big)\prod_m da_m^*\, da_m . \tag{5}$$

By writing a similar expression to (4) with the integration variables y_m we get the representation of the inverse of the determinant in expression (3).

It is now straightforward to carry out the integrations over the matrix elements of H. Using expressions (2)–(5) we get

$$g(z) = (N\pi^N i^N)^{-1}\frac{\partial}{\partial\xi}\int \exp\Big[-\frac{1}{4}\sum_k (x_k^2+y_k^2)^2 + iz\sum_k (x_k^2+y_k^2) + \sum_{k<j}(x_kx_j+y_ky_j)^2 - \xi\sum_k a_k^*a_k - \frac{i}{2}\sum_k (x_k^2+y_k^2)a_k^*a_k + \frac{1}{2}\sum_{k<j} a_k^*a_ka_j^*a_j + i\sum_{k<j}(x_kx_j+y_ky_j)(a_k^*a_j+a_j^*a_k)\Big]\bigg|_{\xi=z}\prod_k dx_k\, dy_k\, da_k^*\, da_k . \tag{6}$$

The next step in the derivation is to integrate over x_k, y_k, a_k^*, and a_k. The Lagrangian formalism this step is carried out using the generalized Hubbard–Stratonovitch transformation.[1] In the present formulation we can collect terms of one kind, e.g., $\Sigma_k x_k^4$ and $\Sigma_{k<j}(x_kx_j)^2$, and express them as $(\Sigma x_k^2)^2$ and then apply the usual transformation,[5,6] which converts a Gaussian into an exponential to each such term individually. Thus the first three terms in the exponent in expression (6) can be rewritten as

$$\pi^{-3/2}\int\prod_{i=1}^{3} dt_i \exp\Big(-\sum_{i=1}^{3} t_i^2\Big)\exp\sum_k [(iz+it_1)x_k^2 + (iz+it_2)y_k^2 + i\sqrt{2}t_3x_ky_k] . \tag{7a}$$

A similar transformation can be written down for the term $\frac{1}{2}\Sigma_{k<j}a_k^*a_ka_j^*a_j$, which will be written as an integral over a new variable t_4.

The only remaining terms are now the ones that are products of x_k or y_k with a_k^*, a_k, namely the fifth and the seventh term in the exponent. These terms are taken care of by introducing four new Grassmann variables $\eta^*, \eta, \theta^*, \theta$ and are rewritten as

0022-2488/86/061530-03$02.50

$$\int \exp \sqrt{\frac{i}{2}} \left[\sum_k (\eta^* a_k^* - \eta a_k) x_k + \sum_k (\theta^* a_k^* - \theta a_k) y_k - \eta^* \eta - \theta^* \theta \right] d\eta^* \, d\eta \, d\theta^* \, d\theta . \tag{7b}$$

We can now easily carry out the integrations over the variables x_k, y_k, a_k^*, a_k. Using expressions (6), (7a), and (7b) we can write $g(z)$ as

$$g(z) = \pi^{-2} \int \exp\left[- \sum_{i=1}^{4} t_i^2 - \eta^* \eta - \theta^* \theta \right] \left[(z - t_1)(z - t_2) - \frac{1}{2} t_3^2 \right]^{-(3N-2)/2} \left[\left(z - \frac{it_4}{\sqrt{2}} \right) \left[(z - t_1)(z - t_2) - \frac{1}{2} t_3^2 \right] \right.$$
$$\left. + \frac{1}{4} \left\{ (z - t_1) \theta^* \theta + (z - t_2) \eta^* \eta + \frac{t_3}{\sqrt{2}} (\eta^* \theta - \eta \theta^*) \right\} \right]^{N-1} \prod_{i=1}^{4} dt_i \, d\eta^* \, d\eta \, d\theta^* \, d\theta . \tag{8}$$

Thus the average resolvent can be expressed as an eight-dimensional integral given by expression (8).

For further discussion of the average resolvent we introduce the matrix σ given by

$$\sigma = \begin{bmatrix} t_1 & t_3/\sqrt{2} & \eta/2 & \eta^*/2 \\ t_3/\sqrt{2} & t_2 & \theta/2 & \theta^*/2 \\ -\eta^*/2 & -\theta^*/2 & it_4/\sqrt{2} & 0 \\ \eta/2 & \theta/2 & 0 & it_4/\sqrt{2} \end{bmatrix} , \tag{9}$$

and the matrix

$$F = z - \sigma . \tag{10}$$

The matrix F in the block form is written as

$$F = \begin{bmatrix} a & \Sigma \\ \rho & b \end{bmatrix} . \tag{11}$$

Introducing the graded determinant of F given by[1]

$$\operatorname{detg} F = \det(a - \Sigma b^{-1} \rho) (\det b)^{-1} , \tag{12}$$

expression (8) can also be written as

$$g(z) = \int \exp - \operatorname{Trg} \left[\sigma^2 + \frac{N}{2} \ln(z - \sigma) \right] \left(z - \frac{it_4}{\sqrt{2}} \right)^{-1} \left[1 - \frac{1}{4} \left(z - \frac{it_4}{\sqrt{2}} \right)^{-1} \left[(z - t_1)(z - t_2) - \frac{t_3^2}{2} \right]^{-1} \right]$$
$$\times \left[\eta^* \eta (z - t_2) + \theta^* \theta (z - t_1) + \frac{t_3}{\sqrt{2}} (\eta^* \theta - \eta \theta^*) \right.$$
$$\left. + \frac{1}{8} \left(z - \frac{it_4}{\sqrt{2}} \right)^{-2} \left[(z - t_1)(z - t_2) - \frac{t_3^2}{2} \right]^{-1} \eta^* \eta \, \theta^* \theta \right] d[\sigma] , \tag{13}$$

where

$$d[\sigma] = \pi^{-2} \prod_{i=1}^{4} dt_i \, d\eta^* \, d\eta \, d\theta^* \, d\theta .$$

From expressions (9)–(11) it can be shown that expression (13) further can be written as

$$g(z) = \int \exp - \operatorname{Trg}\left[\sigma^2 + \frac{N}{2} \ln(z - \sigma) \right]$$
$$\times \frac{1}{2} \operatorname{Tr}(b - \rho a^{-1} \Sigma)^{-1} d[\sigma] . \tag{14}$$

Thus the present formulation gives the average resolvent in terms of the trace of the lower block of the matrix $(z - \sigma)^{-1}$.

III. CONCLUDING REMARKS

As a check on expression (8) for the average resolvent we calculated it explicitly for $N = 2$ and compared it with its exact form from the known two-dimensional distribution of the single eigenvalue. The two expressions checked as they should. We further calculated the ensemble averages of quantities like $(1/N)\operatorname{Tr}\langle H^2 \rangle$ by expanding $g(z)$ given by expression (8) in powers of $1/z$ and found that they also check with their exact values calculated directly using the distribution of the Hamiltonian matrix elements.

We now compare expression (14) with the one obtained using Lagrangian formalism.[1] In our notation it is given by

$$g(z) = \int \exp - \operatorname{Trg}\left[\sigma^2 + \frac{N}{2} \ln(z - \sigma) \right]$$
$$\times \frac{1}{4} \operatorname{Tr}(z - \sigma)^{-1} d[\sigma] . \tag{15}$$

Since both expressions (14) and (15) are exact we conclude that the integral

$$\int \exp\left\{ - \operatorname{Trg}\left[\sigma^2 + \frac{N}{2} \ln(z - \sigma) \right] \operatorname{Trg}(z - \sigma)^{-1} d[\sigma] \right\} , \tag{16}$$

must vanish.

We have not been able to find a simple way to prove this result but have shown it to be true by explicit calculation for the two-dimensional case. For the general case by expanding ln and $(z - \sigma)^{-1}$ in powers of (σ/z) we have shown that the integral is zero for the few lowest powers of σ/z.

Lastly we remark that both expressions (14) and (15) give the same asymptotic form of $g(z)$, namely,

$$g(z) = (2/N)\left[z - \sqrt{z^2 - N}\right],$$

as they should.

[1] J. J. M. Verbaarschot, H. A. Weidenmüller, and M. R. Zirnbauer, "Grassmann integration in stochastic Quantum Physics: The case of compound-nucleus scattering," MPIH preprint No. V5, 1985.

[2] *Methods in Field Theory*, Les Houches Ecole d'Etude Physique Theorique, Session XXVIII, edited by R. Balian and J. Zinn-Justin (North-Holland, Amsterdam, 1976).

[3] J. J. M. Verbaarschot, H. A. Weidenmüller, and M. R. Zirnbauer, Phys. Rev. Lett. **52**, 1597 (1984).

[4] M. L. Mehta, *Random Matrices* (Academic, New York, 1967).

[5] J. Hubbard, Phys. Rev. Lett. **3**, 77 (1959); R. L. Statanovich, Sov. Phys. Dokl. **2**, 416 (1958).

[6] N. Ullah and K. K. Gupta, Nucl. Phys. A **186**, 331 (1972).

Asymptotics of the maximum number of repulsive particles on a spherical surface

Alexander A. Berezin
Department of Engineering Physics, McMaster University, Hamilton, Ontario, L8S 4M1, Canada

(Received 8 October 1985; accepted for publication 5 February 1986)

There are N equal particles interacting through a repulsive potential $V(r) = A/r^{\beta}$ ($\beta > 0$) placed insided the sphere. For $\beta = 1$ (Coulomb case) and for all $\beta < 1$, the minimum energy configuration will have all N particles on the inner surface of the sphere for any integer N. It is shown that starting from $N = 13$ and for $\beta > \beta_{\text{crit}}$ the minimum energy configuration will have only a fraction of particles on the inner surface while the rest of the particles will hang in the equilibrium inside the volume of the sphere. As far as $\beta \rightarrow 1+$, the maximum number of charges that can be held on the surface has an asymptotic $N_0 \sim \text{const}/(\beta - 1)^2$. For const, the gap $6 < C < 8$ was found. The theory may be applicable to, e.g., "magic numbers" in small atomic clusters. Some problems that are not yet solved are listed.

I. INTRODUCTION

Let us place N point charges inside the spherical surface. If the law of interaction between the particles is the repulsive Coulomb potential $V(r) = q^2/r$, then the particles will arrange themselves in such a way as to minimize the total potential energy. The resulting configuration of minimum energy (CME) will change in a quite peculiar jumpwise manner with each time N increases from N to $N + 1$.

Apparently, this classical problem was first clearly stated by Thomson at about 1900 (see, e.g., Ref. 1). Despite some progress that has been achieved,[2–7] the precise structure of the CME is still known only for some few values of N and in many cases as a plausible conjection only.[7] No "general algorithm" that is equally valid for any integer N was offered to describe how the CME will change with each increase of N by 1, and this problem is, by and large, still unsolved. Even for the "simple" cases $N = 4$, 6, 8, 12, and 20 (numbers of the vertices of five Platonic bodies), two out of five regular polyhedrons fail to provide the required CME. While the triad of tetrahedron, octahedron, and icosahedron ($N = 4$, 6, 12) with triangular faces do indeed provide the CME,[3,4,7] the cube ($N = 8$) and dodecahedron ($N = 20$) are *not* the minimum energy configurations! This startling fact, which seems to contradict common sense, illustrates well that this problem is very far from being trivial.

It is not, however, the purpose of this paper to contribute further to the above problem of purely Coulombic point charges on the sphere. Here we will consider another, but related, problem: What will happen with N point particles inside the sphere if the law of interaction deviates from the pure Coulombic and can be represented by the potential

$$V(r) = A/r^{\beta}, \tag{1}$$

where $\beta > 1$? This may be of relevance to, e.g., Pauli (exchange) forces[5] or to the "magic numbers" in small clusters, etc. Here we will, however, consider this problem irrespective of the concrete (and, perhaps, numerous) physical implications it may have.

Although for $\beta = 1$ (Coulomb case) we do not know the stable arrangements of points on a sphere for all integers N, one fact is, nevertheless, certain: No matter how large N is, all charges will *always* be resting on the inner surface of the sphere once the global minimum of the potential energy is reached. In contrast to the pseudo-two-dimensional charge confinement within the circle,[8] in a truly three-dimensional case there will be no "charge ejection" from the surface into the volume,[9] since for the Coulombic particles, such "ejection" would violate the Earnshaw stability theorem.

However, in the case when the interparticle interaction is given by Eq. (1), we are no longer bound by the Earnshaw theorem, which is valid only for $\beta = 1$. It may seem intuitively plausible, even without any calculations, that for the sufficiently large values of β (i.e., for the sharply falling-off interactions) one may place only a *finite* number of particles on the spherical surface. Here we will determine this critical value N_0 as a function of β. As we will see, for every $\beta > 1$ there indeed exists an upper limit for the number of particles that can be held on the surface at the most stable (minimum energy) configuration before the spontaneous ejection towards the center of the sphere will take place.

Another slightly different, but related, line of development was originated by Fejes Tóth[10] and has gained an interesting literature, of which we will mention here Refs. 11–14. It is sometime referred to as the "problem of inimical dictators" and can be put in the following form[7,14]: "A spherical planet (without oceans) is governed by N mutually inimical dictators. How should they place their residences in order to get as far as possible from one another?" Equivalently: "How can N fuel depots be arranged on a planet so that an accidental explosion of one of them will least endanger the rest?"

When formalized, this problem is equivalent to the finding of the configuration of N points on a sphere that will maximize the minimum distance between any two points. Similarly to the previous case, the solution of this problem is also known only for some specific values of N.

II. ENERGY OF N REPULSIVE PARTICLES ON A SPHERE

Let us now find an approximate expression for the minimum energy of N particles placed on a surface of a unit

0022-2488/86/061533-04$02.50

sphere if the law of interaction is given by Eq. (1). In the limit $N \to \infty$, we were able to obtain a compact and easily analyzed expression.

Of course, except for $N = 2$, 3, 4, 6, 8, 12, and 20, it is impossible to place N points fully symmetrically on the surface of a sphere in such a way that all positions are exactly equivalent. However, it is intuitively clear (and can be argued more rigorously[12]) that for large N, each given point will be surrounded almost precisely by six others and these six will form an (almost) regular hexagon around this given point. The inevitable distortions of the hexagons (which is necessary for the complete filling) will become relatively smaller and smaller with the increase of N. To stress that the sphere can be covered by equal hexagons only approximately, we will call the above honeycomblike coverage quasihexagonal.

The total energy of N point particles interacting through the potential of Eq. (1) is

$$W(N) = \sum_{1 < i < j < N} \frac{1}{|\mathbf{R}_i - \mathbf{R}_j|^{\beta}} \cong \frac{N\Phi}{2}. \tag{2}$$

Here $\mathbf{R}_i$ is the radius vector of the particle i ($i = 1,2,...,N$) and Φ is the potential created by all other $N-1$ particles at the position of any given one. The second (approximate) equality becomes exact in the limit $N \to \infty$, indicating that in this limit all particles are equivalent, i.e., Φ is becoming the same for all particles.

Note, also, that for the simplicity we put $A = 1$ in Eq. (1). This does not limit the generality of our consideration: for arbitrary constant A and the radius of the sphere R all following energies have to be expressed in units A/R^{β}. Therefore, the present formulation of the problem is dimensionless: all following results depend only on the value of β, regardless of the scale for A and R.

Let us assume now that N points form a quasihexagonal coverage over the surface of a unit sphere in the above given sense and place, for convenience, one particle at the northern pole (we call this particle the pole particle). For large N the nearest neighbor distance is $a = (8\pi/N\sqrt{3})^{1/2}$. The last can be easily seen from equating the total area (4π) to $N{\cdot}S$ ($S = a^2\sqrt{3}/2$ is the area of one hexagon, i.e., the area per one particle).

To write the expression for Φ we will use the semicontinuum approximation. Examples of its use may be found, e.g., in some old papers on the theory of electronic F centers in ionic crystals (see, e.g., the review in Ref. 15, pp. 188ff). Its basic idea is to calculate the interaction of a given particle with some number of close neighbors "exactly," while replacing the sum of all pointwise interactions with all other particles of the system by an appropriate integral expression.

In the semicontinuum approximation, Φ can be written as

$$\Phi = \frac{6}{a^{\beta}} + \int_{S'} \frac{ds}{a^2\sqrt{3}/2} \frac{1}{r^{\beta}}. \tag{3}$$

The first term is the interaction energy of the pole particle with six nearest neighbors. The second term represents the interaction of the pole particle with the remaining $N-7$ particles of the system. To calculate the second term, the sum was replaced by the integral over S', which is the entire spherical surface with the exception of the small circle of radius R_0 around the pole ($R_0 \ll 1$). In Eq. (3), $a^2\sqrt{3}/2$ is area per particle (flat approximation) and r is the Euclidean (i.e., straight, not spherical) distance from the northern pole to the floating point. There are seven particles inside this small circle, so its radius (in radians) can be taken as $2(7/N)^{1/2}$ (from $7a^2\sqrt{3}/2 = \pi R_0^2$).

Within the framework of the above approximation, Eq. (3) develops into the following expression for the total potential energy of the equilibrium configuration of N particles:

$$W(N) = \frac{N^2}{2} \frac{2^{1-\beta}}{2-\beta}\left[1 - \left(\frac{7}{N}\right)^{1-\beta/2}\right] + 3N\left(\frac{N\sqrt{3}}{8\pi}\right)^{\beta/2}. \tag{4}$$

Equation (4), therefore, gives the total energy as an explicit function of N and β.

The quality of the approximations of W by Eq. (4) increases with the increase of N. However, even for small N its quality is surprisingly satisfactory, e.g., for $N = 12$ and $\beta = 1.5$, $W = 44.056\,07$, while the exact value is 43.465 18. The last can be calculated easily if one notes that in the icosahedron inscribed into a unit sphere there are three different intervertex distances, which are $\sqrt{\frac{2}{5}(5-\sqrt{5})} = 1.051\,462...$, $\sqrt{\frac{2}{5}(5+\sqrt{5})} = 1.701\,301...$, and 2. This leads to the following (exact) expression for the total energy of the icosahedral configuration:

$$W(\text{exact}) = \frac{30}{\left[\frac{2}{5}(5-\sqrt{5})\right]^{\beta/2}} + \frac{30}{\left[\frac{2}{5}(5+\sqrt{5})\right]^{\beta/2}} + \frac{6}{2^{\beta}}. \tag{5}$$

It is possible, of course, to extend the "exact" summation over not just one, but several consecutive concentric circumferences of nearest neighbors around the pole. This will reduce the relative weight of the second (integral) term in Eq. (3) and will make the results more precise. Straightforward consideration of the hexagonal flat tile gives the following numbers of points lying on the consecutive concentric circumferences surrounding the pole particle: 6, 6, 6, 12, 6, 6, 12,... . Their corresponding separations from the origin are a, $a\sqrt{3}$, $2a$, $a\sqrt{7}$, $3a$, $2\sqrt{3}{\cdot}a$, $a\sqrt{13}$,..., respectively.

Taking this into account, Eq. (4) develops into

$$W = \frac{N^2}{2} \frac{2^{1-\beta}}{2-\beta}\left[1 - \left(\frac{n}{N}\right)^{1-\beta/2}\right] + \frac{N}{2}\left(\frac{N\sqrt{3}}{8\pi}\right)^{\beta/2} \times\left[6 + \frac{6}{(\sqrt{3})^{\beta}} + \frac{6}{2^{\beta}} + \frac{12}{(\sqrt{7})^{\beta}} + \frac{6}{3^{\beta}} + \frac{6}{(\sqrt{12})^{\beta}} + \frac{12}{(\sqrt{13})^{\beta}} + \cdots\right]. \tag{6}$$

Here n is the number of particles that are considered pointwise (i.e., $6 + 6 + 6 + 12 + \cdots + 1$; the last 1 relates to the pole particle). We are still assuming that $n \ll N$, i.e., the circle of radius R_0 can be seen as arbitrary flat.

As we will see in Sec. III, in the limit $N \to \infty$, Eq. (6) leads to a minor numerical correction in comparison with Eq. (4). This correction does not really affect any of the following conclusions.

III. THRESHOLD OF THE PARTICLE EJECTION FROM THE SPHERE

Let us consider the energy of another arrangement of N particles when one particle has been moved to the center of the sphere. In this, second, configuration we have $N-1$ particles still on the surface of the sphere. Force acting from the central particle to any one on the surface will be directed along the radius of the sphere. Consequently, the central particle will have no effect on the establishing of the equilibrium among $N-1$ particles remaining on the surface. Therefore, the total energy of this configuration can be easily expressed as

$$E = W(N-1) + N - 1. \tag{7}$$

The second term in Eq. (7) is the interaction energy of the central particle with the $N-1$ remaining on the surface (since $R=1$ and $A=1$ each interaction contributes a unit value).

Suppose now that by changing N or β we reached a certain combination of N and β for which the difference $E-W$ switches from positive to negative. This will mean that the second configuration took over as the CME, i.e., this combination of N and β will correspond to the reaching of the ejection threshold. For larger values of N (for fixed β) or, alternatively, for larger values of β (for fixed N) the "ground state" may, of course, have more than one particle inside the volume. Apparently, with the further increase of β the inner particles will start forming concentric shells with each new shell appearing at a certain critical combination of N and β. This gradual buildup of inner shells can, indeed, be considered as a classical analog of the periodic filling of atomic shells in Mendeleev's table or of the shell model in nuclear physics.

Due to the use of the semicontinuum approximation, the precision of Eqs. (4) and (7), increases with the increase of N. Using the Eqs. (4) and (7) one can establish the occurrence of this "spontaneous ejection" for all $N \geqslant 14$. For each value of N there is a critical value of β such that for all $\beta > \beta_{\text{crit}}$ there will be a spontaneous ejection of at least one particle into the volume.

As one can find using the values from Table 1 of Ref. 14 the ejection actually should start at $N=13$. Our equation for $E-W$ fails to demonstrate the beginning of the ejection at $N=13$ and the first value for which we see it happening is $N=14$. Of course, the exact minimization with all 13 points treated explicitly [and not with the approximate equations (4) and (7)] should exhibit such ejection and could establish the critical value of β for $N=13$ as well. We are not performing this cumbersome calculation here.

For a few representative values of the ejection thresholds N, Eqs. (4) and (6) lead to the critical values of β given in Table I (the cross-point values of W are given in brackets).

IV. ASYMPTOTICS OF THE EJECTION THRESHOLD FOR LARGE N

From the above figures it is easy to see that for large N the value of β approaches 1 (from above) and simple analysis suggests that the threshold of ejection N_0 behaves as

$$N_0 \sim \text{const}/(\beta - 1)^2. \tag{8}$$

TABLE I. Critical values of β. The cross-point values of W are given in brackets.

$N=14$	$N=17$	$N=20$	$N=40$
$\beta=10.4779$	4.8623	3.6945	2.0710
(35.1057)	(83.8355)	(134.6776)	(685.2563)
$N=100$	$N=500$	$N=1000$	$N=10\,000$
$\beta=1.4789$	1.1592	1.104 58	1.028 78
(4691.686)	(122 680.2)	(494 102.3)	(4.9848 E7)

These asymptotics can be obtained from the series expansions of Eqs. (5) and (7) in the limit $N \to \infty$ (or, equivalently, $\beta \to 1+$). In the approximation contained in Eq. (4) (six nearest neighbors treated pointwisely while the rest are treated continuously; i.e., $n=7$) the value of const in Eq. (8) is

$$\text{const} = \frac{9}{16(1-\ln 2)^2}\left[\sqrt{7} - 6\left(\frac{\sqrt{3}}{8\pi}\right)^{1/2}\right]^2 = 6.847\,75\ldots. \tag{9}$$

If, instead, we consider pointwise several circumferences surrounding the pole particle, the const in Eq. (8) changes to

$$\text{const} = \frac{9}{16(1-\ln 2)^2}\left\{\sqrt{n} - \left(\frac{\sqrt{3}}{8\pi}\right)^{1/2} \times\left[\frac{6}{\sqrt{1}} + \frac{6}{\sqrt{3}} + \frac{6}{\sqrt{4}} + \frac{12}{\sqrt{7}} + \frac{6}{\sqrt{9}} + \frac{6}{\sqrt{12}} + \frac{12}{\sqrt{13}} + \cdots\right]\right\}^2 \tag{10}$$

(the sum of numerators in the last bracket should be $n-1$). Note, that in Eq. (10) we already took a limit $\beta \to 1+$, so all powers in denominators are square roots instead of $\beta/2$ as in Eq. (6).

For 1, 2, 3, 4, 5, 6, or 7 circumferences, respectively, $n = 7, 13, 19, 31, 37, 43$, and 55. This leads to the numerical value of const as 6.847 75, 7.507 75, 7.056 54, 7.294 75, 7.162 80, 7.426 58, and 7.228 69, respectively. Although we did not prove this rigorously, it seems likely that when $n \to \infty$ (but still $n \ll N$), the above expression converges to the value somewhere between 6 and 8, or, possibly, between 6.5 and 7.5. More scrupulous analysis to reduce this gap, of course, can be performed, but we are leaving it as an open exercise. It is also, of course, possible to get rid of the flat approximation and take into account the finite curvature of the sphere's top.

V. CONCLUSIONS

One may wonder if the ejection always starts (at fixed N and gradually increasing β) from the jump of just one particle to the center of the sphere. It seems a likely conjecture that we, however, did not prove. In the less likely event that the opposite is true, our asymptote for N_0 still provides an upper limit for the ejection from the surface for a given value of β.

Besides the possible applications for plasma confinement studies, the present model can be of relevance to the recently emerging physics of small clusters, where the existence of some "magic" numbers of the enhanced stability has been recently observed.[16]

We considered here the case of N equal repulsive particles placed inside the spherical surface. Our analysis, though approximate, is, nevertheless, dimensionless: all above numbers depend only on β and do not depend on the radius of the sphere R or constant A in the interaction energy [Eq. (1)] between the two particles.

The methodology suggested in the present paper can be rather straightforwardly refined in several aspects. Here we will point out several possibilities for further studies: multiparametric minimization to account for the formation of the concentric shells, other surfaces than spherical could be treated, particles may not necessarily be equal, etc.

One may consider what will happen if the sphere is gradually deformed into an ellipsoid by contraction (or stretching) of one of the axes. Less symmetric distortion may also appear interesting.

It would be interesting also to add the effect of external fields on the least energy arrangement and critical values of N. Some preliminary results[17] indicate the possibility of interesting effects in the presence of the electric field.

The charges (or, more specifically, constants A) for various particles can be assumed nonequal (e.g., one may consider the case A, $2A$, $3A$,...). The list of unexplored possibilities can, of course, go on and on.

[1] J. J. Thomson, Philos. Mag. **41**, 510 (1921).
[2] L. Föppl, J. Reine. Angew. Math. **141**, 251 (1912).
[3] L. L. Whyte, Am. Math. Month. **59**, 606 (1952).
[4] J. Leech, Math. Gaz. **41**, 81 (1957).
[5] R. J. Gillespie, Canad. J. Chem. **38**, 818 (1960).
[6] M. Goldberg, Math. Comut. **23**, 785 (1969).
[7] T. W. Melnyk, O. Knop, and W. R. Smith, Canad. J. Chem. **55**, 1745 (1977).
[8] A. A. Berezin, Nature (London) **315**, 104 (1985).
[9] A. A. Berezin, Nature (London) **317**, 208 (1985).
[10] L. Fejes Tóth, Am. Math. Month. **56**, 330 (1949).
[11] K. Schutte and B. L. van der Waerden, Math. Ann. **123**, 96 (1951).
[12] W. Habicht and B. L. van der Waerden, Math. Ann. **123**, 223 (1951).
[13] R. M. Robinson, Math. Ann. **144**, 17 (1961).
[14] H. S. M. Coxeter, Trans. NY Acad. Sci. Ser. II **24**, 320 (1962).
[15] B. S. Gourary and F. J. Adrian, in *Solid State Physics*, edited by F. Seitz and D. Turnbull (Academic, New York, 1960), Vol. 10, p. 127.
[16] M. R. Hoare, in *Advances in Chemical Physics*, edited by I. Prigogine and S. A. Rice (Wiley, New York, 1979), Vol. 40, p. 49.
[17] A. A. Berezin, Phys. Canad. **41**(3), 30 (1985).

Logarithmic corrections to the uncertainty principle and infinitude of the number of bound states of *N*-particle systems

J. Fernando Perez, F. A. B. Coutinho, and C. P. Malta
Instituto de Física, Universidade de São Paulo, CP 20516, 01498 São Paulo, SP, Brazil

(Received 13 November 1985; accepted for publication 21 January 1986)

It is shown that critical long-distance behavior for a two-body potential, defining the finiteness or infinitude of the number of negative eigenvalues of Schrödinger operators in ν dimensions, is given by $v_k(r) = -((\nu-2)/2r)^2 - 1/(2r\ln r)^2 + \cdots - 1/(2r\ln r\cdot\ln\ln r\cdots\ln_{(k)} r)^2$, where $k = 0,1,\ldots$ for $\nu\neq 2$ and $k = 1,2,\ldots$ if $\nu = 2$. This result is a consequence of logarithmic corrections to an inequality known as the uncertainty principle. If the continuum threshold in the N-body problem is defined by a two-cluster breakup the results presented generate corrections to the existing sufficient conditions for the existence of infinitely many bound states.

I. INTRODUCTION

It is well known that the finiteness or infinitude of the number of bound states of negative energy of a Schrödinger operator $[-\Delta + V]$ is controlled by the long-distance behavior of the potential.[1-4] For dimension $\nu\neq 2$ a finiteness–infinitude borderline is set by a falloff $\sim((\nu-2)/2r)^2$ as $r\to\infty$. Not coincidentally, for the quadratic form $(\psi,[-\Delta+V]\psi)$, $\psi\in C_0^\infty(\mathbb{R}^\nu\backslash 0)$ and V being a Kato potential,[5] the following results hold: (A) the "Uncertainty Principle Lemma,"[2,6-8] if $V(x) > -((\nu-2)/2r)^2$ then

$$(\psi,[-\Delta+V]\psi)\geqslant 0\,; \tag{1.1}$$

and (B) if, for $\alpha>1$, $r\geqslant R_0>0$, $V(x)\leqslant -\alpha((\nu-2)/2r)^2$ then there exists an infinite sequence $\{\psi_n\in C_0^\infty(\mathbb{R}^\nu\backslash 0)\}_{n>1}$, with disjoint supports, such that

$$(\psi_n,[-\Delta+V]\psi_n)<0\,. \tag{1.2}$$

From (A) (as proved by Simon[1] for $\nu=3$) it follows that if $V(x)\geqslant -((\nu-2)/2r)^2$ for $r\geqslant R_0>0$, then $[-\Delta+V]$ has at most finitely many negative eigenvalues. Under the assumptions of (B), the "min-max principle" implies the existence of infinitely many eigenstates of negative energy.

For $\nu=2$, however, property (A) is trivial and property (B) is false!

The original purpose of our investigation was to determine the critical asymptotic behavior of the potential for $\nu=2$. The answer is that for $\nu=2$ the critical (in the same sense as above) long-distance falloff is $\sim -1/(2r\ln r)^2$. This follows from appropriately modified versions of (A) and (B) above.

Nevertheless, it turns out that the $\nu=2$ result is only the first term of an infinite series of logarithmic corrections for $\nu=1$ and 3 results! This is a consequence of the following chain of facts.

(i) Under *suitable* domain restrictions, the unitary operator

$$T_\nu\colon L^2(\mathbb{R}_+,r^{\nu-1}\,dr)\to L^2(\mathbb{R}_+,dr),$$
$$(T_\nu\psi)(r) = r^{(\nu-1)/2}\psi(r)\,,$$

establishes a unitary equivalence between the radial part of the two-dimensional Laplacian and the critically perturbed radial part of the ν-dimensional Laplacian:

$$\begin{aligned}T_2\Big(-\frac{1}{r}\frac{d}{dr}r\frac{d}{dr}\Big)T_2^{-1}\\ &= T_\nu\Big[-\frac{1}{r^{\nu-1}}\frac{d}{dr}r^{\nu-1}\frac{d}{dr}-\Big(\frac{\nu-2}{2r}\Big)^2\Big]T_\nu^{-1}\\ &= -\frac{d^2}{dr^2}-\frac{1}{4r^2}\,.\end{aligned} \tag{1.3}$$

More generally, if $a\colon \mathbb{R}_+\backslash N_a\to\mathbb{R}_+$ is C^∞ and $a(r)>0$ for all $r\in\mathbb{R}_+\backslash N_a$, where N_a is a finite set, then the unitary operator $U_a\colon L^2(\mathbb{R}_+,a\,dr)\to L^2(\mathbb{R}_+,dr)$, given by $(U_a\psi)(r) = a^{1/2}\psi(r)$ transforms the "radial a-Laplacian" as

$$\begin{aligned}U_a\Big(-\frac{1}{a}\frac{d}{dr}a\frac{d}{dr}\Big)U_a^{-1}\\ &= \Big[-\frac{d^2}{dr^2}-\frac{1}{4}\Big(\frac{a'}{a}\Big)^2+\frac{1}{2}\Big(\frac{a''}{a}\Big)\Big],\end{aligned} \tag{1.4}$$

when restricted, for instance, to $C_0^\infty(\mathbb{R}_+\backslash N_a)$. (From now on we shall use a prime to denote derivatives with respect to r.)

Remark: Since $(-(1/a)(d/dr)a(d/dr))$ is a positive operator, when restricted to $C_0^\infty(\mathbb{R}_+\backslash N_a)$, (1.3) provides a trivial proof of the "Uncertainty Principle Lemma."

(ii) For a class of functions $a(r)$ as above, it is possible to find a critical potential v_a for the a-Laplacian. It is given by

$$v_a(r) = -1/[2a(r)h(r)]^2\,, \tag{1.5}$$

where h is a monotonic function satisfying

$$h'(r) = 1/a(r)\,. \tag{1.6}$$

In fact, denoting by S_a the finite set where a or h are zero, we prove the following lemma.

Lemma 1: If $\psi\in C_0^\infty(\mathbb{R}_+\backslash S_a)$, then

$$\int(\psi')^2a\,dr\geqslant\int v_a\psi^2a\,dr\,. \tag{1.7}$$

Lemma 2: If $\lim\limits_{r\to\infty} h(r) = \infty$, then, given $\epsilon>0$ arbitrary, there exists an infinite family of nonzero functions, with disjoint supports $\{\psi_n\in C_0^\infty(\mathbb{R}_+\backslash S_a)\}_{n>1}$ such that

$$\int(\psi_n')^2a\,dr<(1+\epsilon)\int v_a\,\psi_n^2\,a\,dr\,. \tag{1.8}$$

Remarks: Statement (1.7) is a version of an inequality of Hardy[2,6-8] known as the "Uncertainty Principle Lemma."

Lemma 2 says that the constants appearing in the definition v_a are best possible.

(iii) Finally, the whole procedure may be iterated provided we can find b: $\mathbb{R}_+\setminus N_b \rightarrow \mathbb{R}_+$, with the same assumed properties of $a(r)$, such that

$$U_b\left(-\frac{1}{b}\frac{d}{dr}b\frac{d}{dr}\right)U_b^{-1} = U_a\left[-\frac{1}{a}\frac{d}{dr}a\frac{d}{dr}+v_a(r)\right]U_a^{-1}. \tag{1.9}$$

Starting with $a=r$ and iterating the whole procedure we obtain the result that the potentials

$$v_k(r) = -\left(\frac{\nu-2}{2r}\right)^2 - \left(\frac{1}{2r\ln r}\right)^2 - \cdots - \left(\frac{1}{2r\ln r\ln_{(2)}r\cdots\ln_{(k)}r}\right)^2, \tag{1.10}$$

for $k \geqslant 0$ if $\nu \neq 2$ and $k \geqslant 1$ if $\nu = 2$ are critical; i.e., for some $r \geqslant R_0 > 0$, (a) if $V(x) > (1+\epsilon)v_k(r)$ then $[-\Delta + V]$ has finitely many negative eigenvalues, or (b) if

$$V(x) \leqslant v_{k-1}(r) - \frac{1+\epsilon}{(2r\ln r\cdots\ln_{(k)}r)^2},$$

for some $\epsilon > 0$, then $[-\Delta + V]$ has infinitely many negative eigenvalues.

Notation: For $k \geqslant 2$, $\ln_{(k)}r = \ln\ln_{(k-1)}r$, and $\ln_{(1)}r = \ln r$.

Our results amount, in fact, to logarithmic corrections to the "Uncertainty Principle," a widely used tool in the proofs of self-adjointness of strongly singular potentials (see, for instance, Refs. 2, 7, and 8). In a separate paper[9] we discuss the implications of our results to this problem.

Relative to the two-body problem, the N-body problem presents the extra difficulty of locating the threshold (the infimum of the essential spectrum of the N-body Hamiltonian with center of mass motion removed). However, if the threshold, as given by Hunziker's theorem,[10] is defined by a two-cluster breakup we can extend the results of Simon[1] concerning sufficient conditions for the existence of infinitely many bound states.

This paper is organized as follows. In Sec. II we prove Lemmas 1 and 2 and discuss the two-body problem. In Sec. III the N-body problem is briefly discussed.

II. THE TWO-BODY PROBLEM: FINITENESS AND INFINITUDE

A general proof of inequalities of type (1.7) can be found in Ref. 7. For completeness we present the following simple proof.

Proof of Lemma 1: Let $\psi(r) = g(r)\varphi(r)$, where $g^2 = h$. Then

$$\int(\psi')^2 a\,dr \geqslant \int\varphi^2(g')^2 a\,dr + 2\int gg'\varphi\varphi' a\,dr$$

$$= \int\psi^2\left(\frac{g'}{g}\right)^2 a\,dr + \frac{1}{2}\int(\varphi^2)'(g^2)'a\,dr$$

$$= \int\psi^2 v_a a\,dr. \qquad \text{Q.E.D.}$$

Proof of Lemma 2: (1) Let us first consider the case $a(r) = 1$ and $h(r) = r$. Since, for $\psi = r^{1/2}\varphi$,

$$\int(\psi')^2 dr = \left\{1 + \frac{\int(\varphi')^2 r\,dr}{\int(\varphi^2/r)dr}\right\}\int\psi^2 v_a\,dr,$$

it is enough to show the existence of an infinite sequence $\{\varphi_n \in C_0^\infty(\mathbb{R}_+\setminus S_a)\}_{n>1}$ such that

$$\frac{\int(\varphi_n')^2 r\,dr}{\int(\varphi_n^2/r)dr} < \epsilon. \tag{2.1}$$

The left-hand side of (2.1) is scale invariant, i.e.,

$$\frac{\int(\varphi_\alpha')^2 r\,dr}{\int(\varphi_\alpha^2/r)dr} = \frac{\int(\varphi')^2 r\,dr}{\int(\varphi^2/r)dr},$$

where $\varphi_\alpha(r) = \varphi(\alpha r)$. It is, therefore, sufficient to find just one $\varphi \in C_0^\infty(\mathbb{R}_+\setminus S_a)$ satisfying (2.1) and the infinite sequence $\varphi_n = \varphi(\alpha_n r)$ will be generated by suitably choosing α_n to make the supports disjoint. A possible choice of φ is, given in Ref. 11,

$$\varphi(r) = \begin{cases} 0, & r \leqslant R_0, \\ \rho(r-R_0), & R_0 \leqslant r \leqslant R_0+1, \\ 1, & R_0+1 \leqslant r \leqslant R_0+N, \\ \rho(1-(r-R_0)/N), & R_0+N \leqslant r \leqslant R_0+2N, \\ 0, & r \geqslant R_0+2N, \end{cases}$$

with $R_0 > \max_{r\in S_a} r$ and $\rho \in C^\infty(\mathbb{R}_+)$ with $\rho(r) = 0$ if $0 \leqslant r \leqslant \frac{1}{4}$, $\rho(r) = 1$ if $r \geqslant \frac{3}{4}$. Since $\lim_{N\rightarrow\infty}(\int(\varphi')^2 r\,dr/\int(\varphi^2/r)dr) = 0$, it is enough to take N sufficiently large to verify (2.1).

(2) Let now $\tilde{\psi}(r) = \psi(h(r))$. Then

$$\int(\tilde{\psi}')^2 a\,dr = \int(\psi')^2 dr$$

and

$$\int(\tilde{\psi})^2 v_a a\,dr = \frac{1}{4}\int\frac{\psi^2}{r^2}dr.$$

Taking then $\tilde{\psi}_n = \psi_n \circ f$ with ψ_n as given in part (1) makes the proof complete. Q.E.D.

Remarks: The assumption $\lim_{r\rightarrow\infty} h(r) = \infty$ is used to guarantee that the functions $\tilde{\psi}_n(r) = \psi_n(h(r))$ are not identically zero. It is not the best possible assumption for the result is still true if $a(r) = r^n$, $n \geqslant 1$. However, some assumption on $a(r)$ is required as the result is false if $a(r)h(r) = r^n$, $n \geqslant 1$.

We now describe how, starting with $a_0 = r$, it is possible to generate an infinite chain of logarithmic corrections to the "Uncertainty Principle" as described by Lemmas 1 and 2.

Let $a_n(r) = a_{n-1}(r)\ln_{(n)}r$, $n = 1,2,\ldots$. A straightforward computation gives, for all $\psi \in C_0^\infty(\mathbb{R}_+\setminus S_{a_n})$,

$$U_{a_n}\left(-\frac{1}{a_n}\frac{d}{dr}a_n\frac{d}{dr}\right)U_{a_n}^{-1}\psi = U_{a_{n-1}}\left(-\frac{1}{a_{n-1}}\frac{d}{dr}a_{n-1}\frac{d}{dr}+v_{a_{n-1}}\right)U_{a_{n-1}}^{-1}\psi, \tag{2.2}$$

with U_a as given in the Introduction. Therefore, applying Lemmas 1 and 2 to a_n we obtain the following lemma.

Lemma 3: Let $v_k(r)$ be given by

$$v_0(r) = -(\nu-2)^2/4r^2\,, \tag{2.3a}$$

$$v_k(r) = v_{k-1}(r) - \frac{1}{(2r\Pi_{n=1}^{k}\ \ln_{(n)} r)^2}\,,\quad k = 1,2,\dots\,. \tag{2.3b}$$

Then (a) for $\psi\in C_0^\infty(\mathbf{R}_+\backslash S_{a_k})$,

$$\int(\psi')^2 dr \geqslant \int \psi^2 v_k\, dr \tag{2.4}$$

and (b) for $\epsilon > 0$, there exists an infinite sequence of nonzero functions, with disjoint supports, $\{\psi_n\in C_0^\infty(\mathbf{R}_+\backslash S_{a_k})\}$ such that

$$\int(\psi_n')^2 dr \leqslant \int(\psi_n)^2\left[v_k - \frac{1+\epsilon}{(2r\,\Pi_{n=1}^{k}\ \ln_{(n)} r)^2}\right] dr\,. \tag{2.5}$$

One of the main ingredients in our discussion below is the so called "min–max principle": Let H be a self-adjoint operator in Hilbert space with quadratic form domain $Q(H)$, and for $n = 1,2,\dots$ let

$$\mu_n(H) = \sup_{\varphi_1,\dots,\varphi_{n-1}}\ \inf_{\substack{\psi\in[\varphi_1,\dots,\varphi_{n-1}]^\perp\\ \|\psi\|=1,\ \psi\in Q(H)}} (\psi,H\psi)\,, \tag{2.6}$$

where $[\varphi_1,\dots,\varphi_{n-1}]^\perp$ indicates the orthogonal complement of the subspace generated by $\varphi_1,\dots,\varphi_{n-1}$. Then, for each n, either (a) there are n eigenvalues (counting multiplicities) below the bottom of the essential spectrum, and $\mu_n(H)$ is the nth eigenvalue counting multiplicity in increasing order or (b) μ_n is the bottom of the essential spectrum, and in this case, $\mu_n = \mu_{n+1} = \cdots$ and there are at most $(n-1)$ eigenvalues (counting multiplicity) below μ_n.

We are now prepared to state and prove our main results.

Theorem 1: Let V be a Kato potential in $L^2(\mathbf{R}^\nu)$, $\nu = 1,2,3$, such that for some $R_0 > 1$ and $\epsilon > 0$,

$$V(x) \leqslant v_k(r) - \frac{1+\epsilon}{(2r\Pi_{n=1}^{k+1}\ \ln_{(n)} r)^2}\,,$$
$$k = 0,1,\dots \quad \text{if } \nu\neq 2,\quad k = 1,2,\dots \quad \text{if } \nu = 2.$$

Then, the operator $[-\Delta + V]$ has infinitely many negative eigenvalues.

Proof: By the min–max principle, it is sufficient to exhibit an infinite sequence $\{\psi_n\in Q(-\Delta + V)\}_{n>1}$, with disjoint supports, such that $(\psi_n,\ [-\Delta + V]\psi_n) < 0$. The existence of such a sequence follows from Lemma 3. Q.E.D.

Theorem 2: Let V be a Kato potential in $L^2(\mathbf{R}^\nu)$, $\nu = 1,2,3$, such that, for $R_0 > 1$, $c < 1$, and k,

$$V(x) \geqslant C\, v_k(r),\quad \text{if } r > R_0\,,$$

where $k = 0,1,\dots$ if $\nu\neq 2$ and $k = 1,2,\dots$ if $\nu = 2$. Then $[-\Delta + V]$ has at most finitely many negative eigenvalues.

Proof: We first decompose our operator into

$$-\Delta + V = (-C\Delta + V\chi_2) + (-(1-C)\Delta + V\chi_1)\,,$$

where $\chi_1\in C_0^\infty$, $\chi_1(x) = 1$ if $r\leqslant R_0$, $0\leqslant\chi_1\leqslant 1$, and $\chi_2(x) = 1-\chi_1(x)$.

From a simple application of the min–max principle, it follows that if both operators $A = -(1-C)\Delta + V\chi_1$ and $B = -C\Delta + V\chi_2$ (which are essentially self-adjoint in the same domain and have the same essential spectrum) have finitely many negative eigenvalues then the same holds for $-\Delta + V = A + B$ (see, for instance, Ref. 2, Vol. IV, exercise 129, p. 379).

That the operator A has finitely many negative eigenvalues is a standard result since the potential $V\chi_1$ has compact support (see, for instance, Ref. 2, Vol. IV, exercise 20, p. 366). On the other hand, by assumption, $B > C \times(-\Delta + \chi_2 v_k)$ and it is therefore sufficient to show that the operator $-\Delta + \chi_2 v_k$ has finitely many negative eigenvalues. If $\nu\geqslant 2$ it is sufficient to consider the operator $-\Delta + \chi_2 v_k$ restricted to the subspace $\mathscr{H}_0$ of spherically symmetric functions since in $\mathscr{H}_0^\perp$ the operator is positive! The restriction to $\mathscr{H}_0$ is given by the operator

$$H_k = \left\{-\frac{1}{r^{\nu-1}}\frac{d}{dr}r^{\nu-1}\frac{d}{dr} + \chi_2 v_k(r)\right\}.$$

For $\nu = 1$ we consider the operator $(-d^2/dx^2)_D + \chi_2 v_k$, with Dirichlet boundary conditions on $\pm R_0$.

For $\nu = 2,3$, a similar argument applies for the operator H_k restricted to $\mathscr{H}_0$, thus concluding the proof. Q.E.D.

Remarks: From the proofs it is clear that the finiteness or infinitude is controlled by the following limits:

$$u_0 = \lim_{r\to\infty}(2r)^2 V(x),\dots,$$

$$u_k = \lim_{r\to\infty}\left\{\left(2r\prod_{n=1}^{k}\ln_{(n)} r\right)^2[V(x) - v_{k-1}(r)]\right\}.$$

Indeed finiteness is implied by

$$u_0 = \cdots = u_{k-1} = -1,$$
$$u_k > -1,\quad \text{for some } k\geqslant 0 \quad \text{if } \nu\neq 2\,,$$
$$u_1 = \cdots = u_{k-1} = -1,$$
$$u_k > -1,\quad \text{for some } k\geqslant 1 \quad \text{if } \nu = 2\,,$$

whereas infinitude is guaranteed by

$$u_0 = \cdots = u_{k-1} = -1,$$
$$u_k < -1,\quad \text{for some } k\geqslant 0 \quad \text{if } \nu\neq 2\,,$$
$$u_1 = \cdots = u_{k-1} = -1,$$
$$u_k < -1,\quad \text{for some } k\geqslant 1 \quad \text{if } \nu = 2\,.$$

III. THE *N*-BODY PROBLEM: INFINITUDE

This section constitutes a sort of appendix of Sec. 3 of Simon's work.[1] Therefore we shall not give all the details and instead we shall be rather sketchy.

Let us consider a system of N particles, with masses m_i, $i = 1,\dots,N$, in $\nu = 1,2$ or 3 dimensions, interacting via two-body Kato potentials $V_{ij}(\mathbf{r}_i - \mathbf{r}_j)$. The Hamiltonian H_N, after removal of the center of mass motion,

$$H_N = \sum_{i=1}^{N}\frac{p_i^2}{2m_i} + \sum_{i<j}^{N} V_{ij}(\mathbf{r}_i - \mathbf{r}_j) - \frac{(\Sigma\mathbf{p}_i)^2}{2(\Sigma m_i)}\,,$$

has the infimum Σ of its essential spectrum given by Hunziker's theorem[10]:

$$\Sigma = \min_{\substack{D_1\cap D_2=\phi\\ D_1\cup D_2=\{1,\dots,N\}}} [\sigma_{D_1}+\sigma_{D_2}]\,,$$

where $\sigma_D =$ infimum spectrum H_D; here H_D denotes the Hamiltonian of the cluster $D\subset\{1,\dots,N\}$, with center of mass kinetic energy removed. If $\sigma=\sigma_{D_1}+\sigma_{D_2}$ and H_{D_1} and H_{D_2} have discrete ground states at the bottom of their spectra we say, after Ref. 1, that the system has a "two-cluster continuum limit."

It should be remarked that there are a number of situations for which it can be proved that the system has a "two-cluster continuum limit," namely (a) for $\nu=1,2$, a sufficient condition is that $\int V_{ij}(\mathbf{x})d^\nu x<0$ (see Ref. 12); and (b) for $\nu=3$, a sufficient condition is that V_{ij}'s are purely attractive and hold a bound state.[13]

As in Ref. 1, if we are in the two-cluster limit case, sufficient conditions for infinitude can be obtained by reducing the analysis to that of an effective two-body problem.

Theorem 3: Let V_{ij} be Kato potentials that are C^∞ functions on an open set of $\mathbb{R}^\nu$ whose complement has zero measure and let Σ be given by a two-cluster breakup (D_1, D_2), with reduced mass $\mu_{D_1,D_2}=(1/\Sigma_{i\in D_1}m_i+1/\Sigma_{j\in D_2}m_j)^{-1}$. Denoting by $\mathbf{R}$ the relative coordinate of the center of masses of clusters D_1 and D_2, if the potential

$$\widetilde{V}_{D_1D_2}(\mathbf{R}) = 2\mu_{D_1D_2}\sum_{\substack{i\in D_1\\ j\in D_2}} V_{ij}(\mathbf{R})$$

satisfies the assumptions of Theorem 1, then H_N has infinitely many eigenvalues below Σ.

Remark: We believe that this theorem holds for Kato potentials without that extra smoothness assumption.

Proof: Since $H_N=H_{D_1}+H_{D_2}+V_{D_1D_2}-(1/2\mu_{D_1D_2})\Delta_{\mathbf{R}}$, where

$$V_{D_1D_2}=\sum_{\substack{i\in D_1\\ j\in D_2}} V_{ij}(\mathbf{x}_i-\mathbf{x}_j)$$

is the intercluster potential, for $\psi=\psi_{D_1}\psi_{D_2}\phi$, we have

$$(\psi, H_N\psi)=E_{D_1}+E_{D_2}+(\phi,[-(1/2\mu_{D_1D_2})\Delta_{\mathbf{R}}+\overline{V}]\phi)\,,$$

where

$$\overline{V}(\mathbf{R})=\sum_{\substack{i\in D_1\\ j\in D_2}}(\psi_{D_1}\psi_{D_2},V_{ij}(\mathbf{x}_i-\mathbf{x}_j)\psi_{D_1}\psi_{D_2})$$

is the effective intercluster potential when the clusters D_1 and D_2 are in their bound states ψ_{D_1} and ψ_{D_2}, respectively, with corresponding energies E_{D_1} and E_{D_2}.

The proof of the theorem is completed by the following generalization of Proposition 5 in Ref. 1.

Lemma 4: Let ψ_{D_i} be a bound state of H_{D_i}, a k_i-body system with Kato potentials that are C^∞ functions on an open set of $\mathbb{R}^\nu$ whose complement has zero measure. Let V_{ij} be Kato potentials such that for some $\gamma\leqslant 2$ and $l\geqslant 1$

$$\overline{\lim_{r\to\infty}}\left(2r\prod_{n=1}^{l}\ln_{(n)} r\right)^\gamma[V_{ij}(\mathbf{x})-v_{l-1}(r)]\leqslant C_l\,,$$

v_l given by (2.3). Let

$$\overline{V}_{ij}(\mathbf{R})=\int|\psi_{D_1}(\mathbf{r}_1)|^2|\psi_{D_2}(\mathbf{r}_2)|^2 \times V_{ij}(\mathbf{r}_{ij}(\mathbf{R},\mathbf{r}_1,\mathbf{r}_2))d^{\nu(k_1-1)}r_1d^{\nu(k_2-1)}r_2\,,$$

where $\mathbf{r}_{ij}(\mathbf{R},\mathbf{r}_1,\mathbf{r}_2)$ is the distance between particles $i\in D_1$ and $j\in D_2$, in terms of the internal coordinates $\mathbf{r}_1(\mathbf{r}_2)$ of $D_1(D_2)$ and the distance $\mathbf{R}$ between the centers of mass of D_1 and D_2. Then

$$\overline{\lim_{R\to\infty}}\left(2R\prod_{n=1}^{l}\ln_{(n)} r\right)^\gamma[\overline{V}_{ij}(\mathbf{R})-v_{l-1}]\leqslant C_l\,.$$

Proof: The proof follows by repetition of the steps in Ref. 1, Proposition 5, having in mind that the extra smoothness assumption on the potentials ensures that the function

$$\rho(\mathbf{r}_0)=\int d^{\nu(k_1+k_2-\nu)}r|\psi_{D_1}(\mathbf{r}_1)|^2|\psi_{D_2}(\mathbf{r}_2)|^2$$

[with integration over all coordinates but $\mathbf{r}_0$: $\mathbf{r}_{ij}(\mathbf{R},\mathbf{r}_1,\mathbf{r}_2)=\mathbf{R}-\mathbf{r}_0$] decays faster than any power:

$$\sup_{\mathbf{r}_0}|\rho(\mathbf{r}_0)(1+r_0^n)|<\infty\,,$$

for all n. This is a result by Hunziker.[10] Q.E.D.

ACKNOWLEDGMENT

Two of us (J. F. B. and F. A. B. C.) were partially supported by CNPq.

[1] B. Simon, "On the infinitude or finiteness of the number of bound states of an N-body quantum system," Helv. Phys. Acta **43**, 607 (1970).

[2] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vols. II and IV (Academic, New York, 1975, 1978).

[3] M. Schechter, *Operator Methods in Quantum Mechanics* (North-Holland, New York, 1981).

[4] L. Landau and E. M. Lifshitz, *Mécanique Quantique* (Mir, Moscow, 1966).

[5] Through this paper we will assume that the potential functions satisfy the Kato condition, $V\in L^2(\mathbb{R}^\nu)+L^\infty_\epsilon(\mathbb{R}^\nu)$, i.e., for any $\epsilon>0$, there exists a decomposition $V=V_{1,\epsilon}+V_{2,\epsilon}$ with $V_{1,\epsilon}\in L^2(\mathbb{R}^\nu)$, $V_{2,\epsilon}\in L^\infty(\mathbb{R}^\nu)$ and $\|V_{2,\epsilon}\|_\infty<\epsilon$. This condition will ensure self-adjointness of the relevant Hamiltonians, both for the two-body and the N-body case (see Ref. 2).

[6] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. I (Interscience, New York, 1953).

[7] H. Kalf and J. Walter, "Strongly singular potentials and essential self-adjointness of singular elliptic operators," J. Funct. Anal. **10**, 114 (1972).

[8] H. Kalf, U. W. Schminke, J. Walter, and R. Wüst, in *Spectral Theory and Differential Equations, Lecture Notes in Mathematics*, Vol. 448, edited by W. N. Everitt (Springer, Berlin, 1975).

[9] J. Fernando Perez, F. A. B. Coutinho, and C. P. Malta, paper in preparation.

[10] W. Hunziker, Helv. Phys. Acta **39**, 451 (1966).

[11] J. Uchiyama, Publ. Res. Inst. Math. Sci. Kyoto A **2**, 117 (1966).

[12] F. A. B. Coutinho, C. P. Malta, and J. Fernando Perez, Phys. Lett. A **97**, 242 (1983).

[13] J. Fernando Perez, C. P. Malta, and F. A. B. Coutinho, J. Math. Phys. **26**, 2262 (1985).

On the hydrogen-oscillator connection: Passage formulas between wave functions

M. Kibler
Institut de Physique Nucléaire (et IN2P3), Université Claude Bernard Lyon-1, 43, Boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

A. Ronveaux
Département de Physique, Facultés Universitaires Notre-Dame de la Paix, 5000 Namur, Belgium

T. Négadi
Institut de Physique, Université d'Oran, Es-Sénia, Oran, Algeria

(Received 30 July 1985; accepted for publication 21 February 1986)

Recent works on the hydrogen-oscillator connection are extended to cover in a systematic (and easily computarizable) way the problem of the expansion of an $\mathbb{R}^3$ hydrogen wave function in terms of $\mathbb{R}^4$ oscillator wave functions. Passage formulas from oscillator to hydrogen wave functions are obtained in six cases resulting from the combination of the following coordinate systems: spherical and parabolic coordinate systems for the hydrogen atom in three dimensions, and Cartesian, double polar, and hyperspherical coordinate systems for the isotropic harmonic oscillator in four dimensions. These coordinate systems are particularly useful in physical applications (e.g., Zeeman and Stark effects for hydrogenlike ions and coherent state approaches to the Coulomb problem).

I. INTRODUCTION

The connection between the hydrogen atom in three dimensions and the isotropic harmonic oscillator in four dimensions has been a subject of considerable interest in the last fifteen years.[1-16] Such a connection has been studied in the framework of four formulations of nonrelativistic quantum mechanics. More precisely, the connection between the $\mathbb{R}^3$ hydrogen atom and the $\mathbb{R}^4$ harmonic oscillator has been worked out in (i) a (Schrödinger) partial-differential-equation approach,[1-8] (ii) a (Feynman) path-integral approach,[9-14] (iii) a (Weyl–Wigner–Moyal) phase-space approach,[15] and (iv) an (Heisenberg–Born–Jordan) operator approach based on an investigation of the Pauli equations via Schwinger calculus.[16] In most of the works in Refs. 1-16, the so-called Kustaanheimo–Stiefel transformation[17] (that corresponds to the Hopf fibration $S^3/S^1 = S^2$) is used in the derivation of the connection between the $\mathbb{R}^3$ hydrogen atom and the $\mathbb{R}^4$ harmonic oscillator. We note in passing that the extension to higher dimensions of this nonbijective canonical transformation is limited to a transformation (that corresponds to the Hopf fibration $S^7/S^3 = S^4$) from which it is possible to reach a connection between the $\mathbb{R}^5$ hydrogen atom and the $\mathbb{R}^8$ harmonic oscillator. In contradistinction, it might be interesting to note that the work in Ref. 16 could be extended without any *a priori* dimensional limitation.

Going back to the $\mathbb{R}^3$-$\mathbb{R}^4$ hydrogen-oscillator connection, it is to be emphasized that this connection is of paramount importance in the study of atomic systems subjected to electric and/or magnetic fields (cf. Refs. 18–21). In this respect, the problem of a hydrogenlike atom in an electric field or a strong magnetic field can be transformed, by means of the Kustaanheimo–Stiefel transformation, into the problem of a four-dimensional isotropic oscillator presenting anharmonicity of degree 4 or 6, respectively.[18] Therefore, many calculations arising in the Zeeman and Stark effects for the hydrogen atom may be conducted in the oscillator representation. It is thus desirable to have the wave functions for the hydrogen atom expressed in terms of wave functions for the corresponding four-dimensional oscillator.

As is well known, the Schrödinger equation for the hydrogen atom in $\mathbb{R}^3$ is separable in four coordinate systems, viz., the spherical, parabolic, spheroconical, and prolate spheroidal coordinate systems.[22-25] The spherical and parabolic coordinates are probably the most important as far as physical applications are concerned. On the other hand, the Schrödinger equation for an isotropic harmonic oscillator in $\mathbb{R}^4$ is separable in numerous coordinate systems. In the quaternionic (or Euler-angle) coordinates $(R, \theta, \varphi, \psi)$, defined on $\mathbb{R}^4$ through

$$u_1 = R\cos\frac{\theta}{2}\cos\frac{\varphi+\psi}{2}, \quad u_2 = R\cos\frac{\theta}{2}\sin\frac{\varphi+\psi}{2},$$

$$u_3 = R\sin\frac{\theta}{2}\cos\frac{\varphi-\psi}{2}, \quad u_4 = R\sin\frac{\theta}{2}\sin\frac{\varphi-\psi}{2},$$

the wave functions for the four-dimensional oscillator assume an especially simple form.[1] In addition, the passage formulas between the latter wave functions and those for the hydrogen atom in spherical coordinates are trivial in the sense that each hydrogen wave function corresponds to a particular oscillator wave function.[1]

It is the aim of this paper to obtain other passage formulas in the case of the discrete spectrum of hydrogen. For physical purposes, attention will be drawn to spherical and parabolic coordinates for the hydrogen atom and to Cartesian, double polar, and hyperspherical coordinates for the four-dimensional oscillator. The passage formulas are developed in Sec. V. We begin in Sec. II with the Kustaanheimo–Stiefel transformation and continue in Sec. III with the wave functions for the discrete spectra of the hydrogen atom and

0022-2488/86/061541-08$02.50

the harmonic oscillator in the coordinate systems under consideration. We close this paper in Sec. VI with some computing aspects.

II. THE HYDROGEN-OSCILLATOR CONNECTION

A. The KS transformation

We start from the $\mathbb{R}^4 \to \mathbb{R}^3$ transformation defined by

$$\begin{aligned} x_1(\equiv x) &= 2(u_1u_3 - u_2u_4)\,, \\ x_2(\equiv y) &= 2(u_1u_4 + u_2u_3)\,, \\ x_3(\equiv z) &= u_1^2 + u_2^2 - u_3^2 - u_4^2\,, \\ 0 &= u_2\,du_1 - u_1\,du_2 - u_4\,du_3 + u_3\,du_4\,. \end{aligned} \tag{1}$$

Equations (1) constitute, indeed, a simple rewriting, up to π_3 (on x_i, $i = 1, 2, 3$) and π_4 (on u_α, $\alpha = 1, 2, 3, 4$) permutations, of the transformation used by Kustaanheimo and Stiefel[17] in their work on the regularization of the three-dimensional (classical) Kepler problem. Following many authors, we refer to this transformation as the KS transformation. The KS transformation is a transformation of magnitude 2 since

$$r \equiv (x_1^2 + x_2^2 + x_3^2)^{1/2} = u_1^2 + u_2^2 + u_3^2 + u_4^2 \equiv u^2\,. \tag{2}$$

Furthermore, for any function $f(x_i)$ of the variables x_i ($i = 1, 2, 3$), at least twofold differentiable, we have

$$\begin{aligned} \Delta_3 f(x_i) &= (4r)^{-1}\Delta_4 f(x_i(u_\alpha))\,, \\ Xf(x_i(u_\alpha)) &= 0\,, \end{aligned} \tag{3}$$

where Δ_3 and Δ_4 are the Laplacian operators in three and four dimensions and where

$$X = u_2\,\partial_1 - u_1\,\partial_2 - u_4\,\partial_3 + u_3\,\partial_4 \tag{4}$$

is the infinitesimal operator of a Lie subgroup of type U(1) of the group SO(4) that leaves the form $u_1^2 + u_2^2 + u_3^2 + u_4^2$ invariant. Note that the introduction of the constraint condition $X = 0$ into the Lie algebra so(4) produces an *under constraint* Lie algebra, which is isomorphic to so(3) (cf. Ref. 16).

B. Application to the Schrödinger equation

We now consider the Schrödinger equation

$$-\tfrac{1}{2}(\hbar^2/\mu)\Delta_3\Psi + V\Psi = E\Psi \tag{5}$$

for a (one-particle) problem corresponding to an arbitrary potential V. The KS transformation allows converting this equation into the following:

$$-\tfrac{1}{2}(\hbar^2/\mu)\Delta_4\Psi - 4Eu^2\Psi = -4u^2V\Psi\,, \tag{6}$$

accompanied by the constraint relation

$$X\Psi = 0\,. \tag{7}$$

In the particular case of a hydrogenlike atom with reduced mass μ and nucleus charge Ze, we have the spherically symmetric potential (energy)

$$V = -Ze^2/r\,, \tag{8}$$

so that Eq. (6) reduces to the Schrödinger equation for an isotropic harmonic oscillator in four dimensions, the energy of which is ϵ,

$$\epsilon = 4Ze^2\,, \tag{9}$$

and the angular frequency of which is ω given by

$$-4E = \tfrac{1}{2}\mu\omega^2\,. \tag{10}$$

This four-dimensional oscillator is an (ordinary) *attractive* oscillator for the discrete spectrum ($E < 0$) of the hydrogen atom, a *repulsive oscillator* for the continuous spectrum ($E > 0$) of the hydrogen atom, and collapses into a *free particle* for the zero-energy point ($E = 0$) of the hydrogen atom. Equation (7) bears a nice interpretation from a group-theoretical point of view. In fact, the introduction of the constraint condition $X = 0$ into the Lie algebra of the noninvariance group Sp(8, $\mathbb{R}$) associated to the four-dimensional oscillator produces an *under constraint* Lie algebra, which turns out to be isomorphic to the Lie algebra of the group SU(2,2). This result has been independently obtained[16] from a boson realization of the Pauli equations corresponding to the whole spectrum ($E < 0$, $E > 0$, and $E = 0$) of the hydrogen atom. It provides us with a further way to understand the relevance of the group SO(4, 2), locally isomorphic to SU(2, 2), for the hydrogen atom problem.

We shall devote the rest of this paper to the bound states of the hydrogen atom. In the situation where $E < 0$, the angular frequency ω is indeed quantized as can be seen from Eq. (10). More precisely, for the discrete spectrum of the hydrogen atom, we obtain

$$\frac{\mu\omega}{\hbar} = 2\frac{1}{n}\frac{\mu Ze^2}{\hbar^2}\,, \qquad n \in \mathbb{N} - \{0\}\,, \tag{11}$$

a relation that will prove useful in Sec. V.

III. $\mathbf{R}^3$ HYDROGEN ATOM

We give here the wave functions on $L^2(\mathbf{R}^3)$, in spherical and parabolic coordinates, associated to the energy level

$$\begin{aligned} E \equiv E_n &= -(\alpha^2/2)(\hbar^2/\mu)\,, \\ \alpha &= (1/n)(\mu Ze^2/\hbar^2) \end{aligned} \tag{12}$$

of the discrete spectrum of the (three-dimensional) hydrogen atom. The notation used is a self-explanatory hybridization of the ones in Refs. 22–25.

A. Spherical coordinates

In the (conventional) spherical coordinates (r, θ, φ), we take the n^2 wave functions Ψ_{nlm} corresponding to the level E_n in the form

$$\Psi_{nlm} = N_{nlm}\rho^l Y_{lm}(\theta, \varphi)e^{-\rho/2}L^{2l+1}_{n+l}(\rho)\,, \tag{13}$$

where

$$\begin{aligned} N_{nlm} &= (-1)^a\left\{(2\alpha)^3\frac{(n-l-1)!}{2n[(n+l)!]^3}\right\}^{1/2}, \\ \rho &= 2\alpha r, \qquad l = 0, 1, ..., n-1, \qquad m = -l, -l+1, ..., l\,. \end{aligned} \tag{14}$$

In Eq. (14), a is an arbitrary phase. (Note that $a = 1$ in Ref. 22 and $a = n$ in Ref. 25.) We adopt the phase convention of Condon and Shortley for the spherical harmonics Y_{lm} and the definition of Ref. 22 for the associated Laguerre polynomials L^{2l+1}_{n+l}.

B. Parabolic coordinates

In the parabolic coordinates

$$\xi = r + z, \qquad \eta = r - z, \qquad \varphi = \arctan(y/x), \tag{15}$$

the n^2 wave functions $\Psi_{p_1p_2m}$ corresponding to the level E_n with

$$\begin{gathered} n = p_1 + p_2 + |m| + 1 , \\ p_1 \in \mathbf{N}, \quad p_2 \in \mathbf{N}, \quad m \in \mathbf{Z} \end{gathered} \tag{16}$$

assume the form

$$\begin{aligned} \Psi_{p_1p_2m} &= N_{p_1p_2m} (\xi\,\eta)^{|m|/2} e^{im\varphi} e^{-\alpha(\xi+\eta)/2} \\ &\quad \times L^{|m|}_{p_1+|m|}(\alpha\,\xi) L^{|m|}_{p_2+|m|}(\alpha\,\eta) . \end{aligned} \tag{17}$$

We leave the normalization constant $N_{p_1p_2m}$ as

$$\begin{aligned} N_{p_1p_2m} &= (-1)^b (n\pi)^{-1/2} \alpha^{|m|+3/2} \\ &\quad \times (p_1!p_2!)^{1/2} [(p_1+|m|)!(p_2+|m|)!]^{-3/2} , \end{aligned} \tag{18}$$

where b is an arbitrary phase. [Note that $b = 0$ in Ref. 23 and $b = p_1 + (m - |m|)/2$ in Ref. 25.]

IV. $\mathbb{R}^4$ HARMONIC OSCILLATOR

We give here the wave functions on $L^2(\mathbb{R}^4)$, in Cartesian, double polar, and hyperspherical coordinates, associated with the energy level

$$\begin{gathered} \epsilon \equiv \epsilon_{n_t} = (n_t + 2)\lambda(\hbar^2/\mu) , \\ \lambda = \mu\omega/\hbar, \qquad n_t \in \mathbf{N}, \end{gathered} \tag{19}$$

of an isotropic harmonic oscillator in four dimensions.

A. Cartesian coordinates

In Cartesian coordinates, we know that

$$\begin{gathered} n_t = n_1 + n_2 + n_3 + n_4 , \\ n_\alpha \in \mathbf{N}, \qquad \alpha = 1, 2, 3, 4 . \end{gathered} \tag{20}$$

For n_t fixed, the $C^{n_t}_{n_t+3}$ wave functions $\Psi_{n_1n_2n_3n_4}$ for the $\mathbb{R}^4$ oscillator are easily deduced from the well-known wave functions for an isotropic harmonic oscillator in one dimension. In detail, we have

$$\Psi_{n_1n_2n_3n_4} = N_{n_1n_2n_3n_4} \prod_{\alpha=1}^{4} e^{-(\lambda/2)u_\alpha^2} H_{n_\alpha}(\lambda^{1/2} u_\alpha) , \tag{21}$$

where

$$N_{n_1n_2n_3n_4} = \frac{\lambda}{\pi} \prod_{\alpha=1}^{4} (2^{n_\alpha} n_\alpha!)^{-1/2} . \tag{22}$$

In Eq. (21), H_{n_α} stands for a (conventional) Hermite polynomial.

B. Double polar coordinates

We may look for the wave functions of the $\mathbb{R}^4$ oscillator in terms of wave functions of a pair of isotropic harmonic oscillators in two dimensions, each of the wave functions of the two $\mathbb{R}^2$ oscillators being expressed in polar coordinates. Following standard procedures (as, for example, the one connected to Whittaker invariants[26,27]), we obtain for the first $\mathbb{R}^2$ oscillator (variables: u_1, u_2) the eigenstates

$$\begin{aligned} \epsilon(1) &= (2k_1 + |m_1| + 1)\lambda(\hbar^2/\mu) , \\ \Psi_{k_1m_1} &= N_{k_1m_1} [u_1 + i\,\mathrm{sgn}(m_1) u_2]^{|m_1|} \\ &\quad \times \exp[-(\lambda/2)(u_1^2 + u_2^2)] L^{|m_1|}_{k_1+|m_1|} \\ &\quad \times [\lambda(u_1^2 + u_2^2)] , \end{aligned} \tag{23}$$

where

$$\begin{aligned} N_{k_1m_1} &= \pi^{-1/2} \lambda^{(|m_1|+1)/2} (k_1!)^{1/2} \\ &\quad \times [(k_1 + |m_1|)!]^{-3/2} , \\ & k_1 \in \mathbf{N} , \; m_1 \in \mathbf{Z} . \end{aligned} \tag{24}$$

Remark that the polar coordinates

$$\rho_1 = (u_1^2 + u_2^2)^{1/2}, \qquad \varphi_1 = \arctan(u_2/u_1) \tag{25}$$

are easily seen to occur in Eq. (23) especially because

$$[u_1 + i\,\mathrm{sgn}(m_1) u_2]^{|m_1|} = \rho_1^{|m_1|} e^{im_1\varphi_1} . \tag{26}$$

Similar results are obtainable for the second $\mathbb{R}^2$ oscillator (variables: u_3, u_4) owing to the substitutions

$$\begin{gathered} \epsilon(1) \to \epsilon(2), \quad k_1 \to k_2, \quad m_1 \to m_2 , \\ u_1 \to u_3, \quad u_2 \to u_4 , \\ \rho_1 \to \rho_2, \quad \varphi_1 \to \varphi_2 . \end{gathered} \tag{27}$$

As a result for the $\mathbb{R}^4$ oscillator, the $C^{n_t}_{n_t+3}$ wave functions $\Psi_{k_1m_1k_2m_2}$ corresponding to the level $\epsilon_{n_t} = \epsilon(1) + \epsilon(2)$ with

$$n_t = 2k_1 + 2k_2 + |m_1| + |m_2| , \tag{28}$$

can be written down as

$$\Psi_{k_1m_1k_2m_2} = N_{k_1m_1} N_{k_2m_2} \exp[i(m_1\varphi_1 + m_2\varphi_2)] \rho_1^{|m_1|} \rho_2^{|m_2|} \exp[-(\lambda/2)(\rho_1^2 + \rho_2^2)] L^{|m_1|}_{k_1+|m_1|}(\lambda\rho_1^2) L^{|m_2|}_{k_2+|m_2|}(\lambda\rho_2^2) . \tag{29}$$

C. Hyperspherical coordinates

In the hyperspherical coordinates $(u, \psi, \theta, \varphi)$ defined through

$$\begin{aligned} u_1 &= u \sin\psi \sin\theta \cos\varphi , \\ u_2 &= u \sin\psi \sin\theta \sin\varphi , \\ u_3 &= u \sin\psi \cos\theta , \\ u_4 &= u \cos\psi , \end{aligned} \tag{30}$$

the wave functions of the $\mathbb{R}^4$ oscillator may be obtained from standard procedures (as, for example, the one connected to Whittaker invariants[26,27]). This yields the following eigenvalues:

$$\epsilon = (N + 2K + 1)\lambda(\hbar^2/\mu) , \quad N \in \mathbb{N} - \{0\}, \quad K \in \mathbb{N} . \tag{31}$$

Further, the $C^{n_t}_{n_t+3}$ wave functions Ψ_{NLMK}, corresponding to the level $\epsilon_{n_t} = \epsilon$ with

$$n_t = N + 2K - 1 , \tag{32}$$

are found to be

$$\begin{aligned} \Psi_{NLMK} &= N_{NLMK} u^{N-1} Y_{NLM}(\psi, \theta, \varphi) \\ &\quad \times e^{-(\lambda/2)u^2} L^{N}_{K+N}(\lambda u^2) , \end{aligned} \tag{33}$$

where

$$N_{NLMK} = 2^{1/2} \lambda^{(N+1)/2} (K!)^{1/2} [(K+N)!]^{-3/2} , \tag{34}$$

$$L = 0, 1, \ldots, N-1, \quad M = -L, -L+1, \ldots, L .$$

We adopt the definition of Refs. 28 and 29 for the hyperspherical harmonics Y_{NLM} appearing, in Eq. (33).

V. TRANSFORMATION FORMULAS

We are now in a position to develop any wave function (in a given system of coordinates) of the hydrogen atom as a linear combination of wave functions (in a given system of coordinates) of the four-dimensional oscillator. In a formal way, we may write

$$\Psi \text{ (hydrogen)} = \sum I\,[xxxxxxx]\ \Psi \text{ (oscillator)} , \tag{35}$$

where the expansion coefficients $I\,[xxxxxxx]$ depend on the systems of coordinates chosen for the hydrogen atom and the harmonic oscillator. Each coefficient $I\,[xxxxxxx]$ is thus a function of seven quantum numbers: three for the hydrogen and four for the oscillator. For a given choice of the systems of coordinates both for the hydrogen atom and the four-dimensional oscillator, the calculation of the expansion coefficients $I\,[xxxxxxx]$ may be achieved in two steps. First, we apply the KS transformation on Ψ(hydrogen) in order to produce a function on $L^2(\mathbb{R}^4)$. Second, we treat Eq. (35) as an equality in $L^2(\mathbb{R}^4)$ and use the orthonormality property of Ψ(oscillator) to obtain $I\,[xxxxxxx]$. In all cases, the latter equality may be simplified in view of the fact that

$$\alpha = \lambda/2 \quad \text{or} \quad n_t + 2 = 2n , \tag{36}$$

as can be seen from Eqs. (9), (11), (12), and (19). Furthermore, the definitions and relations

$$\begin{aligned} t_\alpha &= \lambda^{1/2} u_\alpha , \quad \alpha = 1, 2, 3, 4, \\ t^2 &= t_1^2 + t_2^2 + t_3^2 + t_4^2 , \\ \alpha\xi &= \lambda(u_1^2 + u_2^2) = t_1^2 + t_2^2 , \\ \alpha\eta &= \lambda(u_3^2 + u_4^2) = t_3^2 + t_4^2 , \\ \alpha(\xi + \eta)/2 &= (\lambda/2)(u_1^2 + u_2^2 + u_3^2 + u_4^2) \\ &= (\lambda/2)u^2 = (\lambda/2)r = \tfrac{1}{2}t^2 , \\ \xi\eta &= 4(u_1^2 + u_2^2)(u_3^2 + u_4^2) \\ &= (4/\lambda^2)(t_1^2 + t_2^2)(t_3^2 + t_4^2) , \\ \cos\theta &= (t_1^2 + t_2^2 - t_3^2 - t_4^2)/t^2 , \\ e^{i\varphi} &= (t_1 + it_2)(t_3 + it_4)/[(t_1^2 + t_2^2)(t_3^2 + t_4^2)]^{1/2} , \end{aligned} \tag{37}$$

which arise by applying the KS transformation to the parabolic and spherical coordinates of $\mathbb{R}^3$, are also of importance for handling the obtained equality. Finally, it is to be noted that in Eq. (35) the sum over the four quantum numbers for the oscillator is strongly limited by Eq. (36).

A. $\mathbb{R}^4$ double polar→$\mathbb{R}^3$ parabolic passage formulas

We expect the result for the coefficients $I\,[p_1p_2mk_1m_1k_2m_2]$ to be very simple (although the proof lies on a nonstraightforward point) since the Lie algebra of SU(2) ⊗ SU(2) enters the problem of the hydrogen atom in parabolic coordinates[30] and, on the other hand, the four-dimensional oscillator in double polar coordinates clearly exhibits an SU(2) ⊗ SU(2) symmetry. By introducing Eqs. (17) and (29) into Eq. (35), we end up with

$$\begin{aligned} &2^{|m|} N_{p_1p_2m} \rho_1^{|m|} \rho_2^{|m|} e^{im\varphi} L^{|m|}_{p_1+|m|}(\lambda\rho_1^2) L^{|m|}_{p_2+|m|}(\lambda\rho_2^2) \\ &\quad = \sum_{k_1m_1k_2m_2} I\,[p_1p_2mk_1m_1k_2m_2] N_{k_1m_1} N_{k_2m_2} \rho_1^{|m_1|} \rho_2^{|m_2|} e^{im_1\varphi_1} e^{im_2\varphi_2} L^{|m_1|}_{k_1+|m_1|}(\lambda\rho_1^2) L^{|m_2|}_{k_2+|m_2|}(\lambda\rho_2^2) , \end{aligned} \tag{38}$$

after making use of Eqs. (25), (27), and (37). The decisive point is to realize that

$$\varphi = \varphi_1 + \varphi_2 + 2k\pi, \quad k \in \mathbb{N} , \tag{39}$$

a result contained in Eqs. (25), (27), and (37) and in agreement with an alternative derivation by Chen and Kibler.[31] By combining Eqs. (38) and (39), we get $m_1 = m_2 = m$ from ordinary Fourier analysis. Then, the orthogonality property of the associated Laguerre polynomials shows that $k_1 = p_1$ and $k_2 = p_2$. As a final result, we have

$$\begin{aligned} I\,[p_1p_2mk_1m_1k_2m_2] &= \delta(k_1, p_1)\delta(k_2, p_2)\delta(m_1, m)\delta(m_2, m) \\ &\quad \times 2^{|m|}N_{p_1p_2m}/(N_{p_1m}N_{p_2m}), \qquad (40) \\ I\,[p_1p_2mp_1mp_2m] &= (-1)^b 2^{-1}\pi^{1/2}(\alpha/n)^{1/2}. \end{aligned}$$

In other words, the wave functions for the hydrogen atom in the SU(2) ⊗ SU(2) oscillator basis are essentially the same as those that arise using (rotational) parabolic coordinates (see also Refs. 8 and 31).

B. $\mathbb{R}^4$ Cartesian→$\mathbb{R}^3$ parabolic passage formulas

The coefficients $I\,[p_1p_2mn_1n_2n_3n_4]$ are obtained straightforwardly by introducing Eqs. (17) and (21) into Eq. (35). It is enough to use Eqs. (36) and (37) and the orthogonality property of the Hermite polynomials to get

$$\begin{aligned} I\,[p_1p_2mn_1n_2n_3n_4] &= 2^{-2}\alpha^{-|m|-2}N_{p_1p_2m}N_{n_1n_2n_3n_4}\int_{-\infty}^{+\infty\otimes 4}[t_1 + i\,\mathrm{sgn}(m)t_2]^{|m|}[t_3 + i\,\mathrm{sgn}(m)t_4]^{|m|} \\ &\quad \times L^{|m|}_{p_1+|m|}(t_1^2 + t_2^2)L^{|m|}_{p_2+|m|}(t_3^2 + t_4^2)\prod_{\alpha=1}^{4} e^{-t_\alpha^2}H_{n_\alpha}(t_\alpha)dt_\alpha\,. \end{aligned} \qquad (41)$$

Equation (41) can be rewritten as

$$\begin{aligned} I\,[p_1p_2mn_1n_2n_3n_4] &= 2^{-2}\alpha^{-|m|-2}N_{p_1p_2m}N_{n_1n_2n_3n_4} \\ &\quad \times J\,[p_1mn_1n_2]J\,[p_2mn_3n_4]\,, \end{aligned} \qquad (42)$$

where the integrals J are defined via

$$\begin{aligned} J\,[pmab] &= \int_{-\infty}^{+\infty\otimes 2}[x + i\,\mathrm{sgn}(m)y]^{|m|} \\ &\quad \times L^{|m|}_{p+|m|}(x^2 + y^2)e^{-(x^2+y^2)} \\ &\quad \times H_a(x)H_b(y)dx\,dy\,, \end{aligned} \qquad (43)$$

$$p\in\mathbb{N}, \quad m\in\mathbb{Z}, \qquad a\in\mathbb{N}, \quad b\in\mathbb{N}\,.$$

By using parity considerations, we may prove from Eqs. (41)-(43) that

$$I\,[p_1p_2mn_1n_2n_3n_4] = 0\,,$$

for (44)

$$n_1 + n_2 + |m| \quad \text{or} \quad n_3 + n_4 + |m| = \text{odd integers}\,.$$

We thus have the selection rules that $n_1 + n_2 + |m|$ and $n_3 + n_4 + |m|$ must be even integers. Consequently, we recover that $n_t\,(= n_1 + n_2 + n_3 + n_4)$ must be an even integer [cf. Eq. (36)], a well-known result.[2] In addition, we can derive the selection rules

$$n_1 + n_2 = 2p_1 + |m|, \qquad n_3 + n_4 = 2p_2 + |m|\,, \qquad (45)$$

which give back the previous selection rules.

Finally from Eq. (41) we obtain the symmetry relation

$$\begin{aligned} (I\,[p_1p_2mn_1n_2n_3n_4]/N_{p_1p_2m})^* &= I\,[p_1p_2 - mn_1n_2n_3n_4]/N_{p_1p_2-m}\,. \end{aligned} \qquad (46)$$

C. $\mathbb{R}^4$ Cartesian→$\mathbb{R}^3$ spherical passage formulas

The coefficients $I\,[n\,l\,mn_1n_2n_3n_4]$ immediately follow by combining Eqs. (13), (14), (21), and (35)-(37) and by using the orthogonality property of the Hermite polynomials. This leads to

$$\begin{aligned} I\,[n\,l\,mn_1n_2n_3n_4] &= (2\alpha)^{-2}N_{n\,l\,m}N_{n_1n_2n_3n_4}\int_{-\infty}^{+\infty\otimes 4} L^{2l+1}_{n+l}(t_1^2 + t_2^2 + t_3^2 + t_4^2) \\ &\quad \times \mathrm{YKS}_{l\,m}(t_1, t_2, t_3, t_4)\prod_{\alpha=1}^{4} e^{-t_\alpha^2}H_{n_\alpha}(t_\alpha)dt_\alpha\,. \end{aligned} \qquad (47)$$

The special function $\mathrm{YKS}_{l\,m}$ in Eq. (47) is defined in the following way. Let $\mathrm{YKS}_{l\,m}\,(u_1, u_2, u_3, u_4)$ be the function obtained by applying the KS transformation to the $\mathbb{R}^3$ harmonic polynomial $r^l\,Y_{l\,m}\,(\theta, \varphi)$. Clearly, $\mathrm{YKS}_{l\,m}\,(u_1, u_2, u_3, u_4)$ is an $\mathbb{R}^4$ harmonic polynomial of degree $2l$ (cf. Ref. 32). In Eq. (47), we then have

$$\mathrm{YKS}_{l\,m}(t_1, t_2, t_3, t_4) = \lambda^{\,l}\mathrm{YKS}_{l\,m}(u_1, u_2, u_3, u_4) \qquad (48)$$

that stands for the image of $(2\,\alpha)^l r^l Y_{l\,m}\,(\theta, \varphi)$ via the KS transformation.

By using parity considerations, we can show from Eqs. (37) and (47) that

$$I\,[n\,l\,mn_1n_2n_3n_4] = 0\,,$$

for (49)

$$n_1 + n_2 + m \quad \text{or} \quad n_3 + n_4 + m = \text{odd integers}\,.$$

Therefore, we obtain the selection rules that $n_1 + n_2 + m$ and $n_3 + n_4 + m$ must be even integers, from which we recover again that $n_t\,(= n_1 + n_2 + n_3 + n_4)$ must be an even integer [cf. Eq. (36)].

Finally, from Eq. (47) we obtain the symmetry relation

$$(I\,[n\,l\,mn_1n_2n_3n_4]/N_{n\,l\,m})^* = (-1)^m I\,[n\,l - mn_1n_2n_3n_4]/N_{nl-m}\,. \qquad (50)$$

D. $\mathbb{R}^4$ hyperspherical→$\mathbb{R}^3$ spherical passage formulas

A preliminary relation involving the coefficients $I[n\,l\,m\,N\,L\,M\,K]$ can be set up by introducing Eqs. (13) and (33) into Eq. (35) and by using the orthonormality property of the hyperspherical harmonics. This yields

$$(2\alpha)^{(N-1)/2}A\,[l\,m\,N\,L\,M\,]N_{n\,l\,m}t^{2l+1}L^{2l+1}_{n+l}(t^2)$$
$$=\sum_K I\,[n\,l\,m\,N\,L\,M\,K\,]N_{NLMK}t^{N}L^{N}_{K+N}(t^2)\,, \quad (51)$$

where the integral A is defined by

$$A\,[l\,m\,N\,L\,M\,]$$
$$=\int_0^{\pi}\int_0^{\pi}\int_0^{2\pi}Y_{NLM}(\psi,\theta,\varphi)^*u^{-2l}$$
$$\times\mathrm{YKS}_{lm}(u_1,u_2,u_3,u_4)\sin^2\psi\sin\theta\,d\psi\,d\theta\,d\varphi\,. \quad (52)$$

In Eq. (52), $u^{-2l}\mathrm{YKS}_{l\,m}(u_1,u_2,u_3,u_4)$ should be understood as a function of the angular variables ψ, θ, φ [cf. Eq. (30)]. Thus, A is a basic integral of a completely angular (or geometrical) nature. Next, we apply to Eq. (51) the orthogonality property of the associated Laguerre polynomials to obtain

$$I\,[n\,l\,m\,N\,L\,M\,K\,]$$
$$=(2\alpha)^{(N-1)/2}(N_{n\,l\,m}/N_{N\,L\,M\,K})$$
$$\times A\,[l\,m\,N\,L\,M\,]R\,[n\,l\,N\,K\,]\,, \quad (53)$$

where the radial integral R is given by

$$R\,[n\,l\,N\,K\,]=K!\,[(K+N)!]^{-3}$$
$$\times\int_0^{\infty}L^{N}_{K+N}(x)L^{2l+1}_{n+l}(x)$$
$$\times x^{(2l+1+N)/2}e^{-x}\,dx\,. \quad (54)$$

The last point amounts to use some trivial properties of the functions $\mathrm{YKS}_{l\,m}$ (cf. Ref. 32). As a consequence of the development of $u^{-2l}\mathrm{YKS}_{l\,m}$ in standard hyperspherical harmonics Y_{NLM}, we may prove that $A\,[l\,m\,N\,L\,M]=0$ if $N\neq 2\,l+1$. The introduction of the selection rule $N=2\,l+1$ into Eq. (54) shows in turn that $R\,[n\,l\,N\,K]=0$ if $K\neq n-l-1$, and we thus have a further selection rule, viz., $K=n-l-1$. The latter two selection rules ensure that we obtain

$$I\,[n\,l\,m\,N\,L\,M\,K\,]$$
$$=\delta(N,2l+1)\delta(K,n-l-1)$$
$$\times(2\alpha)^{l}(N_{n\,l\,m}/N_{N\,L\,M\,K})A\,[l\,m\,N\,L\,M\,]\,, \quad (55)$$

since $R\,[n,l,2l+1,n-l-1]=1$.

Finally, we note the symmetry relation

$$(I\,[n\,l\,m\,N\,L\,M\,K\,]/N_{n\,l\,m})^*$$
$$=(-1)^{L-M+m}I\,[n\,l-m\,N\,L-M\,K\,]/N_{n\,l-m}\,. \quad (56)$$

E. $\mathbb{R}^4$ hyperspherical→$\mathbb{R}^3$ parabolic passage formulas

The coefficients $I\,[p_1p_2m\,N\,L\,M\,K]$ can be obtained by combining Eqs. (17), (33), and (35). We formally obtain

$$I\,[p_1p_2m\,N\,L\,M\,K\,]=2^{|m|}(2\alpha)^{-(N+2|m|+3)/2}N_{p_1p_2m}N_{NLMK}\int_{-\infty}^{+\infty\otimes 4}[t_1+i\,\mathrm{sgn}(m)t_2]^{|m|}L^{|m|}_{p_1+|m|}(t_1^2+t_2^2)$$
$$\times[t_3+i\,\mathrm{sgn}(m)t_4]^{|m|}L^{|m|}_{p_2+|m|}(t_3^2+t_4^2)(t_1^2+t_2^2+t_3^2+t_4^2)^{(N-1)/2}Y_{NLM}(\psi,\theta,\varphi)^*$$
$$\times L^{N}_{K+N}(t_1^2+t_2^2+t_3^2+t_4^2)\exp[-(t_1^2+t_2^2+t_3^2+t_4^2)]dt_1\,dt_2\,dt_3\,dt_4\,, \quad (57)$$

where $t^{N-1}Y_{NLM}(\psi,\theta,\varphi)$ should be considered as a function of the dimensionless variables t_1, t_2, t_3, t_4.

E. $\mathbb{R}^4$ double polar→$\mathbb{R}^3$ spherical passage formulas

The coefficients $I\,[n\,l\,m\,k_1\,m_1\,k_2\,m_2]$ can be obtained by combining Eqs. (13), (29), and (35). We formally obtain

$$I\,[n\,l\,m\,k_1\,m_1\,k_2\,m_2]=(2\alpha)^{-(|m_1|+|m_2|+4)/2}N_{n\,l\,m}N_{k_1\,m_1}N_{k_2\,m_2}\int_{-\infty}^{+\infty\otimes 4}[t_1-i\,\mathrm{sgn}(m_1)t_2]^{|m_1|}L^{|m_1|}_{k_1+|m_1|}(t_1^2+t_2^2)$$
$$\times[t_3-i\,\mathrm{sgn}(m_2)t_4]^{|m_2|}L^{|m_2|}_{k_2+|m_2|}(t_3^2+t_4^2)\mathrm{YKS}_{l\,m}(t_1,t_2,t_3,t_4)L^{2l+1}_{n+l}(t_1^2+t_2^2+t_3^2+t_4^2)$$
$$\times\exp[-(t_1^2+t_2^2+t_3^2+t_4^2)]dt_1\,dt_2\,dt_3\,dt_4\,. \quad (58)$$

Arguments similar to the ones used in Sec. V A lead to the selection rules $m_1=m_2=m$ so that Eq. (58) may be simplified in view of

$$I\,[n\,l\,m\,k_1\,m_1\,k_2\,m_2]=\delta(m_1,m)\delta(m_2,m)I\,[n\,l\,m\,k_1\,m\,k_2\,m]\,. \quad (59)$$

Equations (58) and (59) may be worked out to lead to a symmetrical form for the coefficients $I\,[n\,l\,m\,k_1\,m\,k_2\,m]$. Indeed, by transforming the relevant fourfold integral into a twofold integral, we have

$$I\,[n\,l\,m\,k_1\,m\,k_2\,m] = 2^{-|m|-1}\pi^{3/2}\alpha^{-|m|-2}(-1)^m[(2l+1)(l-m)!/(l+m)!]^{1/2}N_{n\,l\,m}N_{k_1m}N_{k_2m}K\,[n\,l\,m\,k_1\,k_2]\,, \tag{60}$$

with

$$K\,[n\,l\,m\,k_1\,k_2] = \int_0^{\infty\,\otimes 2} (xy)^{|m|+1}L^{|m|}_{k_1+|m|}(x^2)L^{|m|}_{k_2+|m|}(y^2)$$
$$\times P_l^m[(x^2-y^2)/(x^2+y^2)]L^{2l+1}_{n+l}(x^2+y^2)(x^2+y^2)^l e^{-(x^2+y^2)}\,dx\,dy\,, \tag{61}$$

where P_l^m is an associated Legendre function.

VI. CLOSING REMARKS ON COMPUTING ASPECTS

To close this paper, we present some developments concerning computing aspects of this work.

A. A MACSYMA approach

The expansion coefficients $I\,[xxxxxxx]$ for low values for the seven quantum numbers corresponding to a given transformation can be easily calculated with the help of the algebraic and symbolic programming system MACSYMA. We have written a program to handle all the basic integrals appearing in this paper. Therefore, the expansions of the type of Eq. (35) can be computer generated in an algebraic (rather than numeric) fashion. By way of illustration, we consider the development of the spherical wave function Ψ_{nlm} with $n=3$, $l=2$, and $m=0$ for the hydrogen atom in terms of Cartesian wave function $\Psi_{n_1n_2n_3n_4}$ for the harmonic oscillator (cf. Sec. V. C). (The corresponding expansion coefficients I $[320n_1n_2n_3n_4]$ are far from being easily obtainable by hand.) Our program gives the expansion

$$\Psi_{320} = 2^{-3}3^{-1/2}\pi^{1/2}\alpha^{1/2}[\Psi_{4000}+\Psi_{0400}+\Psi_{0040}$$
$$+\Psi_{0004} - 2^{3/2}3^{-1/2}(\Psi_{2020}+\Psi_{2002}+\Psi_{0220}$$
$$+\Psi_{0202} - 2^{-1}\Psi_{2200} - 2^{-1}\Psi_{0022})]\,. \tag{62}$$

We note that the selection rules (49) are satisfied in Eq. (62). [It should be realized that the nonobservation of the selection rules (44) and/or (49) in some recent works has lead to errors.] The reader interested in other expansions may write to the authors.

B. Basic integrals

The integrals [see Eqs. (43), (47), (52), (57), (58), and (61)] encountered in this paper are central to the theory of special functions. As a first example, the angular integral A [see Eq. (52)] corresponds to the development of YKS_{lm} in terms of hyperspherical harmonic polynomials $u^{N-1} \times Y_{NLM}$. As a second example, Eq. (43) corresponds to connection formulas between the product of Hermite polynomials $H_a(x)H_b(y)$ and the Laguerre polynomial $L^{|m|}_{p+|m|}(x^2+y^2)$, formulas that are (more or less) known both from a physical[9] and mathematical[33] viewpoint.

C. Relations between expansion coefficients

Finally, we note that the various expansion coefficients $I[xxxxxxx]$ discussed in this paper are not all independent. As a matter of fact, the spherical and parabolic wave functions in $L^2(\mathbb{R}^3)$ for the hydrogen atom are connected by a relationship of type

$$\Psi_{p_1p_2m} = \sum\,(n\,l\,m|p_1p_2m)\Psi_{n\,l\,m}\,, \tag{63}$$

where $(n\,l\,m|p_1p_2\,m)$ is essentially a $\mathrm{SU}(2)\supset\mathrm{U}(1)$ Clebsch–Gordan coefficient (cf. Ref. 25). Consequently, given a coordinate system for the four-dimensional oscillator, the corresponding expansion coefficients $I[p_1p_2mxxxx]$ and $I\,[nlmxxxx]$ are related by

$$I\,[p_1p_2mxxxx] = \sum\,(nlm|p_1p_2m)I\,[nlmxxxx]\,, \tag{64}$$

which is the relation dual to Eq. (63).

We close this paper by noticing that nothing has been said on the use of the dynamic symmetry group O(4) for deriving the Green's function for the Coulomb field in the spirite of the works initiated by Schwinger[34] and further developed by Bander and Itzykson.[35] The reader should consult Refs. 9–14 for developments along this line.

ACKNOWLEDGMENT

One of the authors (M. K.) wishes to thank Dr. C. C. Gerry for sending him his work on coherent states and the Kepler–Coulomb problem[36] prior to publication. In this regard, the authors would like to emphasize that the present work could have an impact on the work of Ref. 36.

[1] M. Ikeda and Y. Miyachi, Math. Jpn. **15**, 127 (1970).

[2] M. Boiteux, C. R. Acad. Sci. Ser. B **274**, 867 (1972). See also Physica **65**, 381 (1973) and J. Math. Phys. **23**, 1311 (1982).

[3] A. O. Barut, C. K. E. Schneider, and R. Wilson, J. Math. Phys. **20**, 2244 (1979). See also A. O. Barut and I. H. Duru, Proc. R. Soc. London Ser. A **333**, 217 (1973), where the two-dimensional Kepler and oscillator problems are related.

[4] A. C. Chen, Phys. Rev. A **22**, 333, 2901(E) (1980); **25**, 2409 (1982). See also J. Math. Phys. **23**, 412 (1982).

[5] T. Iwai, J. Math. Phys. **23**, 1093 (1982).

[6] J. Kennedy, Proc. R. Ir. Acad. Sect. A **82**, 1 (1982).

[7] H. Grinberg, J. Marañón, and H. Vucetich, J. Math. Phys. **25**, 2648 (1984).

[8] F. H. J. Cornish, J. Phys. A **17**, 323 (1984).

[9] I. H. Duru and H. Kleinert, Phys. Lett. B **84**, 185 (1979); Fortschr. Phys. **30**, 401 (1982).

[10] G. A. Ringwood and J. T. Devreese, J. Math. Phys. **21**, 1390 (1980).

[11] R. Ho and A. Inomata, Phys. Rev. Lett. **48**, 231 (1982). See also A. Inomata, Phys. Lett. A **101**, 253 (1984).

[12] H. Grinberg, J. Marañón, and H. Vucetich, J. Chem. Phys. **78**, 839 (1983); Int. J. Quantum Chem. **23**, 379 (1983); Z. Phys. C **20**, 147 (1983).

[13] N. K. Pak and I. Sökmen, Phys. Lett. A **100**, 327 (1984).

[14] F. Steiner, Phys. Lett. A **106**, 363 (1984).

[15] J. M. Gracia-Bondía, Phys. Rev. A **30**, 691 (1984).

[16] M. Kibler and T. Négadi, Lett. Nuovo Cimento **37**, 225 (1983); J. Phys. A **16**, 4265 (1983); Phys. Rev. A **29**, 2891 (1984).

[17] P. Kustaanheimo and E. Stiefel, J. Reine Angew. Math. **218**, 204 (1965).

[18] M. Kibler and T. Négadi, Lett. Nuovo Cimento **39**, 319 (1984).

[19] H. A. Cerdeira, J. Phys. A **18**, 2719 (1985).

[20] D. Delande and J. C. Gay, J. Phys. B **17**, L 335 (1984).

[21] A. C. Chen, Phys. Rev. A **31**, 2685 (1985).

[22] L. I. Schiff, *Quantum Mechanics* (McGraw-Hill, New York, 1955).

[23] H. A. Bethe and E. E. Salpeter, *Quantum Mechanics of One- and Two-Electron Atoms* (Springer, Berlin, 1957).

[24] E. G. Kalnins, W. Miller, Jr., and P. Winternitz, SIAM J. Appl. Math. **30**, 630 (1976). See also P. Winternitz, in *Group Theoretical Methods in Physics*, edited by R. T. Sharp and B. Kolman (Academic, New York, 1977), p. 549.

[25] K. Fujio, T. Maekawa, and S. Nagae, Lett. Nuovo Cimento **20**, 25 (1977). See also S. Nagae, Kumamoto Phys. Rep. **3**, 91 (1978).

[26] A. K. Bose, Phys. Lett. **7**, 245 (1963); Nuovo Cimento **32**, 679 (1964); A. Lemieux and A. K. Bose, Ann. Inst. H. Poincaré Sect. A **10**, 259 (1969).

[27] A. Nikiforov and V. Ouvarov, *Eléments de la Théorie des Fonctions Spéciales* (MIR, Moscow, 1976).

[28] L. C. Biedenharn, J. Math. Phys. **2**, 433 (1961).

[29] T. Shibuya and C. E. Wulfman, Proc. R. Soc. London Ser. A **286**, 376 (1965).

[30] F. Ravndal and T. Toyoda, Nucl. Phys. B **3**, 312 (1967).

[31] A. C. Chen and M. Kibler, Phys. Rev. A **31**, 3960 (1985).

[32] M. Kibler, T. Négadi, and A. Ronveaux, in *Polynômes orthogonaux et applications*, edited by C. Brezinski, A. Draux, A. P. Magnus, P. Maroni, and A. Ronveaux (Springer, Berlin, 1985).

[33] D. Colton and J. Wimp, Complex Variables **3**, 397 (1984).

[34] J. Schwinger, J. Math. Phys. **5**, 1606 (1964).

[35] M. Bander and C. Itzykson, Rev. Mod. Phys. **38**, 330 (1966).

[36] C. C. Gerry, "Coherent states and the Kepler–Coulomb problem," submitted to Phys. Rev. A.

Comment on momentum in stochastic mechanics

Simon Golin
Universität Bielefeld, Fakultät für Physik and Forschungszentrum Bielefeld-Bochum-Stochastik, D-4800 Bielefeld 1, Federal Republic of Germany

(Received 4 June 1985; accepted for publication 14 January 1986)

Stochastic mechanics is a probabilistic description of quantum systems in terms of stochastic differential equations. Davidson [M. Davidson, Lett. Math. Phys. **5**, 523 (1981)] and de Falco, De Martino, and De Siena [D. de Falco, S. De Martino, and S. De Siena, Lett. Nuovo Cimento **36**, 457 (1983)] have introduced momentum variables into this scheme. In this paper a discussion of this attempt is presented and some difficulties concerning the physical interpretation are pointed out.

I. INTRODUCTION

The theory of stochastic mechanics[1-4] provides an alternate mathematical—and possibly physical—representation of nonrelativistic quantum mechanics. This probabilistic quantization procedure has as its basic underlying mathematical object a diffusion process associated to the motion of a quantum mechanical particle. The elevated role of position (or, more generally, of any configurational variable) is in contrast to the L^2 formalism of conventional quantum mechanics, where one treats space and momentum coordinates on the same footing. It seems of interest to understand whether stochastic mechanics allows for an analog of the canonical transformation theory of classical and quantum mechanics.

Guerra and Morato[5] have attempted to approach the position–momentum complementarity in the stochastic frame and managed to deal with the coherent states of the harmonic oscillator. By exploiting the symmetry (in position and momentum) of the Hamiltonian they could construct diffusions associated to momentum. Their strategy, however, does not seem to be capable of generalization to other potentials.

The question of momentum is treated in a completely different fashion by Davidson[6] and de Falco, De Martino, and De Siena.[7] They make use of the asymptotic behavior of the trajectories in the stochastic mechanics of a free particle. In this way one is able to define momentum random variables for a general class of potentials, and their distributions coincide with those of quantum mechanical momentum.

The concern of the present paper is to examine this implementation of momentum in stochastic mechanics. The momentum process is found to be non-Markovian. It reveals some serious unphysical features: First of all, the momentum variables do not meet the requirement that momentum ought to have an operational meaning in physics. Second, the time derivative (if it exists) does not yield force. Furthermore, the definition is so implicit that it is of no use in the derivation of uncertainty relations in stochastic mechanics.

The organization of the paper is as follows. In Sec. II the basic notions of stochastic mechanics are introduced. Then the momentum process is defined in Sec. III, accompanied by the example of the harmonic oscillator ground state in Sec. IV. The phase space formulation of quantum mechanics is reviewed in Sec. V. In Sec. VI we analyze the physics behind the momentum process and we conclude in Sec. VII.

II. CORRESPONDENCE BETWEEN STOCHASTIC MECHANICS AND QUANTUM MECHANICS

The formulation of stochastic mechanics goes back to Nelson.[1] This section is devoted to a brief review of the correspondence between stochastic mechanics and conventional quantum mechanics. For a detailed exposition we refer to Refs. 1–4.

Let us consider a point particle of mass m moving under the influence of a potential $V(x)$. In stochastic mechanics the kinematical aspects of the motion are described by a Markovian diffusion (with values in the configuration space) solution of the stochastic differential equation

$$d\xi_t = b(\xi_t,t)dt + dw_t, \tag{1}$$

where w_t denotes the Wiener process with variance 2ν (independent of ξ_0). The probability density $\rho(x,t)$ of the process ξ_t connects the forward drift $b(x,t)$ to the backward drift $b_*(x,t)$:

$$b_*(x,t) = b(x,t) - 2\nu\nabla \ln \rho(x,t). \tag{2}$$

The drifts simply represent the mean forward and backward velocity of ξ_t, respectively. The osmotic velocity $u(x,t)$ and the current velocity $v(x,t)$ are defined by

$$u(x,t) := \tfrac{1}{2}[b(x,t) - b_*(x,t)] = \nu\nabla \ln \rho(x,t), \tag{3}$$

$$v(x,t) := \tfrac{1}{2}[b(x,t) + b_*(x,t)]. \tag{4}$$

The dynamics has to specify the influence of the potential $V(x)$. This can be accomplished, e.g., by the Guerra–Morato variational principle.[3,8] (An enlightening presentation of this principle can be found in Ref. 9.) It relates the solution of the Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\psi(x,t) = \left[-\frac{\hbar^2}{2m}\Delta + V(x)\right]\psi(x,t) \tag{5}$$

to the diffusion ξ_t, where $\hbar$ is Planck's constant divided by 2π, and one finds

$$u(x,t) = (\hbar/m)\mathrm{Re}\,\nabla \ln \psi(x,t), \tag{6}$$

$$v(x,t) = (\hbar/m)\mathrm{Im}\,\nabla \ln \psi(x,t). \tag{7}$$

The diffusion constant ν turns out to be equal to $\hbar/2m$, and the probability density of the process is related to the (normalized) solution of the Schrödinger equation by

$$\rho(x,t) = |\psi(x,t)|^2. \tag{8}$$

0022-2488/86/061549-07$02.50

III. MOMENTUM PROCESS

In stochastic mechanics there are several stochastic processes whose mean values coincide with the expectation of the quantum mechanical momentum operator P. For instance, we can take the forward and backward drifts or the current velocity:

$$E\,[b(\xi_t,t)] = E\,[b_*(\xi_t,t)] = E\,[v(\xi_t,t)] \\ = \langle \psi(\cdot,t), P\psi(\cdot,t)\rangle. \tag{9}$$

But none of these random variables has the same distribution as the operator P. Already their variances differ from those of P.

The definition of momentum random variables by Davidson[6] and de Falco, De Martino, and De Siena[7] is based essentially on work by Shucker,[10] who has investigated the behavior of the trajectories of the diffusion ξ^f corresponding to the free Schrödinger equation (i.e., in the absence of any potential).

Shucker showed that in the free case the limit

$$\lim_{T\to\infty} (m\xi^f_T/T) \tag{10}$$

exists pointwise almost surely (under moderate technical assumptions) and has probability density equal to the quantum mechanical momentum distribution. Of course, in the free case, the momentum distribution is time independent. Incidentally, Shucker's result has been generalized to the interacting case. Biler[11] considered the one-dimensional case, and the case of three dimensions was treated by Serva[12] for central potentials and by Carlen[13] for potentials of the Kato–Rellich type.

Now we consider a situation where a potential is present. Let ξ be the corresponding position process. Consider also the solution $\psi^{f,t}$ of the free Schrödinger equation with initial condition at time t being given by the interacting wave function ψ at time t:

$$\psi^{f,t}(x,t) = \psi(x,t). \tag{11}$$

This leads to the free position process $\xi^{f,t}$ given by

$$d\xi^{f,t}_T = b^{f,t}(\xi^{f,t}_T,T)dT + dw^{f,t}_T, \tag{12}$$

where $w^{f,t}$ is a Wiener process with variance 2ν (independent of $\xi^{f,t}_t$).

In particular, we can impose Davidson's[6] "by fiat" assumptions

$$\xi^{f,t}_t = \xi_t, \tag{13a}$$

$$w^{f,t}_T = w_T. \tag{13b}$$

On account of (13a), the process $\xi^{f,t}$ can be thought of as being "tangent" to the process ξ at time t.

Following Davidson[6] and de Falco, De Martino, and De Siena,[7] we define

$$\pi_t := \lim_{T\to\infty} (m\xi^{f,t}_T/T). \tag{14}$$

According to Shucker's analysis this limit exists, and it has a probability density equal to the momentum distribution of the quantum state $\psi(\cdot,t)$.

Thus, in the case of arbitrary potential, a random variable has been constructed whose distribution coincides with the momentum distribution in quantum mechanics. This property, of course, if not sufficient to define π_t uniquely, e.g., we could have taken the limit $T\to-\infty$ in the definition of π_t. The resulting random variable has the same distribution. According to a result of Nelson,[3] however, the two limits differ.

We want to elaborate on the last point a little further. Nelson considered a Gaussian wave packet under the free evolution and computed the correlation matrix of the initial momentum ($T\to-\infty$) and the final momentum ($T\to+\infty$), and found it to be $-e^{-\pi}\mathbf{1}$ (independently of the width of the Gaussian). Therefore the two momenta differ (although their densities coincide) and this shows the difficulty of defining a pathwise analog of the scattering matrix in stochastic mechanics. Similarly, if one computes the correlation coefficient of the squares of the momenta, it turns out to be equal to $-e^{-2\pi}$. Hence in stochastic mechanics there is no pathwise energy conservation, i.e., the trajectories of the position process do not exhibit elastic scattering.

We also point out that the definition of the momentum process π_t depends on the state of the quantum system, as does the position process π_t. In contrast to ξ_t, the process π_t will generally not be a diffusion. Moreover, as we shall see, π_t will turn out to be non-Markovian.

Before entering into a discussion of the momentum process, we consider the ground state of the harmonic oscillator as an explicitly calculable example.

IV. HARMONIC OSCILLATOR GROUND STATE

In order to assess the scope of the momentum variables π_t, the ground state of the one-dimensional harmonic oscillator will be discussed. For this specific example π_t can be determined rather explicitly.[6] We point out that the free case is also tractable if one starts with a Gaussian wave packet as initial condition.[7]

The ground state of the harmonic oscillator is given by

$$\psi(x,t) = (2\pi\sigma^2)^{-1/4}\exp\{-\tfrac{1}{2}(i\omega t + x^2/2\sigma^2)\}, \tag{15}$$

where $\sigma^2 := \hbar/2m\omega$ $(\omega > 0)$ is the variance of position. The drift is consequently given by

$$b(x,t) = -\omega x \tag{16}$$

and the position process obeys

$$d\xi_t = -\omega\xi_t\,dt + dw_t, \tag{17}$$

$$\xi_t = e^{-\omega t}\Big(\xi_0 + \int_0^t e^{\omega z}\,dw_z\Big). \tag{18}$$

Next we consider the free particle solution with initial condition

$$\psi^{f,t}(x,t) = \psi(x,t), \tag{19}$$

i.e.,

$$\psi^{f,t}(x,T) = (2\pi\sigma^2)^{-1/4}[1 + i\omega(T-t)]^{-1/2} \\ \times\exp\{-\tfrac{1}{2}[i\omega t + x^2/2\sigma^2[1 + i\omega(T-t)]]\}, \tag{20}$$

$$b^{f,t}(x,T) = -\omega\,\frac{1+\omega(T-t)}{1+\omega^2(T-t)^2}\,x = -\dot\gamma(T-t)x, \tag{21}$$

where we have set

$$\gamma(t) : = \arctan \omega t - \tfrac{1}{2}\ln(1+\omega^2 t^2). \tag{22}$$

As a result of this, we find

$$d\xi_T^{f,t} = -\dot{\gamma}(T-t)\xi_T^{f,t}\,dT + dw_T^{f,t}, \tag{23}$$

$$\xi_T^{f,t} = e^{-\gamma(T-t)}\left[\xi_t^{f,t} + \int_t^T e^{\gamma(z-t)}\,dw_z^{f,t}\right], \tag{24}$$

$$\pi_t = m\omega e^{-\pi/2}\left[\xi_t^{f,t} + \int_t^\infty e^{\gamma(z-t)}\,dw_z^{f,t}\right]. \tag{25}$$

So we have obtained a direct form of the momentum variables. Another formula may be found on application of Itô's lemma:

$$\begin{aligned} d\,[e^{\gamma(z-t)}\xi_z] &= e^{\gamma(z-t)}\,d\xi_z + \dot{\gamma}(z-t)e^{\gamma(z-t)}\xi_z\,dz \\ &= [\dot{\gamma}(z-t) - \omega]e^{\gamma(z-t)}\xi_z\,dz \\ &\quad + e^{\gamma(z-t)}\,dw_z. \end{aligned} \tag{26}$$

It follows from (18) and (22) that

$$\lim_{T\to\infty} e^{\gamma(T-t)}\xi_T = (1/\omega)e^{\pi/2}\lim_{T\to\infty}(\xi_T/T) = 0, \tag{27}$$

$$\xi_t = \int_t^\infty [\omega - \dot{\gamma}(z-t)]e^{\gamma(z-t)}\xi_z\,dz - \int_t^\infty e^{\gamma(z-t)}\,dw_z. \tag{28}$$

If we now also impose Davidson's "by fiat" conditions (13), then we get another formula for the momentum process,

$$\pi_t = m\omega e^{-\pi/2}\int_t^\infty [\omega - \dot{\gamma}(z-t)]e^{\gamma(z-t)}\xi_z\,dz. \tag{29}$$

This representation of π_t makes it clear that π_t has two (but no more) continuous derivatives. We use

$$\dot{\gamma}(t) = \omega[(1-\omega t)/(1+\omega^2 t^2)], \tag{30}$$

$$\ddot{\gamma}(t) = \omega^2[(\omega^2 t^2 - 2\omega t - 1)/(1+\omega^2 t^2)^2] \tag{31}$$

and obtain

$$\dot{\pi}_t = m\omega e^{-\pi/2}\int_t^\infty \frac{\partial}{\partial t}[(\omega - \dot{\gamma}(z-t))e^{\gamma(z-t)}]\xi_z\,dz, \tag{32}$$

$$\ddot{\pi}_t = m\omega e^{-\pi/2}\left\{\int_t^\infty \frac{\partial^2}{\partial t^2}[(\omega - \dot{\gamma}(z-t))e^{\gamma(z-t)}]\xi_z\,dz - \omega^2\xi_t\right\}. \tag{33}$$

We will come back to the results obtained and discuss them at some length in Sec. VI.

V. PHASE SPACE FORMULATION OF QUANTUM MECHANICS

Now the possibility of formulating quantum mechanics in the phase space of position and momentum will be reviewed briefly. In the so-called phase space formulation of quantum mechanics one considers a function $F(x,p)$, which, in the classical limit, is expected to converge to the classical phase space density. (We drop the time dependence in this section.) The existence and properties of such distribution functions are closely related to the possibility of formulating quantum mechanics in terms of classical concepts.[14] But this is exactly what stochastic mechanics claims to achieve, or as Nelson[3] puts it: "Stochastic mechanics attempts to provide a realistic, objective description of physical events in classical terms."

We will relate the joint probability density of the process (ξ_t,π_t) to the phase space formulation of quantum mechanics in Sec. VI.

The first analysis into the direction of quantum distribution functions goes back to Wigner in 1932.[15] Since then this approach has been studied by many authors, both for conceptual reasons and in order to gain an effective means of calculating quantities that are not easily obtainable otherwise. The phase space density $F(x,p)$ is required to yield the proper quantum mechanical marginals (all integrations, unless otherwise noted, are to be extended over the whole space):

$$\int F(x,p)dp = |\psi(x)|^2, \tag{34a}$$

$$\int F(x,p)dx = |\hat{\psi}(p)|^2, \tag{34b}$$

where ψ and $\hat{\psi}$ are the configuration and momentum wave functions, respectively. A milder form of conditions (34) is to require that $F(x,p)$ should give the quantum mechanical expectations for functions depending on one of the phase space variables:

$$\int\int F(x,p)[g_1(x) + g_2(p)]dx\,dp = \int \psi^*(x)\left[g_1(x) + g_2\left(\frac{\hbar}{i}\frac{\partial}{\partial x}\right)\right]\psi(x)dx. \tag{35}$$

Wigner has shown[14] that there is no non-negative distribution function satisfying (35) such that $F(x,p)$ is a Hermitian form in ψ. Thus we cannot have Hermiticity in stochastic mechanics.

Cohen[16] has given an explicit representation of all quantum mechanical distributions satisfying (34):

$$F(x,p) = \int\int\int e^{i(\theta u - \theta x - \tau p)}f(\theta,\tau) \times\psi^*\left(u - \frac{\hbar\tau}{2}\right)\left(u + \frac{\hbar\tau}{2}\right)d\theta\,d\tau\,du, \tag{36}$$

where f is any function satisfying

$$f(0,\tau) = f(\theta,0) = 1. \tag{37}$$

For $f \equiv 1$, one gains the Wigner distribution. The characteristic function $M(\theta,\tau)$ of $F(x,p)$ is defined by

$$M(\theta,\tau) = \int\int e^{i(\theta x + \tau p)}F(x,p)dx\,dp = f(\theta,\tau)\int e^{i\theta u}\psi^*\left(u - \frac{\hbar\tau}{2}\right)\psi\left(u + \frac{\hbar\tau}{2}\right)du. \tag{38}$$

The quantum mechanical distribution function $F(x,p)$ may in general assume negative values (as the Wigner distribution does), or it even may be complex valued. Of course, in stochastic mechanics we deal with proper probability distributions.

As indicated by Eq. (35) the distribution function F allows us to calculate expectations of quantum mechanical observables in a probabilistic manner rather than through the operator formalism of conventional quantum mechanics. Let $g(x,p)$ be a classical function of position and momentum and denote by $G(X,P)$ the corresponding quantum operator. We would like, of course, for F to yield

$$\langle\psi, G(X,P)\psi\rangle = \int \psi^*(x)G(X,P)\psi(x)dx$$
$$= \int g(x,p)F(x,p)dx. \qquad (39)$$

However, there is no distribution F that gives the right expectation values for all quantum mechanical operators. This is because noncommuting observables cannot have a genuine joint distribution.[2] That is, we cannot find a distribution F that satisfies, in addition to (39), the equation

$$\langle\psi, q(G(X,P))\psi\rangle = \int \psi^*(x)q(G(X,P))\psi(x)dx$$
$$= \int q(g(x,p))F(x,p)dx, \qquad (40)$$

for arbitrary functions $q(\cdot)$. We shall elucidate this fact by an example.

To this end we recall that Cohen[16] has established a general relationship between phase space distributions and the rules of associating classical quantities to quantum mechanical operators:

$$g(x,p)\rightarrow G(X,P). \qquad (41)$$

These prescriptions are called correspondence rules or rules of association. A correspondence rule is related to a distribution F and a function g if (39) is satisfied. For instance, Weyl's rule

$$e^{i(\theta x+\tau p)}\rightarrow e^{i(\theta X+\tau P)}, \qquad (42)$$

or equivalently

$$x^n p^m \rightarrow \frac{1}{2^n}\sum_{l=0}^{n}\binom{n}{l}X^{n-l}P^m X^l \qquad (43)$$

is obtained if one deals with the Wigner distribution.

If one computes the variance of the energy of the first excited state of the harmonic oscillator by means of the classical Hamiltonian and the Wigner distribution, we obtain a nonzero value. This is clearly in disagreement with quantum mechanics. The explanation of this "paradox" is easy. Weyl's rule of association promotes the classical Hamiltonian

$$h(x,p) = (1/2m)p^2 + (m\omega^2/2)x^2$$
$$H(X,P) = (1/2m)P^2 + (m\omega^2/2)X^2,$$

but $H^2(X,P)$ is not related to $h^2(x,p)$ in this way, because

$$x^2p^2\rightarrow\tfrac{1}{4}(X^2P^2 + P^2X^2 + 2XP^2X). \qquad (44)$$

Obviously this differs from the term $\frac{1}{2}(X^2P^2 + P^2X^2)$ in $H^2(X,P)$.

What does this example teach us? It exemplifies the fact that a given phase space density is only useful in connection with particular operators for which it yields the proper quantum mechanical expectations. Of course, for operators of either X or P, any quantum mechanical distribution function may be taken [cf. (35)].

We close this section with a remark on the position-momentum uncertainty relations in the phase space formulation of quantum mechanics. It is often maintained (e.g., in Ref. 5) that the existence of a joint distribution of position and momentum is in contradiction with the uncertainty relations. However, this is not so, because in order to establish them one only needs the marginals. Take, e.g., $F(x,p) = |\psi(x)|^2|\hat{\psi}(p)|^2$. This function yields the uncertainty relations.

VI. COMMENT ON THE MOMENTUM PROCESS

What is the physics behind the momentum variables π_t? All we know so far is that they have the right distributions, but we are going to investigate some more of their properties.

A. Operational meaning

Momentum is one of the fundamental concepts both in classical and quantum physics. For instance, this notion enters in an essential way in scattering theory, and momentum is, in fact, what experimentalists frequently claim to be able to measure. From this point of view it seems indispensably necessary that any reasonable definition of momentum has an operational meaning; i.e., a prescription of a measuring procedure must go with it. Normally there are two possible ways of defining momentum.

(a) The classical definition of instantaneous momentum is used whenever the particle trajectories are differentiable.

(b) The time-of-flight definition of Feynman and Hibbs[17] gives another operational prescription for the determination of momentum. This is a common tool in scattering theory.

Instantaneous momentum is clearly of no use in stochastic mechanics, since diffusions are nowhere differentiable, although a more refined theory of Brownian motion containing the microscopic equations should—physically speaking—allow for instantaneous momentum. The time-of-flight technique, however, enters in the statement that the diffusions of stochastic mechanics have the property that each sample path assumes constant velocity asymptotically, and the distribution of this limit coincides with the quantum mechanical initial or final velocity.[9–12]

So for the case of zero potential the definition of π_t is operationally meaningful. If one waits sufficiently, one can measure the momentum approximately. Whenever a nonvanishing potential is present, there is no experimental way of implementing the definition of π_t, because you cannot simply turn off the potential at time t. But this was required in the definition of π_t. For this reason the momentum process π_t is found to have no operational meaning (except in the free case).

B. Non-Markovity

Let us consider the harmonic oscillator in the ground state. From the explicit representation (25) of the momentum process one can already guess that π_t does not have the Markov property because it is the sum of the Markov process ξ_t and a term depending on the future.

To prove this rigorously we consider the covariance of the momentum process. Let $s<t$. After some calculations using (18), (25), and

$$E\,[\xi_s\xi_t] = \sigma^2 e^{-\omega(t-s)}, \tag{45}$$

we find that

$$E\,[\pi_s\pi_t] = m^2\omega^2 e^{-\pi}\Big\{\sigma^2 e^{-\omega(t-s)} + \frac{\hbar}{m} e^{-\omega(t-s)}\int_0^{t-s} e^{\omega z+\gamma(z)}\,dz + \frac{\hbar}{m}\int_{t-s}^{\infty} e^{2\gamma(z)}\,dz\Big\}, \tag{46}$$

i.e., the covariance is a strictly positive function of $t-s$,

$$E\,[\pi_s\pi_t] = m^2\omega^2 e^{-\pi} C(t-s). \tag{47}$$

Moreover, $C(t)$ is not a constant, since

$$\dot C(t) = -\frac{\hbar}{m}\Big[\frac{\omega}{2}e^{-\omega t} + \omega e^{-\omega t}\int_0^t e^{\omega z+\gamma(z)}\,dz - e^{\gamma(t)} + e^{2\gamma(t)}\Big] \tag{48}$$

is negative. In particular, $C(t-s)$ does not split into a product of functions of s and t only.

For a Gaussian process with strictly positive covariance the Markov property is equivalent to the statement that the covariance splits.[18] Thus we have established that π_t is non-Markovian, a result independent of assumption (13).

Whether this must be looked at as a defect of π_t or not is not quite clear. Although stochastic mechanics usually deals with Markovian diffusions, it might be necessary to shift to a non-Markovian framework in order to satisfy the locality principle.[3]

C. Force

We saw that in the simple example of the harmonic oscillator ground state the momentum process was differentiable provided we impose assumption (13). Let us rewrite

$$\dot\pi_t = -m\omega e^{-\pi/2}\int_t^{\infty}\frac{\partial}{\partial z}\,[(\omega-\dot\gamma(z-t))e^{\gamma(z-t)}]\xi_z\,dz \tag{49}$$

by means of Itô's lemma:

$$d([\;\;]\xi_z) = \frac{\partial[\;\;]}{\partial z}\xi_z\,dz + [\;\;]d\xi_z. \tag{50}$$

We also use

$$[\;\;]|_{z=t} = 0, \quad \lim_{z\to\infty}([\;\;]\xi_z) = 0. \tag{51}$$

Then

$$\dot\pi_t = m\omega e^{-\pi/2}\int_t^{\infty}[\;\;]d\xi_z = m\omega e^{-\pi/2}\Big\{-\omega\int_t^{\infty}[\;\;]\xi_z\,dz + \int_t^{\infty}[\;\;]dw_z\Big\}. \tag{52}$$

When inserting (26) this gives

$$\dot\pi_t = -m\omega e^{-\pi/2}\Big[\omega\xi_t - \int_t^{\infty}\dot\gamma(z-t)e^{\gamma(z-t)}\,dw_z\Big]. \tag{53}$$

From this we can deduce that $\dot\pi_t$ is a Gaussian process of mean zero and variance given by

$$\mathrm{Var}\,\dot\pi_t = m^2\omega^2 e^{-\pi}\Big[\omega^2\sigma^2 + \frac{\hbar}{m}\int_0^{\infty}\dot\gamma^2(z)e^{2\gamma(z)}\,dz\Big]. \tag{54}$$

The last integral can be calculated explicitly. Let $y := \arctan \omega z$. Then

$$\int_0^{\infty}\dot\gamma^2(z)e^{2\gamma(z)}\,dz = \omega\int_0^{\pi/2}\Big(\frac{1-\tan y}{1+\tan^2 y}\Big)^2 e^{2y}\,dy. \tag{55}$$

But

$$\Big(\frac{1-\tan y}{1+\tan^2 y}\Big)^2 = \frac{1}{2}\Big(1+\cos 2y - \sin 2y - \frac{1}{2}\sin 4y\Big). \tag{56}$$

Therefore

$$\int_0^{\infty}\dot\gamma^2(z)e^{2\gamma(z)}\,dz = \frac{\omega}{20}(e^{\pi}-11) \tag{57}$$

and finally we obtain

$$\mathrm{Var}\,\dot\pi_t = [(1-e^{-\pi})/10]m^2\omega^4\sigma^2. \tag{58}$$

Of course, $\dot\pi_t$ is expected to be the stochastic equivalent to force. Most unfortunately, however, this is not the case. For the harmonic oscillator, the quantum mechanical force operator F is multiplication by $-m\omega^2 x$. And this implies that the variance of the ground state is equal to

$$\langle(F-\langle F\rangle)^2\rangle = m^2\omega^4\langle x^2\rangle = m^2\omega^4\sigma^2. \tag{59}$$

As this is in disagreement with (58), $\dot\pi_t$ does not have the same distribution as the force operator F. Therefore $\dot\pi_t$ cannot be interpreted as force.

Incidentally, Guerra and Morato[5] define a momentum process for the coherent states of the harmonic oscillator by

$$d\pi_t^{\mathrm{GM}} = F_+(\pi_t^{\mathrm{GM}},t)dt + \sqrt{\hbar m}\omega\,d\tilde w_t, \tag{60}$$

where $\tilde w_t$ is the unit Wiener process. In their construction, however, π_t^{GM} is not acting on the same probability space as ξ_t. For the ground state, the forward and backward forces are given by

$$F_\pm(p,t) = \mp\omega p, \tag{61}$$

and the distributions of both $F_+(\pi_t^{\mathrm{GM}},t)$ and $F_-(\pi_t^{\mathrm{GM}},t)$ coincide with the quantum mechanical force distribution.

D. Uncertainty relations

Without making use of the momentum process, the position–momentum uncertainty relations can be rediscovered in stochastic mechanics, both the usual Heisenberg uncertainty relations[19,20] and their generalized form *à la* Schrödinger[21]:

$$\mathrm{Var}\,X\,\mathrm{Var}\,P \geqslant \mathrm{Cov}^2(X,P) + \hbar^2/4, \tag{62}$$

where

$$\mathrm{Cov}(X,P) : = \tfrac{1}{2}\langle XP+PX\rangle - \langle X\rangle\langle P\rangle. \qquad (63)$$

In the stochastic frame

$$\mathrm{Var}\,\xi(\mathrm{Var}\,u + \mathrm{Var}\,v) \geqslant \mathrm{Cov}^2(\xi,v) + \nu^2. \qquad (64)$$

Therefore the uncertainty relations (58) and (60) are equivalent, since

$$\mathrm{Var}\,X = \mathrm{Var}\,\xi, \qquad (65)$$

$$\mathrm{Var}\,P = m^2(\mathrm{Var}\,u + \mathrm{Var}\,v), \qquad (66)$$

$$\mathrm{Cov}(X,P) = m\,\mathrm{Cov}(\xi,v), \qquad (67)$$

$$\nu = \hbar/2m. \qquad (68)$$

In the example of Sec. IV we were dealing with a coherent state where the Heisenberg inequality turns into an equality, i.e.,

$$\mathrm{Cov}(X,P) = 0. \qquad (69)$$

Can we find the uncertainty relations by means of π_t? We know from the results above that $\mathrm{Var}\,\xi\,\mathrm{Var}\,\pi = \mathrm{Var}\,X\,\mathrm{Var}\,P$, and hence $\mathrm{Var}\,\xi\,\mathrm{Var}\,\pi$ has the same lower bound. But if we try to mimic the proof of (64) we find

$$\begin{aligned}\mathrm{Var}\,\xi\,\mathrm{Var}\,\pi &= E[(\xi - E\xi)^2]E[(\pi - E\pi)^2]\\ &\geqslant |E[(\xi - E\xi)(\pi - E\pi)]|^2\\ &= \mathrm{Cov}^2(\xi,\pi) = (h^2/4)e^{-\pi}. \end{aligned}\qquad (70)$$

The last equality follows from Eq. (25). Of course, in the light of the remarks of Sec. V, there is no surprise that $\mathrm{Cov}(\xi,\pi)$ does not coincide with $\mathrm{Cov}(X,P) = 0$ (cf. also Ref. 22).

Clearly the momentum process has the drawback that it cannot be used in an obvious manner to obtain the uncertainty relations in the stochastic frame, although it does satisfy them.

As for the momentum process π_t^{GM} considered by Guerra and Morato, it also cannot be used directly to obtain the uncertainty relations, because it is not defined on the same probability space as ξ_t.

E. Joint distribution of (ξ_t, π_t)

If we impose assumption (13), the position and momentum variables related to the ground state of the harmonic oscillator are jointly Gaussian and their covariance matrix can be calculated. The explicit formula for the joint probability density of $F(x,p)$ of (ξ_t, π_t) is given by

$$\begin{aligned}F(x,p) = {}&[(1 - e^{-\pi})^{-1/2}/\pi\hbar]\\ &\times\exp\left\{-\frac{1}{2\sigma^2(1-e^{-\pi})}\right.\\ &\left.\times\left[x^2 - \frac{2}{m\omega}e^{-\pi/2}xp + \frac{1}{(m\omega)^2}p^2\right]\right\},\end{aligned}\qquad (71)$$

$$f(\theta,\tau) = \exp(-(\hbar/2)e^{-\pi/2}\theta\tau), \qquad (72)$$

and it fits into Cohen's classification. Of course, there is no time dependence.

F. Variance of the ground state energy

No question, the variance of the ground state energy must vanish, since the ground state is an energy eigenstate. But if we use the density $F(x,p)$ from Eq. (71) along with the classical Hamiltonian, we find the same phenomenon as discussed in Sec. V: The ground state energy has a nonzero variance when computed in phase space.

That is, the joint distribution is not adapted to compute energy dispersions.

VII. CONCLUSION

What are the merits of the momentum process? Of course, it has the (minimal) property that its distribution coincides with the quantum mechanical momentum distribution. But, as we pointed out, this is not the only process with this characteristic.

On the other hand our analysis disclosed some manifestly unphysical features. The most serious ones are as follows: there is no operational implementation, the derivative of the momentum process does not yield force, and there is no straightforward way of gaining the position–momentum uncertainty relations using this process.

These unsatisfactory shortcomings lead us to the conclusion that such a definition of momentum is unacceptable. Yet there is no better stochastic definition known so far. Only configurational observables are satisfactorily embedded in the stochastic frame.

Undoubtedly, any measurement in physics can eventually be reduced to a position measurement. But this does not invalidate the concept of momentum. It is certainly of interest to extract the information on momentum encoded in stochastic mechanics in a way such that position and momentum are on the same footing. It seems to us that a new approach to momentum in stochastic mechanics is in order.

ACKNOWLEDGMENTS

I should like to thank S. Albeverio, Ph. Blanchard, and Ph. Combe for stimulating discussions and R. Brustein for useful correspondence. Thanks are also due to R. Marra for having pointed out Ref. 6 to me.

[1] E. Nelson, "Derivation of the Schrödinger equation from Newtonian mechanics," Phys. Rev. **150**, 1079 (1966).

[2] E. Nelson, *Dynamical Theories of Brownian Motion* (Princeton U. P., Princeton, NJ, 1967).

[3] E. Nelson, *Quantum Fluctuations* (Princeton U. P., Princeton, NJ, 1985).

[4] F. Guerra, "Structural aspects of stochastic mechanics and stochastic field theory," Phys. Rep. 77, 263 (1981).

[5] F. Guerra and L. Morato, "Momentum–position complementarity in stochastic mechanics," in *Stochastic Processes in Quantum Theory and Statistical Physics, Lecture Notes in Physics*, Vol. 173, edited by S. Albeverio, Ph. Combe, and M. Sirugue-Collin (Springer, Berlin, 1982).

[6] M. Davidson, "Momentum in stochastic mechanics," Lett. Math. Phys. **5**, 523 (1981).

[7] D. de Falco, S. De Martino, and S. De Siena, "Momentum from sample paths in stochastic mechanics," Lett. Nuovo Cimento **36**, 457 (1983).

[8] F. Guerra and L. Morato, "Quantization of dynamical systems and stochastic control theory," Phys. Rev. D **27**, 1774 (1983).

[9] F. Guerra, "Probability and quantum mechanics; the conceptual foundations of stochastic mechanics," in *Quantum Probability and Applications to the Quantum Theory of Irreversible Processes, Lecture Notes in Math-*

ematics, Vol. 1055, edited by L. Accardi, A. Frigerio, and V. Gorini (Springer, Berlin, 1984).

[10] D. S. Shucker, "Stochastic mechanics of systems with zero potential," J. Funct. Anal. **38**, 146 (1980).

[11] P. Biler, "Stochastic interpretation of potential scattering in quantum mechanics," Lett. Math. Phys. **8**, 1 (1984).

[12] M. Serva, "Elastic scattering in stochastic mechanics," Lett. Nuovo Cimento **41**, 198 (1984).

[13] E. A. Carlen, "Potential scattering in stochastic mechanics," Ann. Inst. H. Poincaré A **42**, 407 (1985).

[14] E. P. Wigner, "Quantum-mechanical distribution functions revisited," in *Perspectives in Quantum Theory*, edited by W. Yourgrau and A. van der Merwe (M.I.T., Cambridge, MA, 1971).

[15] E. P. Wigner, "On the quantum correction for thermodynamic equilibrium," Phys. Rev. **40**, 749 (1932).

[16] L. Cohen, "Generalized phase-space distribution functions," J. Math. Phys. **7**, 781 (1966).

[17] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw–Hill, New York, 1965).

[18] B. Simon, *Functional Integration and Quantum Physics* (Academic, New York, 1979).

[19] L. de La Peña-Auerbach and M. Cetto, "Stronger form for the position–momentum uncertainty relation," Phys. Lett. A **39**, 65 (1972).

[20] D. de Falco, S. De Martino, and S. De Siena, "Position–Momentum uncertainty in stochastic mechanics," Phys. Rev. Lett. **49**, 181 (1982).

[21] S. Golin, "Uncertainty relations in stochastic mechanics," J. Math. Phys. **26**, 2781 (1985).

[22] H. Margenau and R. N. Hill, "Correlation between measurements in quantum theory," Prog. Theor. Phys. **26**, 722 (1961).

The range of quantum probability

Itamar Pitowsky
Department of Philosophy, The Hebrew University, Jerusalem, Israel

(Received 8 October 1985; accepted for publication 30 December 1985)

The set of all pair (and in fact higher-order) distributions that are representable in quantum mechanics is characterized and compared with the classical range. Various interference phenomena yield pair distributions that are not classical; a few examples are discussed. These results shed light on some fundamental problems concerning the interpretation of quantum mechanics, in particular it is demonstrated how the "quantum logic" of Birkhoff and Von Neumann can be naturally interpreted in terms of truth values. Finally, the possibility of interpreting quantum probability in a realistic "quasiclassical" way is explored.

I. BETWEEN TWO CONVEX POLYHEDRA

Let SR_n denote the set of all $n \times n$ real symmetric matrices. Then SR_n is a linear space with the usual matrix addition and scalar multiplication and $\dim(\mathrm{SR}_n) = \frac{1}{2}n(n+1)$. Also SR_n is a topological space with the usual Euclidean topology. For convenience I shall represent this topology with the supremum norm. Thus if $a = (a_{ij}) \in \mathrm{SR}_n$, we shall put $\|a\| = \max_{ij}|a_{ij}|$. I shall denote by $\mathrm{cl}(A)$ and $\mathrm{int}(A)$ the closure and interior of a subset $A \subseteq \mathrm{SR}_n$, respectively, and by $\mathrm{co}(A)$ the convex hull generated by A (i.e., the intersection of all convex sets containing A).

Definition 1.1: A matrix $p \in \mathrm{SR}_n$ is called a phenomenal pair distribution of order n if, for all $i, j = 1,2,...,n$,

$$0 \leqslant p_{ij} \leqslant \min(p_{ii}, p_{jj}) \leqslant \max(p_{ii}, p_{jj}) \leqslant 1 . \tag{1.1}$$

Let L_n be the set of all phenomenal pair distributions of order n. Then it is easy to see that L_n is a compact convex subset of SR_n. The motivation behind this definition is probabilistic. Let $s_1, s_2, ..., s_n$ be "events" or "states" of some system and put $p_{ii} = \mathrm{prob}(s_i)$ and $p_{ij} = \mathrm{prob}(s_i \& s_j)$ then surely $p = (p_{ij}) \in L_n$. Condition (1.1) is indeed necessary for p to represent classical pair distribution but it is not sufficient. Thus the matrix $p = \left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)$ is in L_2 but it does not represent any classical distribution since the intersection of two events, each having probability 1, has necessarily probability 1 as well. Let us therefore introduce the following definition.

Definition 1.2: A phenomenal distribution $p \in L_n$ has a classical representation if there exists a probability space (X, Σ, μ) and events $A_1, ..., A_n \in \Sigma$ such that $p_{ij} = \mu(A_i \cap A_j)$, $ij = 1,2,...,n$.

Let C_n denote the set of all phenomenal pair distributions of order n that have a classical representation. We shall prove that both C_n and L_n are (compact) convex polyhedra in SR_n with nonempty interior and characterize them in terms of their extreme points. It is a matter of fact that in microphysics one observes pair distributions $p \in L_n$ that do not have a classical representation (see examples in Sec. II A). It is therefore interesting to characterize all those pair distributions that might arise in a quantum mechanical context. So we define the following.

Definition 1.3: A phenomenal pair distribution $p \in L_n$ has a quantum mechanical representation if there exists a separable complex Hilbert space H a density operator (statistical operator) W on H and (continuous) projections $E_1, ..., E_n$ (which do not necessarily pairwise commute) such that

$$p_{ij} = \mathrm{tr}[W(E_i \wedge E_j)], \quad ij = 1,2,...,n , \tag{1.2}$$

where $E_i \wedge E_j$ denotes the projection onto the closed subspace $E_i(H) \cap E_j(H)$.

Let Q_n denote the set of all phenomenal distributions $p \in L_n$, which have a quantum mechanical representation. I shall prove that Q_n is convex and that $Q_n \supseteq C_n$. However, Q_n is not closed and thus it lies between two convex polyhedra $C_n \subset Q_n \subset L_n$. The main result of Sec. I is the demonstration that Q_n contains the whole interior of L_n: $Q_n \supset \mathrm{int}(L_n)$ and thus all phenomenal pair distributions have quantum representation, save for some that lie on the faces of L_n.

The present discussion concerns pair distributions $p_{ij} = \mathrm{prob}(s_i \& s_j)$ but it can be extended easily to triple quadruple or any k-tuple distribution:

$$p_{i_1 i_2 \cdots i_k} = \mathrm{prob}(s_{i_1} \& s_{i_2} \& \cdots \& s_{i_k}) .$$

With the obvious extensions of Definitions 1.1, 1.2, and 1.3, all the proofs given in this article generalize easily. I thus have decided to concentrate on the simple case and save cumbersome notations.

The most common cases of interference phenomena give rise to pair distributions $p \in Q_n$, which lie outside the classical domain C_n. Few typical examples are discussed in Sec. II A. I believe that the formal results of Sec. I shed some new light on fundamental problems concerning the interpretation of quantum mechanics. These are discussed in Sec. II B. In particular, the so-called "quantum logic" can be viewed from a new angle (Sec. II C).

In the third section, I explore the possibility of representing quantum mechanical distributions $p \in Q_n$ by means other than the ones provided by the Hilbert space formalism. In particular the possibility of forming a "realistic" model of such distributions is discussed.

The connection between probability theory and convex analysis is long known. It was explored lately by Grag and Mermin[1] in the case of phenomenal pair distributions associated with spin measurements. This paper was motivated by their results. All basic facts concerning convex sets and their properties that I use are to be found in the monograph by Rockafellar.[2]

0022-2488/86/061556-10$02.50

A. Characterization of L_n

A matrix $u\in \mathrm{SR}_n$ will be called an *extreme matrix* if $u\in L_n$ and $u_{ij}=0$ or $u_{ij}=1$ for all $i,j=1,2,\ldots,n$. Let $\mathscr{E}_n$ be the set of all $n\times n$ extreme matrices then E_n is finite [surely $\mathrm{card}(\mathscr{E}_n)<2^{(1/2)n(n+1)}$] and we have the following therorem.

Theorem 1.1: A matrix $p\in\mathrm{SR}_n$ is an extreme point of the convex set L_n if and only if it is an extreme matrix. Therefore $L_n=\mathrm{co}(\mathscr{E}_n)$.

Proof: It is easy to see that if $u\in L_n$ is an extreme matrix then u is an extreme point of $L_{n.}$ Suppose that we have a representation $u=\lambda p+(1-\lambda)p'$, where $p,p'\in L_n$ and $0<\lambda<1$, then, for all $i,j=1,2,\ldots,n$, $u_{ij}=\lambda p_{ij}+(1-\lambda)p'_{ij}$. We know that $u_{ij}=0$ or $u_{ij}=1$. Since $\{0\}$, $\{1\}$ are extreme points of the interval $[0,1]$ and since $0\leqslant p_{ij},p'_{ij}\leqslant 1$, we have $u_{ij}=0$ entails $p_{ij}=p'_{ij}=0$, $u_{ij}=1$ entails $p_{ij}=p'_{ij}=1$, and, therefore, $p=p'=u$ and u is an extreme point of L_n.

Suppose that p is an extreme point of L_n then since L_n is closed we have $p\in L_n$. Consider the function $f(t)=2t-t^2$. It is an increasing function in the interval $[0,1]$ and $f(0)=0, f(1)=1$. Define a matrix p^1 by $p^1_{ij}=f(p_{ij})$, then from the above-mentioned properties of f it follows that $p^1\in L_n$. Consider the function $g(t)=t^2$. Again it is an increasing function in the interval $[0,1]$ and $g(0)=0$, and $g(1)=1$, therefore the matrix p^2 defined by $p^2_{ij}=g(p_{ij})$ is in L_n. Now $\frac{1}{2}(f(t)+g(t))=t$, therefore $p=\frac{1}{2}p^1+\frac{1}{2}p^2$. Since p is an extreme point we have $p=p^1=p^2$, therefore, in particular, $p_{ij}=p^2_{ij}$ for all $i,j=1,2,\ldots,n$. Hence $p_{ij}=1$ or $p_{ij}=0$ for all i,j and since $p\in L_n$ it follows that p is an extreme matrix. Every compact convex set is the convex hull generated by its set of extreme points, hence $L_n=\mathrm{co}(\mathscr{E}_n)$. Q.E.D.

The set of all $n\times n$ extreme matrices $\mathscr{E}_n$ is finite, therefore L_n is a polyhedron. Suppose $\mathscr{E}_n=\{u^1,u^2,\ldots,u^k\}$, then every $p\in L_n$ has a representation $P=\Sigma_{\nu=1}^{k}\lambda_\nu\mu^\nu$, where $0\leqslant\lambda_\nu\leqslant 1$ and $\Sigma_{\nu=1}^{k}\lambda_\nu=1$. Note that the zero matrix, $u_{ij}=0$, $i,j=1,2,\ldots,n$, is also an element of E_n.

B. Characterization of C_n

C_n, the set of all $p\in L_n$ that have a classical representation, is a convex polyhedron generated by a subset of the set of extreme matrices. Let $\Omega=\{0,1\}^n$ be the set of all n-tuples of zeros and ones, $\mathrm{card}(\Omega)=2^n$. I shall denote by $\epsilon=(\epsilon_1,\epsilon_2,\ldots,\epsilon_n)$ the elements of Ω. For $\epsilon\in\Omega$ let $u(\epsilon)$ be the extreme matrix defined by $u_{ij}(\epsilon)=\epsilon_i\epsilon_j$, $i,j=1,2,\ldots,n$. In this way we obtain 2^n distinct extreme matrices.

Theorem 1.2: C_n is the closed convex hull generated by the set of matrices $\{u(\epsilon);\ \epsilon\in\Omega\}$.

Proof: Suppose that $p\in C_n$. Then there is a probability space (X,Σ,μ) and subsets $A_1,\ldots,A_n\in\Sigma$ such that $p_{ij}=\mu(A_i\cap A_j)$, for $i,j=1,2,\ldots,n$. Let $B\in\Sigma$ be an arbitrary set. Denote $B^1=B$, $\mathrm{B}^0=\mathrm{X}\backslash B=\widetilde{B}$ and for $\epsilon\in\Omega$ denote $A(\epsilon)=A_1^{\epsilon_1}\cap A_2^{\epsilon_2}\cap\cdots\cap A_n^{\epsilon_n}$. Then for $\epsilon\neq\epsilon'$ we have $A(\epsilon)\cap A(\epsilon')=\phi$ and $\cup_{\epsilon\in\Omega}A(\epsilon)=X$. Put $\lambda(\epsilon)=\mu[A(\epsilon)]$, then $0\leqslant\lambda(\epsilon)\leqslant 1$ and $\Sigma_{\epsilon\in\Omega}\lambda(\epsilon)=1$. For all $i,j=1,2,\ldots,n$, we have $A_i\cap A_j=\cup\{A(\epsilon);\ \epsilon_i=\epsilon_j=1\}$. Hence, for $i,j=1,2,\ldots,n$,

$$p_{ij}=\mu(A_i\cap A_j)=\sum_{\{\epsilon|\epsilon_i=\epsilon_j=1\}}\lambda(\epsilon)$$

$$=\sum_{\epsilon\in\Omega}\epsilon_i\epsilon_j\lambda(\epsilon)=\sum_{\epsilon\in\Omega}u_i(\epsilon)\lambda(\epsilon).$$

Therefore $p\in\mathrm{co}\{u(\epsilon);\ \epsilon\in\Omega\}$.

As for the converse, suppose $p\in\mathrm{co}\{u(\epsilon);\ \epsilon\in\Omega\}$, then we can write $p=\Sigma_{\epsilon\in\Omega}\lambda(\epsilon)u(\epsilon)$, where $0\leqslant\lambda(\epsilon)\leqslant 1$ and $\Sigma_{\epsilon\in\Omega}\lambda(\epsilon)=1$. Let $X=\{0,1\}^n=\Omega$ and let $\Sigma=$ the power set of X. Define a measure μ on the singletons of Σ by $\mu(\{\epsilon\})=\lambda(\epsilon)$ and for $B\subseteq X$ put $\mu(B)=\Sigma_{\epsilon\in B}\lambda(\epsilon)$. Then (X,Σ,μ) is a probability space. Put $A_i=\{\epsilon;\ \epsilon_i=1\}$, then

$$\mu(A_i\cap A_j)=\sum_{\{\epsilon;\ \epsilon_i=\epsilon_j=1\}}\lambda(\epsilon)=\sum_{\epsilon\in\Omega}\epsilon_i\epsilon_j\lambda(\epsilon)$$

$$=\sum\lambda(\epsilon)u_{ij}(\epsilon)=p_{ij}\ .\qquad\text{Q.E.D.}$$

As a corollary we see that if $p\in C_n$, then p can be represented on the space $X=\Omega=\{0,1\}^n$ and $\Sigma=$ the power set of X. Let $\epsilon\in\Omega$ and put $|\epsilon|=\Sigma_{i=1}^{n}\epsilon_i$. Consider the set of matrices $u(\epsilon)$ for $\epsilon\in\Omega$ such that $|\epsilon|=1$. There are n such matrices with only one nonzero entry on the diagonal. Consider the set of matrices $u(\epsilon)$ for $\epsilon\in\Omega$ with $|\epsilon|=2$. There are $\binom{n}{2}=n(n-1)/2$ such matrices, each having four nonzero entries. All the above matrices (those with $|\epsilon|=1$ and those with $|\epsilon|=2$) form a linearly independent set in SR_n. But there are $n+\binom{n}{2}=\frac{1}{2}n(n+1)=\dim(\mathrm{SR}_n)$ such matrices and hence the space spanned by C_n in SR_n is SR_n itself. Since the zero matrix is in C_n we have proved the following corollary.

Corollary 1.2: Both C_n and L_n have nonempty interior.

C. Probability in Hilbert spaces

The state of a quantum mechanical system is given, in the most general case, by a density operator on a complex Hilbert space. Let H be a Hilbert space. Then W is a density operator if (a) W is Hermitian, $W^\dagger=W$; (b) W is definite, $\langle\phi|W|\phi\rangle\geqslant 0$, for all $|\phi\rangle\in H$; and (c) the trace of W is well defined and $\mathrm{tr}(W)=1$. Pure states are those states for which W is a projection onto a one-dimensional subspace, $W=|\phi\rangle\ \langle\phi|$, the other states are "mixtures." With every closed subspace of H there corresponds a unique (orthogonal) projection operator E onto this subspace.

For two such projections E_1,E_2 (which do not necessarily commute) let $E_1\wedge E_2$ be the projection onto the subspace $E_1(H)\cap E_2(H)$, let $E_1\vee E_2$ be the projection onto the closed subspace spanned by $E_1(H)\cup E_2(H)$, and let $E_1^{\perp}$ be the projection onto the subspace orthogonal to $E_1(H)$. With these operations the set of all closed (orthogonal) projections forms an orthocomplemented lattice. In quantum mechanics we associate with every projection E as an idealized observable whose expectation on the pure state $|\phi\rangle\ \langle\phi|$ is $\langle\phi|E\,|\phi\rangle$. More generally the expectation of E on the mixture W is given by $\mathrm{tr}(WE)$. Thus given a density operator W puts $\mu(E)=\mathrm{tr}(WE)$ and we obtain

$$\mu(E^{\perp})=1-\mu(E),\quad\text{for all } E;\qquad(1.3)$$

$$\mu(I)=1,\quad\mu(0)=0,\qquad(1.4)$$

where I and 0 denote the identity and zero operators, respectively; and if $E_1,E_2,...,E_n$ are *pairwise orthogonal* then

$$\mu\left(\bigvee_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i) \, . \tag{1.5}$$

This means that if μ is restricted to families of pairwise commuting projections, μ behaves like a regular measure. Due to a deep theorem by Gleason[3] we know that every measure on closed projections in a separable Hilbert space of dimension $\geqslant 3$, which satisfy (1.3)–(1.5), is given by some density matrix W, i.e., by $\mu(E) = \text{tr}(WE)$. Our definition of Q_n, therefore, does not restrict the discussion.

In the following I shall denote by $H_1 \oplus H_2$ the direct sum of two Hilbert spaces, i.e., the set of all pairs $|\phi \oplus \psi\rangle$, with $|\phi\rangle \in H_1$ and $|\psi \in H_2$, with the usual coordinate by coordinate addition and scalar multiplication, and with the inner product

$$\langle \phi \oplus \psi | \phi' \oplus \psi' \rangle = \langle \phi | \phi' \rangle + \langle \psi | \psi' \rangle .$$

Here $H_1 \otimes H_2$ will denote the tensor product of H_1,H_2 if $|\phi_1\rangle \in H_1$, $|\phi_2\rangle H_2$, and $|\phi_1 \psi_2\rangle$ stand for $|\phi_1\rangle \oplus |\phi_2\rangle$.

D. Characterization of Q_n

First I shall prove the following lemma.

Lemma 1.4.1: Q_n is a convex set in SR_n.

Proof: Let $p,p' \in Q_n$ and $0 < \lambda < 1$. We have to show that $\lambda p + (1-\lambda)p' \in Q_n$. Since $p,p' Q_n$, there are Hilbert spaces H,H', density operators W,W' on H,H', respectively, and projections $E,...,E_n$ in H and $E'_1,...,E'_n$ in H' such that

$$p_{ij} = \text{tr}[W(E_i \wedge E_j)], \quad p'_{ij} = \text{tr}[W'(E'_i \wedge E'_j)] \, .$$

Let $H = H \oplus H'$ be the direct sum of H and H' and let $\overline{W}$ be the operator defined on H by $\overline{W}|\phi \oplus \psi\rangle = |\lambda W \phi \oplus (1-\lambda) W' \psi\rangle$. Then $\overline{W}$ is linear, self-adjoint definite, and $\text{tr}(\overline{W}) = \lambda \, \text{tr}(W) + (1-\lambda)\text{tr}(W') = 1$. Let $\overline{E}_i$ be the projection onto the direct sum $E_i(H) \oplus E'_i(H')$, i.e., $\overline{E}_i = E_i \oplus E'_i$. Then since

$$[E_i(H) \oplus E'_i(H')] \cap [E_j(H) \oplus E'_j(H')]$$
$$= [E_i(H) \cap E_j(H)] \oplus [E'_i(H') \cap E'_j(H')] \, ,$$

we have $\overline{E}_i \wedge \overline{E}_j = E_i \wedge E_j \oplus E'_i \wedge E'_j$ and, therefore,

$$\text{tr}[\overline{W}(\overline{E}_i \wedge \overline{E}_j)]$$
$$= \lambda \, \text{tr}[W(E_i \wedge E_j)] + (1-\lambda)\text{tr}[W'(E'_i \wedge E'_j)]$$
$$= \lambda p_{ij} + (1-\lambda)p'_{ij} \, . \qquad \text{Q.E.D.}$$

Lemma 1.4.2: $C_n \subset Q_n$.

Proof: Let H be a Hilbert space of dimension 2^n. Let $\{|\psi(\epsilon)\rangle; \ \epsilon \in \Omega\}$ be an orthonormal basis of H parametrized in an arbitrary way by $\epsilon \in \Omega = \{0,1\}^n$. As a consequence of Theorem 1.2 we know that if $p \in C_n$ then there are numbers $0 \leqslant \lambda(\epsilon) \leqslant 1$ such that $\Sigma_{\epsilon \in \Omega} \lambda(\epsilon) = 1$ and

$$p_{ij} = \sum_{\epsilon \in \Omega} \lambda(\epsilon) u_{ij}(\epsilon) = \sum_{\epsilon \in \Omega} \epsilon_i \epsilon_j \lambda(\epsilon) \, .$$

Let W be the density operator that, relative to the basis $\{|\psi(\epsilon)\rangle; \ \epsilon \in \Omega\}$, is given by the matrix $\langle \psi(\epsilon)|W|\psi(\epsilon')\rangle = 0$, if $\epsilon \neq \epsilon'$ and $\langle \psi(\epsilon)|W|\psi(\epsilon)\rangle = \lambda(\epsilon)$. Let E_i be the projection into the subspace spanned by all $|\psi(\epsilon)\rangle$ with $\epsilon_i = 1$, then $E_i \wedge E_j$ is the projection onto the subspace spanned by all $|\psi(\epsilon)\rangle$ with $\epsilon_i = \epsilon_j = 1$ and thus

$$\text{tr}[W(E_i \wedge E_j)] = \sum_{\{\epsilon | \epsilon_i = \epsilon_j = 1\}} \lambda(\epsilon) = \sum_{\epsilon \in \Omega} \lambda(\epsilon) \epsilon_i \epsilon_j = p_{ij} \, .$$

Q.E.D.

As a consequence we also proved the following corollary.

Corollary 1.4.3: A sufficient condition for $p \in Q_n$ to be in C_n is that $E_1,...,E_n$ pairwise commute.

Corollary 1.4.4: Q_n has a nonempty interior.

Since $C_n \subseteq Q_n$ and int $C_n \neq 0$, We shall now show that Q_n is "almost all" of L_n.

Theorem 1.4: int $(L_n) \subset Q_n \subset L_n$.

Proof: It suffices to prove that Q_n is densed in L_n. For if A is a convex set in a Euclidean space then $\text{int}(\text{cl}(A)) = \text{int}(A)$ (see Ref. 4). Thus if Q_n is densed, then, since Q_n is convex, we have

$$\text{int}(L_n) = \text{int}(\text{cl}(Q_n)) = \text{int}(Q_n) \subseteq Q_n \, .$$

In order to show that Q_n is densed in L_n it is sufficient to demonstrate that for every $\epsilon > 0$ and each extreme matrix $u \in \mathscr{E}_n$ there exists $q \in Q_n$ such that $\|u - q\| = \max_{ij} |u_{ij} - q_{ij}| < \epsilon$. For if this is the case let $\mathscr{E}_n = \{u^1,...,u^k\}$ be the set of all $n \times n$ extreme matrices and let $\epsilon > 0$. Then for each $\nu = 1,2,...,k$ there is a matrix $q^\nu \in Q_n$ such that $\|u^\nu - q^\nu\| < \epsilon$. Let $p \in L_n$. Then we can represent $p = \Sigma_{\nu=1}^k \lambda_\nu u^\nu$ for $0 \leqslant \lambda_\nu \leqslant 1$, $\Sigma_{\nu=1}^k \lambda_\nu = 1$ (Theorem 1.1). Put $q = \Sigma_{\nu=}^k \lambda_\nu q^\nu$. Then $q \in Q_n$ since Q_n is convex and

$$\|p - q\| \leqslant \sum_{\nu=1}^k \lambda_\nu \|u^\nu - q^\nu\| < \sum_{\nu=1}^k \lambda_\nu \epsilon = \epsilon .$$

Hence for all $\epsilon > 0$, $p \in L_n$, there is $q \in Q_n$ with $\|p - q\| < \epsilon$ and Q_n is densed in L_n.

So, by induction on $n = 2,3,...$, we shall prove that for all (sufficiently small) $\epsilon > 0$ and every $u \in \mathscr{E}_n$ there is a $q \in Q_n$ with $\|u - q\| < \epsilon$.

For $n = 2$, the extreme matrices are

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \ \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$
$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The first four matrices are in C_2 and $C_2 \subseteq Q_2$ so no problem arises here. As for the fifth matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ let $0 < \epsilon < 1$ and let $0 < \theta < \pi/2$ be such that $\cos^2 \theta = 1 - \epsilon$. Let H be a two-dimensional Hilbert space with the orthogonal basis $|1\rangle$, $|2\rangle$. Let $W = |1\rangle\langle 1|$. Let E_1 be the projection into the one-dimensional subspace spanned by the vector $\cos\theta|1\rangle + \sin\theta|2\rangle$ and E_2 the projection onto the subspace spanned by $\cos\theta|1\rangle - \sin\theta|2\rangle$. Then $E_1 \wedge E_2 = 0$ and thus $p_{12} = \text{tr}[W(E_1 \wedge E_2)] = 0$ also

$$p_{11} = \text{tr}[WE_1] = \text{tr}[WE_2] = p_{22} = \cos^2\theta = 1 - \epsilon \, .$$

Hence the matrix

$$\begin{pmatrix} 1-\epsilon & 0 \\ 0 & 1-\epsilon \end{pmatrix} \in Q_2,$$

for all $\epsilon > 0$, and we have proved the claim for $n = 2$.

Assume that we proved the result for every $\epsilon > 0$ and $u \in \mathscr{E}_{n-1}$ (for $n > 2$). Let $u \in \mathscr{E}_n$. For $i = 1,2,...,n$. Let u^i be the $n \times n$ matrix given by $u^i_{ij} = u^i_{ji} = 0$, for all $j = 1,2,...,n$, and $u^i_{jk} = u_{jk}$, for $j,k \neq i$. Then for all i, u^i is an element of $\mathscr{E}_n$

and u^i has zeros in its ith row and column. By the induction hypothesis there exists $q^i \in Q_n$ such that $\|u^i - q^i\| < \epsilon/n$. [To see that, remove the ith row and column from u^i to obtain an $(n-1)\times(n-1)$ extreme matrix $\bar{u}^i \in \mathscr{E}_{n-1}$. By the induction hypothesis there is a matrix $\bar{q}^i \in Q_{n-1}$ such that $\|\bar{u}^i - \bar{q}^i\| < \epsilon/n$. Now add an ith row and column of zeros to $\bar{q}^i$ to obtain a matrix q^i. Surely $\|q^i - u^i\| = \|\bar{q}^i - \bar{u}^i\| < \epsilon/n$. Also $q^i \in Q_n$ since we can always add a projection $E_i = 0$ to the representation of $\bar{q}^i$.]

Since $q^i \in Q_n$ there is a Hilbert space H_i, n projections $E^i_1, E^i_2, \ldots, E^i_n$, and a density operator W_i on H_i such that $q^i_{jk} = \mathrm{tr}[W_i(E^i_j \wedge E^i_k)]$. (In particular we can take $E^i_i = 0$.) Let $H = H_1 \otimes H_2 \otimes \cdots \otimes H_n$ be the tensor product of the n Hilbert spaces and let $W = W_1 \otimes W_2 \otimes \cdots \otimes W_n$. Then W is a density operator on H. For $i = 1,2,\ldots,n$, let

$$E_i = E^1_i \otimes E^2_i \otimes \cdots \otimes E^{i-1}_i \otimes I_i \otimes E^{i+1}_i \otimes \cdots \otimes E^n_i,$$

where I_i is the identity operator on H_i, and let $s \in Q_n$ be defined by $s_{jk} = \mathrm{tr}[W(E_i \wedge E_j)]$. Then by definition we have

$$s_{ii} = q^1_{ii}\, q^2_{ii} \cdots q^{i-1}_{ii}\, 1\, q^{i+1}_{ii} \cdots q^n_{ii} = \prod_{k \neq i} q^k_{ii},$$

and, for $i < j$,

$$s_{ij} = q^1_{ij} \cdots q^{i-1}_{ij}\, q^j_{ii}\, q^{i+1}_{ij} \cdots q^{j-1}_{ij}\, q^i_{jj}\, q^{j+1}_{ij} \cdots q^n_{ij}$$

$$= q^j_{ii}\, q^i_{jj} \prod_{k \neq ij} q^k_{ij}.$$

The proof is concluded when we observe the following: If $0 \leqslant a_1, \ldots, a_m, b_1, \ldots, b_m \leqslant 1$ are numbers such that $|a_i - b_i| < \epsilon$, $i = 1, \ldots, m$, then

$$\left| \prod_{i=1}^m a_i - \prod_{i=1}^m b_i \right| < m\epsilon.$$

Now since $u \in \mathscr{E}_n$ and since, for $k \neq i$, $u^k_{ii} = u_{ii} = (u_{ii})^{n-1}$, we have

$$|s_{ii} - u_{ii}| = |s_{ii} - (u_{ii})^{n-1}|$$

$$= \left| \prod_{k \neq i} q^k_{ii} - \prod_{k \neq i} u^k_{ii} \right| < (n-1)\frac{\epsilon}{n} < \epsilon.$$

Let $i < j$. Since for $k \neq ij$ we have $u^k_{ij} = u_{ij} = (u^k_{ij})^{n-2}$ and since $u_{ij} = u^k_{ij} \leqslant u^j_{ii}\, u^i_{jj} = u_{ii} u_{jj}$, we have

$$|s_{ij} - u_{ij}| = \left| q^j_{ii}\, q^i_{jj} \prod_{k \neq ij} q^k_{ij} - u^j_{ii}\, u^i_{jj} \prod_{k \neq ij} u^k_{ij} \right| < n\frac{\epsilon}{n} = \epsilon.$$

Hence $\|s - u\| < \epsilon$, for $s \in Q_n$. Q.E.D.

Corollary 1.4: Let $p \in L_n$. A sufficient condition that $p \in Q_n$ is

$$0 < p_{ij} \leqslant \min(p_{ii}, p_{jj}) \leqslant \max(p_{ii}, p_{jj}) < 1, \qquad (1.6)$$

for $i,j = 1,2,\ldots,n$.

We see that the entire interior of L_n is in Q_n. But $Q_n \neq L_n$, for example, the matrix $(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}) \notin Q_2$ because if

$$\mathrm{tr}(WE_1) = \mathrm{tr}(WE_2) = 1$$

then necessarily $\mathrm{tr}[W(E_1 \wedge E_2)] = 1$. But for every $1 > \epsilon > 0$ the matrices

$$\begin{pmatrix}1-\epsilon & 0\\ 0 & 1\end{pmatrix} \text{ and } \begin{pmatrix}1 & 0\\ 0 & 1-\epsilon\end{pmatrix}$$

are elements of Q_2 and hence Q_2 is L_2 except the extreme matrix $(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix})$. In higher dimensions the situation may be more complicated and Q_n does not contain some nonextreme matrices on the faces of L_n as well. For example,

$$\begin{pmatrix}0.9 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix} \notin Q_3.$$

The following, however, remains valid: *If u is an extreme matrix that is not classical (i.e., $u \notin C_n$) then $u \notin Q_n$.*

For $n = 2$, we can actually draw a picture of L_2, Q_2, and C_2 when we identify the matrix

$$\begin{pmatrix}p_{11} & p_{12}\\ p_{12} & p_{22}\end{pmatrix}$$

with the three-dimensional vector (p_{11}, p_{22}, p_{12}) (Fig. 1).

In general we shall call $Q_n \setminus C_n$, that is, the set of all the matrices that have a quantum mechanical representation but not a classical representation, "the interference region."

II. EXAMPLES AND INTERPRETATIONS

A. Examples

Suppose that we are given a closed convex polyhedron A in a Euclidean space and a vector p in that space and we are asked to determine whether p is an element of A or not. Such a problem can be replaced by an equivalent decision problem: Determine whether p is a solution to set of linear inequalities (which depends on A).[5] Bell's inequality[6] and the Clauser–Horne inequalities[7] are typical examples of such a procedure (associated with C_3 and C_4, respectively). In that respect the results obtained in the previous chapter are extensions of Bell's work.

Yet the violation of Bell's inequality in the Einstein–Podolsky–Rosen (EPR) experiment is only one example of a pair distribution that lies in the interference region. Interference phenomena of various kinds often give rise to pair (or higher-order) distributions that possess this character. In addition to the EPR case I shall discuss three examples: the two-slits experiment, the scattering of identical particles, and the interference of paths of a free particle (in Feynman's path integral formalism). The discussion will be brief with

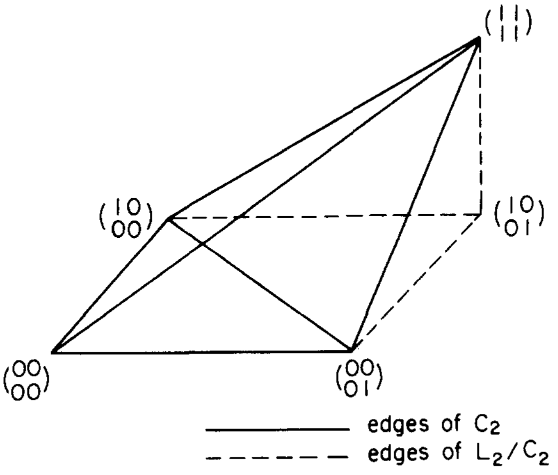

FIG. 1.

many technical details omitted since those problems are worked out in great detail in many textbooks. My purpose is to associate these cases with the analysis of pair distributions provided in the previous chapter. Note that the pair distributions in some of these examples is "phenomenal" only in a theoretical sense. The "experiments" described are only thought experiments. (This applies in particular to the two-slits and free-particle cases.)

(a) EPR experiment: The experiment involves a measurement of spin on a pair of electrons (or other particles) in the singlet state while the particles are sufficiently separated. I shall not describe the experimental setup, which is well known, and only analyze the corresponding pair distribution.

For a given direction w in physical space let $|+w\rangle, |-w\rangle$ be the states "spin up" and "spin down" in the direction w of a spin-$\frac{1}{2}$ system. Let $E_w = |+w\rangle\langle +w|$ and $E_{-w} = |-w\rangle\langle -w|$ be the projections onto these states. Given a pair of electrons, the one-dimensional projection $E_w \otimes E_{-w}$ corresponds to the pure state "spin up in the w direction for the first (left) electron and spin down in the w direction for the second (right) electron" with similar interpretation for similar expressions.

Let W, the density matrix, be the projection onto the singlet state. Since this state is rotationally invariant it can be represented as

$$|\psi\rangle = (1/\sqrt{2})[|+w\rangle \otimes |-w\rangle - |-w\rangle \otimes |+w\rangle]$$

for every direction w. Let x,y,z be three distinct directions and consider the projections

$$\begin{aligned} E_1 &= E_x \otimes E_x \vee E_x \otimes E_{-x}\,, \\ E_2 &= (E_{-y} \otimes E_y)^{\perp} \\ &= (E_y \otimes E_y) \vee (E_y \otimes E_{-y}) \vee (E_{-y} \otimes E_{-y})\,, \\ E_3 &= (E_z \otimes E_z) \vee (E_{-z} \otimes E_z)\,. \end{aligned}$$

Then $E_1 \wedge E_2$ is the projection into the space spanned by $|+x\rangle \otimes |-y\rangle$, $E_1 \wedge E_3$ is the projection into the span $(|+x\rangle \otimes |+z\rangle)$, and $E_2 \wedge E_3$ is the projection into the span $(|+y\rangle \otimes |+z\rangle)$. The matrix $p_{ij} = \mathrm{tr}[W(E_iE_j)]$ is given by

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2}\cos^2(\widehat{xy}/2) & \frac{1}{2}\sin^2(\widehat{xz}/2) \\ \frac{1}{2}\cos^2(\widehat{xy}/2) & \frac{1}{2} & \frac{1}{2}\sin^2(\widehat{yz}/2) \\ \frac{1}{2}\sin^2(\widehat{xz}/2) & \frac{1}{2}\sin^2(\widehat{yz}/2) & \frac{1}{2} \end{pmatrix},$$

where $\widehat{xy}$ is the angle between the directions x and y. We have the following theorem.

Bell's Theorem[6]: There is a choice of directions x,y,z such that $p \notin C_3$.

Proof: Suppose there exists a probability space (X,Σ,μ) and events, which I shall also denote by E_1,E_2,E_3, such that $\mu(E_i \cap E_j) = p_{ij}$. Then put $\widetilde{E}_i = X \setminus E_i$. We have $\mu(E_1 \cap \widetilde{E}_2) \geqslant \mu(E_1 \cap \widetilde{E}_2 \cap E_3)$ and $\mu(E_2 \cap E_3) \geqslant \mu(E_1 \cap E_2 \cap E_3)$. Adding these inequalities we get

$$\mu(E_1 \cap \widetilde{E}_2) + \mu(E_2 \cap E_3) \geqslant \mu(E_1 \cap E_3)\,,$$

but $\mu(E_1 \cap \widetilde{E}_2) = \mu(E_1) - \mu(E_1 \cap E_2)$. Substituting from the matrix we obtain that $p \in C_3$ only if

$$\tfrac{1}{2}\sin^2\left(\frac{\widehat{xy}}{2}\right) + \tfrac{1}{2}\sin^2\left(\frac{\widehat{yz}}{2}\right) \geqslant \tfrac{1}{2}\sin\left(\frac{\widehat{xz}}{2}\right).$$

Take x,y,z, which lie in the same plane with $\theta = \widehat{xy} = \widehat{yz} = \frac{1}{2}\widehat{xz}$. We get that $p \in C_3$ only if $\sin^2(\theta/2) \geqslant \frac{1}{2}\sin^2\theta$, an equality that is violated for, say $\theta = 60°$.

(b) Two-slits experiment: Consider a beam of photons (or another type of particle) scattered by two slits and arrive at a screen behind. Let R_1,R_2,R_3 be the regions on the screen above the upper slit, between the two slits, and below the lower slit, respectively. Consider the following events:

s_1 - photon passes through the upper slit,
s_2 - photon arrives at region R_1,
s_3 - photon arrives at region R_2,
s_4 - photon arrives at region R_3.

Formally these events correspond to projections in a Hilbert space E_1,E_2,E_3,E_4 such that $E_i \perp E_j$, for $i,j = 2,3,4$, and $[E_1,E_j] \neq 0$, for $j = 2,3,4$. I shall not bother to identify these subspaces and concentrate on the phenomenal level.

Suppose that the coming photons are all in a fixed pure state so that the density operator W is just the projection onto this state. In order to measure probability $(s_1) = \mathrm{tr}(WE_1)$ we put detectors behind the two screens and count the number of incoming photons [Fig. 2(a)]. Assume that the slits are symmetric relative to the source so that $p_{11} = \text{probability } (s_1) = \frac{1}{2}$.

The measurement of $p_{ii} = \text{probability } (s_i)$, for $i = 2,3,4$, consists of counting the number of photons on the screen when interference occurs [Fig. 2(b)]. By symmetry we have $p_{22} = p_{44} = \frac{1}{2}(1 - p_{33})$. In order to measure p_{1j}, for $j = 2,3,4$, we cannot use detectors as in Fig. 1(a), since, in the best case, the detector will further scatter the incoming particles. The following suggestion due to Einstein[8] provides a thought experiment which does the job. Let the screen behind the slits be so constructed that it can move up and down parallel to the line connecting the slits. Then we can measure both the position of the incoming photon by detecting it on the screen and determine which slit it came from by measuring the direction of its momentum component parallel to the screen. Due to the uncertainty principle the interference will be destroyed and the distribution on the screen will be just the average of two normal curves about the slits [Fig. 2(c)]. We have $p_{12} = \mathrm{prob}(s_1 \& s_2) = \frac{1}{4}$. Here p_{13} is slightly less than $\frac{1}{4}$, say, $p_{13} = \frac{1}{4} - \delta$, and $p_{14} = \delta$. The matrix is therefore

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} - \delta & \delta \\ \frac{1}{4} & (1 - p_{33})/2 & 0 & 0 \\ \frac{1}{4} - \delta & 0 & p_{33} & 0 \\ \delta & 0 & 0 & (1 - p_{33})/2 \end{pmatrix}.$$

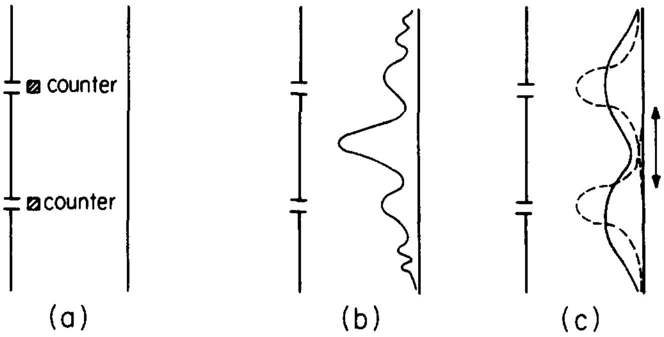

FIG. 2.

Suppose that $p\in C_4$. Then there is a probability space (X,Σ,μ) and events E_1,E_2,E_3,E_4 such that $p_{ij}=\mu(E_i\cap E_j)$, $i,j=1,2,3,4$. In particular

$$p_{33}=\mu(E_3)=\mu(E_3\cap E_1)+\mu(E_3\cap\tilde{E}_1)\,.$$

Now $\mu(E_3\cap E_1)=\frac{1}{4}-\delta$ and, by symmetry, $\mu(E_3\cap\tilde{E}_1)=\frac{1}{4}-\delta$. Hence $p_{33}=\frac{1}{2}-2\delta$. This condition may be violated when interference is sufficiently strong. In this case it may even have $p_{33}>\frac{1}{2}$.

(c) Scattering of identical particles: Both the scattering cross section and the square of the absolute value of the free-particle propagator diverge when integrated in their domain of definition so that probabilities in these cases are conditional. I shall not bother with precise normalization and assume that an appropriate normalization was chosen. Consider a completely elastic proton–proton scattering in which only Coulomb forces play an effective role. We thus assume that the interaction is spin independent. Let R_1,R_2 be the upper and lower halves of the scattering plan and consider the events

s_1 - the left proton is scattered into R_1,

s_2 - the left proton is scattered into R_2.

Let $0<\theta<\pi/2$ and let $\Delta\theta$ be a small angle. The third event is

s_3 - a proton is detected at $(\theta-\frac{1}{2}\Delta\theta,\theta+\frac{1}{2}\Delta\theta)$.

The event s_3 can occur in two ways (Fig. 3). By symmetry we have

$$p_{11}=\text{prob }(s_1)$$
$$=\text{prob }(s_2)=p_{22}=\tfrac{1}{2},\ p_{12}=0\,,$$

and

$$p_{33}=C\,|f(\theta)-f(\pi-\theta)|^2\Delta\theta\,,$$

where C is an appropriate normalization constant and $f(\theta)$ is the scattering amplitude.

In order to measure p_{13} and p_{23} we have to attach "labels" to the protons in order to identify which process was taking place, the one in Fig. 3(a) or in Fig. 3(b). Since the forces do not cause spin interchange we can use opposite spins on the left and right beam and then we have $p_{13}=C\,|f(\theta)|^2\Delta\theta$ and $p_{23}=C\,|f(\pi-\theta)|^2\Delta\theta$. Since

$$|f(\theta)-f(\pi-\theta)|^2\neq|f(\theta)|^2+|f(\pi-\theta)|^2\,,$$

we have $p_{33}\neq p_{13}+p_{23}$, even though $p_{11}+p_{22}=1$, $p_{12}=0$, so that $p\notin C_3$. Note that this violation of classicality does not occur because the spins are coupled with the forces, it is solely due to the interference in the identity of the particles in the measurement of p_{33}.

(d) Free-particle propagator: Consider a free particle moving in one dimension x. Suppose the particle is starting at $x=0$ at time $t=0$. Let $0<t_1<t_2$ and consider the two events

s_1 - the particle is at (x_1-a,x_1+a) at time t_1,

s_2 - the particle is at (x_2-b,x_2+b) at time t_2.

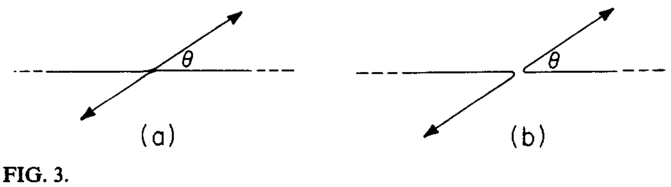

FIG. 3.

Let

$$K(x_2,t_1;\ x_2t_2)$$
$$=\left[\frac{2\pi i\hbar(t_2-t_1)}{m}\right]^{-1/2}\exp\left[\frac{im(x_2-x_1)^2}{2\hbar(t_2-t_1)}\right]$$

be the free particle propagator (m is the particle mass) then

$$p_{11}=C\left|\int_{x_1-a}^{x_1+a}K(0,0;\ \xi,t_1)d\xi\right|^2,$$

$$p_{22}=C\left|\int_{x_2-b}^{x_2+b}K(0,0;\ \xi t_2)d\xi\right|^2,$$

where C is an appropriate normalization. In order to measure p_{12} we have to introduce an infinite potential barrier (or slit) of size $2a$ around x_1 at time t_1, remove it immediately after, and count the rate at which particles arrive at (x_2-b, x_2+b). We obtain

$$p_{12}=C\left|\int_{x_1-b}^{x_2+b}\int_{x_1-a}^{x_2+a}K(0,0;\ \xi_1,t_1)\right.$$
$$\left.\times K(\xi_1,t_1;\ \xi_2-\xi_1,t_2-t_1)d\xi_1\,d\xi_2\right|^2.$$

The value p_{12} depends crucially on the size of the slit $(=2a)$. Various examples are given in Feynman and Hibbs,[9] where the reader can see that for an appropriate choice of the parameters we shall obtain

$$p=\begin{pmatrix}p_{11} & p_{12}\\ p_{21} & p_{22}\end{pmatrix}\notin C_2\,.$$

B. Consequences and interpretation

Considered from a purely formal perspective the results of Sec. I demonstrate that the classical concept of probability is more restrictive then the quantum notion. Moreover, save for some boundary cases, every pair (and in fact every multiple) phenomenal distribution has a quantum representation. The only requirements are that "probability" be a number between zero and one and that the probability of a joint of two events is less or equal to the probability of each of the events. We see therefore that the Hilbert space formalism associated with quantum theory is *in itself* void of any physical content. The content of quantum theory is introduced to this tautological background by identifying certain *particular* operators as physical observables and by the unitary representations of physical symmetries. This remark is concerned with the numerous attempts to axiomatize quantum mechanics while overstressing the role of the background Hilbert space formalism. A large portion of the axioms, in particular the identification of *every* self-adjoint operator (and thus every projection) as an "observable," are nothing but a sophisticated guise for the triviality presented in (1.6).

The classical notion of probability is indeed more restrictive but it is nevertheless rooted in some very basic intuition. Probability, at least in the discrete case, has always been associated with the proportion or relative size of a certain subset of a given set. Thus the probability of drawing a red ball from an urn containing a well-mixed finite variety of balls is just the proportion of red balls in the urn. In fact every discrete (rationally valued) classical distribution can be simulated or interpreted in terms of a "drawing balls from an urn" experiment and surely in every such experiment we shall obtain a classical distribution. The fact that interference phenomena give rise to pair distributions that lie outside C_n clashes with this preanalytic notion. What is called "an interpretation of quantum mechanics" is usually an attempt to make sense of nonclassical deviant probabilities. I shall briefly describe a few attempts of this kind.

(a) The nonrealist view: The trouble with the classical concept lies with its insistence on ascribing properties to objects independently of observation. The "balls in the urn" model of probability is based on such an assumption since it takes for granted the idea that one always deals with a sample (balls), which has definite properties (color, composition, etc.), prior to the experiment. When this assumption is removed the existence of nonclassical pair distributions is no longer problematic since the distribution of properties among the particles in the sample is not fixed prior to experiment but depends on the type of experiment that one chooses to perform. This interpretation due to N. Bohr "explains away" the difficulty by appealing to a metaphysical principle, namely the denial of the reality of experimentally independent properties.

(b) Nonlocal view: If one takes the terms "interference" and even "collapse of the wave function" to signify real physical processes one may arrive at the conclusion that those processes are *caused* by certain ill-understood physical mechanisms. Perhaps there is a "field" of a very peculiar sort that influences the behavior of particles and whose properties depend radically on slight changes in the boundary conditions (e.g., the closing of a slit in the two-slits experiment). In this case there is nothing wrong with "probability" *per se* and the appearance of nonclassical distributions is an illusion that results from the presence of these unfamiliar casual influences. Such a "field," if it exists, seems to violate the principles of relativity since its influence is felt instantaneously all over space (and moreover in certain cases its influence seems not to decrease with distance). No detailed theory of the sort exists, but, from a logical point of view, this general approach seems perfectly consistent.

(c) Classical probability theory should be abandoned: The observation of phenomenal pair distributions which lie in the "interference region" should be taken as a primitive basic fact. As in the case of non-Euclidean space-time geometry this fact contradicts some of our basic intuitions and again as in the case of geometry this fact does not force upon us a nonrealistic view (see Sec. III). This approach can be traced to Feynman[10] and is shared by numerous authors.[11] There is an even more radical school that maintains that classical logic should be abandoned.[12] More on this in the following section.

C. Extreme matrices and quantum logic

If we identify probability 1 with "truth" and probability 0 with "falsity" we see that the classical extreme matrices $u(\epsilon)\in C_n$ play the role of truth functions. (Indeed "truth" and "falsity" are usually assigned to propositions, not events, but we can overcome this difficulty by substituting for each event a proposition that describes it in an appropriate language.)

Remember that the interpretation of a matrix $p\in L_n$ is $p_{ii} = \text{prob}(s_i)$, $p_{ij} = \text{prob}(s_i\&s_j)$, where $s_1,...,s_n$ are propositions (events). So for $u(\epsilon)\in C_n$ we obtain

$$u_{ii}(\epsilon) = \begin{cases} 1, & \text{if } s_i \text{ is true}, \\ 0, & \text{if } s_i \text{ is false}, \end{cases}$$

and $u_{ij}(\epsilon) = \epsilon_i\epsilon_j = u_{ii}(\epsilon)u_{jj}(\epsilon)$. This is just the classical rule "the truth value of a conjuction is the product of the truth values of the conjuncts." It follows that every classical pair distribution $p\in C_n$ is nothing but a weighted average of all possible truth assignments to the propositions $s_1,...,s_n$. (This will remain valid when we consider higher multiple distributions p_{ijk}, p_{ijkl}, etc.) If $p\in Q_n$ is in the interference region this observation ceases to be true. In this case p is a weighted average of extreme matrices, some of which do not correspond to classical truth values (we may have $u_{ii} = u_{jj} = 1$, but $u_{ij} = 0$). The fact that $Q_n \supset \text{int}(L_n)$ and the analogy with the classical case compel one to consider, at least tentatively, the idea that all the extreme matrices, not just the classical ones, correspond to "truth values" in some extended sense. This means, among other things, that one should replace the rule "the truth rule of a conjunction is the product of the truth values of the conjuncts" by "the truth value of a conjunction is less or equal to the product of the truth values of the conjuncts." This is a radical move. It entails, for example, that the truth value of a complicated proposition is not a (one-valued) function of the truth values of its parts. If one extends the concept of "truth" along these lines one obtains a new propositional logic, "quantum logic," which I shall discuss below. Before doing that, however, I should point to an important *conceptual* difference between classical and quantum logic. In the classical case the concept of "truth" is usually taken to be more profound and basic than the concept "probability." The latter is a derived concept, which presupposes the validity of the classical propositional calculus (as the entire body of mathematics presupposes it). This can be seen easily when we observe that probabilities are weighted averages of truth values and thus "probability" comes to represent our ignorance of the actual state of affairs. (This interpretation is apparent in classical statistical mechanics. One cannot measure the initial conditions of $\sim 10^{23}$ particles and solve the same number of differential equations, though a solution is known to exist. The way to circumvent this ignorance is to take the weighted average of all the possible states given by the Maxwell–Boltzmann formula.)

It is difficult to conceive of the general quantum case in the same way simply because the nonclassical extreme matrices *are not* in Q_n that is, no quantum system can ever realize those "truth functions." Since Q_n is densed in L_n one can approximate by probability every such "truth value" but no more. Hence quantum logic is an idealization or "limit case" of quantum probability. The former presupposes the latter and not the other way around.

Keeping this qualification in mind we can proceed to describe the new formal logic. As in the classical case, let $s_1,s_2,...,s_n,...$ be a set of symbols called atomic propositions.

The set Π of (well-formed) propositions is defined by induction on length of formulas as follows: (a) every atomic proposition is a proposition, (b) if a is a proposition so is $\neg a$ (read "not a"), and (c) if a,b are propositions so is $a\&b$ (read "a and b"). Here $a \vee b$ (read a or b) is a shorthand for $\neg(\neg a\&\neg b)$ and $a\rightarrow b$ ("a entails b") is a shorthand for $\neg a \vee b$.

We have already remarked that in quantum logic the truth value of a complicated proposition is not a function of the truth value of its atomic constituents. Thus it will not suffice to define the notion of "truth" for atomic propositions alone and we have to define it by induction on length of formulas as follows.

Definition 2.1: A quantum truth function is a function θ from the set of well-formed formulas Π into $\{0,1\}$ satisfying the following:

(i) for every atomic proposition s_i,

$$\theta(s_i) = 0 \text{ or } \theta(s_i) = 1;$$

(ii) $\theta(\neg a) = 1 - \theta(a)$;

(iii) $\theta(a\&b) = \theta(b\&a) \leqslant \theta(a)\theta(b), \quad \theta(a\&a) = \theta(a)$;

(iv) if $\theta'(b) \leqslant \theta'(c)$ for all quantum truth functions θ' defined for b and c, then $\theta(a\&b) \leqslant \theta(a\&c)$, for all $a \in \Pi$.

The inductive character of this definition is evident from (iv). From the above axioms we have by definition

$$\theta(a \vee b) = 1 - \theta(\neg a\&\neg b) \geqslant \theta(a) + \theta(b) - \theta(a)\theta(b)$$

and also $\theta(a\rightarrow b) = \theta(\neg a \vee b)$. Hence $\theta(a\rightarrow b) = 0$ entails $\theta(a) = 1$ $\theta(b) = 0$ (but not necessarily the other way around). If equality obtains in (iii) for all $a,b \in \Pi$ then θ is a classical truth function.

A classical tautology (logical falsity) is a proposition a such that $\theta(a) = 1$ [resp. $\theta(a) = 0$] for all classical truth values θ. Hence a quantum tautology (falsity) is a proposition for which $\theta(a) = 1$ [resp. $\theta(a) = 0$] for all quantum truth functions. Therefore there are fewer quantum tautologies and falsities then classical ones. In order to decide whether a given proposition $a \in \Pi$ is a classical tautology it is sufficient to check all 2^n possible truth value assignments to its atomic constituents $s_1,...,s_n$. This is no longer true in the quantum case but the number of quantum truth assignments is nevertheless bounded by 2^k, where k is the number of well-formed subpropositions of a. [This also means that the inductive definition (Definition 3.1) can be effectively applied.]

Many classical tautologies are in fact quantum tautologies. These include $a \vee \neg a$, $a \rightarrow a \vee b$, $a\&b \rightarrow a$, $\neg\neg a \leftrightarrow a$, $\neg(a\&b) \leftrightarrow (\neg a \vee \neg b)$, (De Morgan rule), and $(a\&b) \vee (a\&c) \rightarrow a\&(b \vee c)$. The proof of the last two tautologies is based on a repetitive use of the inductive rule (iv). Note, however, that $a\&(b \vee c) \rightarrow (a\&b) \vee (a\&c)$ is not a quantum tautology though it is a classical one. (To see that, it suffices to consider atomic propositions a,b,c.) Therefore the distributivity law is not a quantum tautology. Indeed it was the nondistributivity of the lattice of closed subspaces of a Hilbert space that led Birkhoff and Von Neumann[12] to their (admittedly heuristic) argument that quantum mechanics may "force" upon us a novel logic. With the identification of extreme matrices as "truth functions" this conclusion becomes more motivated. Definition 3.1 does not suffice to derive all the logical relations that obtain in the lattice of closed subspaces of a Hilbert space and it is not clear whether any finite or even recursive set of rules will achieve that. In this respect Definition 3.1 is only minimal.

III. QUASICLASSICAL MODELS OF QUANTUM PROBABILITY

A. The geometric analogy

Various pair distributions observed in microphenomena and predicted by quantum theory are not classical and this fact clashes with a strong "realistic" intuition regarding probability. It appears *prima facie* that one cannot reconstruct quantum probability if one assumes the "balls in an urn" picture that we usually associate with this notion. The question that I shall address in this section is whether we can nevertheless extend the classical concept of probability to cover the whole range of Q_n, while still retaining the essential realistic aspects of the classical notion. Can we, in other words, conceive of physical objects with fixed properties distributed so as to give rise to deviant probabilities on the phenomenal level?

A somewhat analogous problem faced mathematicians a century ago—I refer to non-Euclidean geometry. After centuries of futile attempts to derive Euclid's fifth postulate (the parallel axiom) from the other four postulates Lobatchevsky decided to turn the tables and assume the validity of a negation of that postulate. He obtained a *formal* system, hyperbolic geometry, which appeared to be consistent, though Lobatchevsky was not able to prove that. The completeness theorem of predicate logic (due to K. Gödel) states that a formal system of axioms is consistent if and only if it has a model. So the problem of consistency boils down to the following question: Can one conceive of geometric objects (call them "lines") whose properties and relations exemplify the formal properties and relations postulated in hyperbolic geometry? The affirmative answer was given by F. Kline. He constructed a so-called "Euclidean model of hyperbolic geometry" where hyperbolic lines are in fact segments of Euclidean curves and the Euclidean metric is replaced by another metric (which is nevertheless defined in terms of the former).

In our case we face no formal problem of consistency but the question is otherwise similar. We want to construct a "Kolmogorovian (i.e., classical) model of quantum probability" where "events" are subsets of a given set so that the family of all events forms a Boolean algebra, and where "probability" is the measure of the relative size of these sets. These events will play a role analogous to that of the hyperbolic lines and the "probability measure" will play a role analogous to that of the hyperbolic metric.

On a less formal level this will enable us to conceive of quantum probability in terms of a "balls in an urn" model. Surely we cannot expect to achieve our goal with finite sets of objects (balls) since the laws of classical probability in the discrete case are simple consequences of arithmetic. When we move from the finite domain to the continuum an appropriate model can be constructed. I shall proceed first with the formal properties of the model and discuss its physical interpretation immediately after.

B. Representing L_n by outer measure

Consider the (classical) probability space $([0,1],\Sigma,m)$, where Σ is the family of Lebesgue-measurable subsets of the interval $[0,1]$ and m the Lebesgue measure. One of the consequences of the axiom of choice is that not every subset of $[0,1]$ is an element of Σ. With a given arbitrary set $A \subseteq [0,1]$ we can nevertheless always associate an outer measure:

$$\bar{m}(A) = \inf\{m(G);\ G \supseteq A,\ G \text{ open}\}. \qquad (3.1)$$

If A is Lebesgue measurable, then $\bar{m}(A) = m(A)$. Note that the outer measure is subadditive and not always additive, that is if $A_1,A_2 \subseteq [0,1]$, $A_1 \cap A_2 = \varnothing$, then

$$\bar{m}(A_1 \cup A_2) \leqslant \bar{m}(A_1) + \bar{m}(A_2) \qquad (3.2)$$

and a sharp inequality may obtain when $A_1,A_2 \notin \Sigma$. This property is desirable. If E_1,E_2 are two projections in a Hilbert space H such that $E_1 \wedge E_2 = 0$ and if W is a density operator on H we have

$$\mathrm{tr}[W(E_1 \vee E_2)] \leqslant \mathrm{tr}(WE_1) + \mathrm{tr}(WE_2) \qquad (3.3)$$

and the inequality may be sharp (a sufficient condition for equality is that $[E_1,E_2] = 0$, which means in this case $E_1 \perp E_2$). Another formal property that indicates that the outer measure can serve as a model for quantum probability is given by the following theorem.

Theorem 3.1: There exists a decomposition of the interval $[0,1]$ into a continuum of pairwise disjoint subsets each having outer measure 1.

The proof involves the axiom of choice.[13] Needless to say, none of the sets whose existence is postulated in the above theorem is Lesbesgue measurable. So let $A_1,A_2 \subseteq [0,1]$ be two subsets such that $A_1 \cap A_2 = \varnothing$, $\bar{m}(A_1) = \bar{m}(A_2) = 1$. The matrix $p_{ij} = \bar{m}(A_i \cap A_j)$, for $i,j = 1,2$, is just $\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)$, the only nonclassical extreme matrix of L_2. This is, of course, no accident.

Definition 3.1: A matrix $p \in L_n$ has an outer measure representation if there are subsets $A_1,A_2,\ldots,A_n \subseteq [0,1]$ such that $p_{ij} = \bar{m}(A_i \cap A_j)$.

Let O_n be the set of all $p \in L_n$ which have an outer measure representation, then we have the following theorem.

Theorem 3.2: $O_n = L_n$.

Proof: I shall show first that O_n is convex and then that every extreme matrix $u \in \mathscr{E}_n$ is an element of O_n. This will complete the proof since $L_n \supseteq O_n \supseteq \mathrm{co}(\mathscr{E}_n) = L_n$. Let $p,p' \in O_n$ and $0 < \lambda < 1$. I shall show that $\lambda p + (1-\lambda)p' \in O_n$. By definition there are sets $A_1,\ldots,A_n, A'_1,\ldots,A'_n \subseteq [0,1]$ such that $p_{ij} = \bar{m}(A_i \cap A_j)$, $p'_{ij} = \bar{m}(A'_1 \cap A'_j)$, for $i,j = 1,2,\ldots,n$. Consider the function $f(t) = \lambda t$. It maps the interval $[0,1]$ onto the interval $[0,\lambda]$. For $B \subseteq [0,1]$ let $f(B) = \{f(E);\ t \in B\}$. Since f is linear we have $f(B \cap C) = f(B) \cap f(C)$ and, for every $B \subseteq [01]$, $\bar{m}(f(B)) = \lambda \bar{m}(B)$ (this can be proved first for open intervals, then for open sets and the claim follows). Put $C_i = f(A_i)$. Then from the above remark we obtain $\bar{m}(C_i \cap C_j) = \lambda(A_i \cap A_j) = \lambda p_{ij}$.

Consider the function $g(t) = \lambda + (1-\lambda)t$. From a similar argument we get, for $C'_i = g(A'_i) = \{g(t);\ t \in A'_i\}$,

$$\bar{m}(C'_i \cap C'_j) = (1-\lambda)\bar{m}(A'_i \cap A'_j) = (1-\lambda)p'_{ij}.$$

Now put $D_i = C_i \cup C'_i$. Since $C_i \subseteq [0,\lambda]$ and $C'_i \subseteq [\lambda,1]$, for $i = 1,2,\ldots,n$, we have

$$\bar{m}(D_i \cap D_j) = \bar{m}(C_i \cap C_j) + \bar{m}(C'_i \cap C'_j)$$
$$= \lambda p_{ij} + (1-\lambda)p'_{ij},$$
$$\text{for } i = 1,2,\ldots,n,$$

and thus O_n is convex.

Let $u \in \mathscr{E}_n$ be an extreme matrix. We shall prove that $u \in O_n$. By Theorem 3.1, we can decompose interval $[0,1]$ into n^2 pairwise disjoint subsets each having outer measure 1. Denote these sets by $B_{11},\ldots,B_{1n},B_{21},\ldots,B_{2n},\ldots,B_{n1},\ldots,B_{nn}$. We have $B_{ij} \cap B_{kl} = \varnothing$ for $(ij) \neq (kl)$ and $\bar{m}(B_{ij}) = 1$, for $i,j = 1,2,\ldots,n$. For an arbitrary set $B \subseteq [0,1]$, put $B^0 = \varnothing$ and $B^1 = B$ and define

$$A_i = \bigcup_{k=1}^{n} B_{ik}^{u_{ik}} \cup B_{ki}^{u_{ki}}.$$

Then $\bar{m}(A_i) = 0$ if and only if $u_{ik} = u_{ki} = 0$, for all $k = 1,2,\ldots,n$. Since $u_{ki} \leqslant u_{ii}$ this occurs if and only if $u_{ii} = 0$. Let $i \neq j$. Then $A_i \cap A_j = B_{ij}^{u_{ij}} \cup B_{ji}^{u_{ji}}$ then $\bar{m}(A_i \cap A_j) = 0$ iff $u_{ij} = 0$. Since in all other cases, $\bar{m}(A_i \cap A_j) = 1$, we conclude that $\bar{m}(A_i \cap A_j) = u_{ij}$, for $i,j = 1,2,\ldots,n$. Q.E.D.

C. Outer measure and probability

I have proved that every $p \in L_n$ (and in particular every $p \in Q_n$) can be represented by subsets of $[0,1]$ together with their outer measure. On a less formal level I shall indicate how this fact can be interpreted in terms of events and their probabilities. Rather then providing an actual physical example[14] I shall describe an artificial setup whose advantage lies in its simplicity. I hope this will convince the reader that a "balls in an urn" model of quantum probability is possible.

The "urn" is just the interval $[0,1]$ and the "balls" are its points. Assume that every point has a definite fixed color red or blue (not both). This means that every point radiates either red or blue light. Consider first the classical case where the set of red points A and the set of blue points B are Lebesgue measurable. By definition, $A \cap B = \varnothing$, $A \cup B = [0,1]$. Suppose that the intensity of red (blue) light from an interval is proportional to the measure of the set of red (blue) points in that interval. If we choose normalization such that the total intensity (red + blue) is 1 we obtain that the intensity of red radiation is $m(A)$ and the intensity of blue radiation is $m(B)$. These intensities can be measured by introducing appropriate filters. (We filter out the blue light when we measure the intensity of red and vice versa.) Suppose now that A,B are not measurable and moreover that the rule associated with intensities now reads "the intensity of (red, blue, or any) light from an interval is proportional to the *outer* measure of the set of radiating points in that interval." If we keep the normalization convention as above we see that intensities are no longer additive, the total intensity is 1 while the red and blue intensities are $\bar{m}(A)$, $\bar{m}(B)$ and we may have $\bar{m}(A) + \bar{m}(B) > 1$. In particular if $\bar{m}(A) = \bar{m}(B) = 1$ we obtain the following strange situation: If we introduce both filters we detect no radiation ($A \cap B = \varnothing$); if we introduce any one of the filters no reduction in the radiation intensity is observed.

This is of course only a metaphor. The question is what can prevent us from associating with every physical system a fixed set of "hidden properties" (call them "colors") and interpret every physical experiment as a process by which some of these properties are filtered out. If the distribution of these properties is sufficiently wild we shall obtain nonclassical probabilities on the phenomenal level. In light of Theorem 3.2, this can be done for every $p\in L_n$, that is, for every conceivable experimental result. I do not claim that one ought to interpret quantum mechanics in this way or even that it makes physical sense to do so. My only claim is that one *can* interpret microphysical phenomena by such "hidden variables" and obtain a perfectly consistent picture.

IV. CONCLUSION

What the previous discussion demonstrates, I believe, is that the problem of realism, the reality of microphysical properties, is irrelevant to the understanding of quantum mechanics. One can, *if one wishes*, conceive of quantum probabilities as resulting from distributions of fixed properties. One can also decline to do so. *In that respect quantum mechanics is not different from any classical physical theory.* One can conceive of classical gravitational fields, say, as real objects "feeling up" space (a realistic attitude) or one can think of gravitational fields as mathematical tools whose purpose is to systematize and organize a complicated set of phenomena (an instrumentalistic approach). In classical as in quantum physics, both alternatives are compatible with the experimental results. The problem of physical realism is a metaphysical problem and no observation or experiment bears direct relevance to it. An argument for or against realism is thus inherently a metaphysical argument and should be evaluated accordingly. To claim otherwise is to commit what philosphers call a category mistake.

Bohr's interpretation, I believe, falls in this trap. His unnecessary focus on the metaphysical problems blurs the issue at hand. A careful rational analysis would reveal, I believe, that only two coherent alternatives exist.

(a) "Interference" and "collapse of the wave function" are physical processes caused by unknown physical mechanisms. One can take a realistic or instrumentalistic approach regarding that "mechanism" but regardless of one's attitude one can hope to control and manipulate these processes and obtain results which transcend or even contradict quantum mechanics.

(b) "Inteference" and "collapse of the wave function" are names for the fact that the probability theory associated with quantum theory is nonclassical. As in the case of space-time geometry one can understand this probability theory in realistic or in instrumentalistic terms, but, regardless of one's views, one should try and understand the properties and extension of this new strange notion.

No decision between these two alternatives currently exists but the question of which is the correct interpretation is in large part empirical, that is, if alternative (a) is the correct one we shall find it out sooner or later.[15]

[1] A. Grag and N. D. Mermin, Found. Phys. **14**, 1 (1984).

[2] R. T. Rockafellar, *Convex Analysis* (Princeton U.P., Princeton, NJ, 1970).

[3] A. M. Gleason, J. Math. Mech. **6**, 885 (1957).

[4] Reference 2, p. 46.

[5] Reference 2, Secs. 21 and 22.

[6] J. S. Bell, Physics (NY) **1**, 195 (1964); E. P. Wigner, Am. J. Phys. **38**, 1005 (1970).

[7] J. F. Clauser and M. A. Horne, Phys. Rev. D **10**, 526 (1974).

[8] Reported by N. Bohr, in "Discussions with Einstein on epistemological problems in atomic physics," in P. A. Schilpp, *Albert Einstein Philosopher Scientist* (Library of Living Philosophers, Evanston, IL, 1949).

[9] R. P. Feynman and A. K. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).

[10] R. P. Feynman, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (U. California Press, Berkeley, 1951).

[11] L. Accardi, Phys. Rep. **77**, 169 (1981); S. P. Gudder, Proc. Am. Math. Soc. **21**, 557 (1973); Found. Phys. **3**, 399 (1980).

[12] G. Birkhoff and J. Von Neumann, Ann. Math. **37**, 823 (1936); S. Kochen and E. P. Specker, J. Math. Mech. **17**, 59 (1967).

[13] E. Hewitt and K. A. Ross, *Abstract Harmonic Analysis* (Springer, Berlin, 1979), Lemma 16.7, p. 218. Without the axiom of choice one can construct a model of the reals in which every subset is Lebesgue measurable, see R. M. Solovey, Ann. Math. **92**, 1 (1970).

[14] Concrete physical models constructed along similar lines are in I. Pitowsky, Phys. Rev. Lett. **48**, 1216, 1299 (1982); Phys. Rev. D **27**, 2316 (1983); *Proceedings of the Conference on Fundamental Problems in Quantum Mechanics*, Albany, 1984 (to appear); S. P. Gudder, J. Math. Phys. **25**, 2397 (1984); Found. Phys. (1984). The example below appeared in I. Pitowsky, Synthese **63**, 233 (1985).

[15] These alternative views resemble the problem of the conventionality of space-time geometry, see I. Pitowsky, Philos. Sci. **51**, 685 (1984).

Non-Abelian Aharonov–Bohm effects, Feynman paths, and topology

Raman Sundrum[a] and L. J. Tassie
Department of Theoretical Physics, Research School of Physical Sciences, Australian National University, G. P. O. Box 4, Canberra, Australian Capital Territory 2601, Australia

(Received 3 September 1985; accepted for publication 3 January 1986)

The Aharonov–Bohm effect in general gauge theories, for particles in gauge-curvature-free regions, is studied using the quantum mechanical propagator in the form of a Feynman sum over paths. Following Schulman [L. S. Schulman, *Techniques and Applications of Path Integration* (Wiley, New York, 1981)], such paths are divided into their homotopy equivalence classes, and the contributions from each class of paths of the Feynman sum are identified with propagators of a wave equation in the universal covering manifold of M, resulting in a simple form for the propagator on M. A group homomorphism from $\mathscr{H}$, the fundamental homotopy group of M, to the gauge group G is shown to characterize possible Aharonov–Bohm effects, which can be divided into two types, Abelian and non-Abelian, according to whether $\mathscr{H}^*$, the image of this homomorphism, is Abelian or non-Abelian. For a non-Abelian Aharonov–Bohm effect, it is necessary that both $\mathscr{H}$ and G be non-Abelian. Simple examples illustrate the theory.

I. INTRODUCTION

In this paper, we study a particle moving in a multiply connected manifold M, under the influence of a gauge potential for a matrix gauge group G, given by the one-form $\mathbf{b}^k X_k \cdot d\mathbf{l}$, where the X_k are generators of G. Here G is taken to act on particle wave functions by matrix multiplication. It should be noted that M describes only the spatial position of the particle, time being treated separately, $\mathbf{b}^k X_k \cdot d\mathbf{l}$, representing only the spatial components of a gauge potential, the time component (scalar potential) assumed zero. The gauge potential is also assumed to have zero gauge curvature.

$$D(\mathbf{b}^k X_k \cdot d\mathbf{l}) = 0. \tag{1.1}$$

Such a situation occurs when the particle is excluded from some region of space in which magnetic fluxes are confined. The physically observable difference between the case $\mathbf{b}^k X_k \cdot d\mathbf{l} \neq 0$ and the free-particle case $\mathbf{b}^k X_k \cdot d\mathbf{l} = 0$ is the Aharonov–Bohm (hereafter abbreviated by AB) effect.[1]

We use the Feynman approach to quantum mechanics, in which the propagator is given as a sum over paths. Following Schulman,[2] the paths are partitioned into their homotopy equivalence classes, Feynman sums over paths in each class giving *homotopy propagators*, the whole effect of the gauge potential being to multiply these homotopy propagators by different gauge phase factors. For a simply connected space, all paths between two points are in the same homotopy class, and the effect of the potential is to multiply the free-particle propagator by a single gauge phase factor, so the potential has no physical effect. For a multiply connected manifold, the potential can have a physical effect because the gauge phase factors can be different for different homotopy classes.

The homotopy propagators are related to propagators on the universal covering manifold of M, leading to an expansion of the propagators in terms of eigenfunctions of a Hamiltonian on the covering manifold.

It is shown that an AB effect is characterized by a homomorphism from the fundamental homotopy group of M to the holonomy subgroup of G, the effect being Abelian or non-Abelian depending on whether the holonomy group is Abelian or not. The effect proposed by Yang and Wu[3] for $G = \mathrm{SU}(2)$ is shown to be Abelian despite the fact that G is non-Abelian.

To illustrate the theory, the case originally dealt with by Aharanov and Bohm[1] of a charge near a long straight solenoid is used. Also an SU(3) gauge potential is given to show how a non-Abelian AB effect might arise.

Units in which $\hbar$, c, and the gauge coupling constant are set to 1, are used throughout.

II. THE PROPAGATOR

The Feynman sum form of the nonrelativistic propagator for the motion of a particle under the influence of the spatial components of a gauge potential $\mathbf{b}^k X_k \cdot d\mathbf{l}$ ($b_0^k X_k$, the time component assumed zero), is taken as

$$K_b(x',t';x,t) = \int_{x,t}^{x',t'} \mathscr{D}\Gamma \exp\left(\int_\Gamma \mathbf{b}^k X_k \cdot d\mathbf{l}\right) \exp(iS(\Gamma)), \tag{2.1}$$

a sum over all paths Γ: $[t,t'] \to M$ from x to x' $[\Gamma(t) = x, \Gamma(t') = x']$, where S is the free-particle classical action on particle trajectories, and

$$\exp\left(\int_\Gamma \mathbf{b}^k X_k \cdot d\mathbf{l}\right)$$

is Yang's gauge phase factor,[3,4] generalizing the phase factor used in electromagnetism.[5] It is always considered as path (or time) ordered. The propagator satisfies

$$K_b(x',t;x,t) = \delta_M(x',x) \tag{2.2}$$

and

$$\left(i\frac{\partial}{\partial t'} - H_b\right) K_b(x',t';0,t) = 0, \quad t' > t, \tag{2.3}$$

where δ_M is the delta function on M and H_b is the Hamiltonian operator on M, acting on the first variable x'. Here K_0 and H_0 are used to denote K_b and H_b when $\mathbf{b}^k X_k \cdot d\mathbf{l} = 0$.

[a] Present address: Department of Physics, Yale University, New Haven, Connecticut 06520.

III. HOMOTOPY PROPAGATORS

Following Schulman,[2] we rewrite the propagator in the form

$$K_b(x',t';x,t) = \sum_{f\in F(x',t';x,t)} \int_{x,t}^{x',t'} \mathscr{D}_f\Gamma \exp\left(\int_\Gamma \mathbf{b}^k X_k \cdot d\mathbf{l}\right) \exp(iS(\Gamma)), \tag{3.1}$$

where $F(x_1,t_1;x,t)$ is the collection of all homotopy classes of paths Γ: $[t,t'] \to M$ from x to x', and $\mathscr{D}_f\Gamma$ indicates a Feynman sum over only those paths in class f. The homotopy class of a path Γ is denoted $[\Gamma]$. Physically, $[\Gamma]$ is important because for gauge fields where (1.1) holds, the gauge phase factor depends only on $[\Gamma]$. That is, if $\Gamma' \in [\Gamma]$, then

$$\exp\left(\int_{\Gamma'} \mathbf{b}^k X_k \cdot d\mathbf{l}\right) = \exp\left(\int_\Gamma \mathbf{b}^k X_k \cdot d\mathbf{l}\right). \tag{3.2}$$

This is proven in the Appendix. In view of this fact, we can define, for $f \in F(x',t';x,t)$,

$$\exp\left(\int_f \mathbf{b}^k X_k \cdot d\mathbf{l}\right) = \exp\left(\int_\Gamma \mathbf{b}^k X_k \cdot d\mathbf{l}\right), \tag{3.3}$$

for any $\Gamma \in f$ (i.e., $f = [\Gamma]$). Then (3.1) becomes

$$K_b(x',t';x,t) = \sum_{f\in F(x',t';x,t)} \exp\left(\int_f \mathbf{b}^k X_k \cdot d\mathbf{l}\right) K^f(x',t';x,t), \tag{3.4}$$

where the *homotopy propagator of class f* is defined by

$$K^f(x',t';x,t) = \int_{x,t}^{x',t'} \mathscr{D}_f\Gamma \exp(iS(\Gamma)). \tag{3.5}$$

Thus the AB effect is obtained by multiplying the homotopy propagators by gauge phase factors. In general, these factors will vary for different homotopy classes, permitting $|K_b|^2 \neq |K_0|^2$, and producing an observable effect.

IV. THE COVERING MANIFOLD OF *M*

Let C be the universal covering manifold of M, with covering projection π: $C \to M$. Again following Schulman[2] in noting the correspondence under projection between paths in M of a particular class $f \in F(x',t';x,t)$ and paths in C from $y \in \pi^{-1}(x)$ to a particular point $y'_f \in \pi^{-1}(x')$,

$$K^f(x',t';x,t) = \int_{y,t}^{y'_f,t'} \mathscr{D}_\gamma \exp(iS(\pi\gamma)), \tag{4.1}$$

where the right-hand side is now a sum over all paths in C from y to y'_f. Also from the correspondence we can define

$$g_b(y',y) = \exp\left(\int_{\pi\gamma} \mathbf{b}^k X_k \cdot d\mathbf{l}\right) \tag{4.2}$$

for any path γ in C, from y to y', two points in C. (Note the line integral above is performed over path $\pi\gamma$ that lies in M.) Equation (3.4) then becomes

$$K_{\mathbf{b}}(x',t';x,t) = \sum_{f\in F(x',t';x,t)} g_b(y'_f,y) \int_{y,t}^{y'_f,t} \mathscr{D}\gamma \exp(iS(\pi\gamma))$$
$$= \sum_{y'\in\pi^{-1}(x)} g_b(y',y) \int_{y,t}^{y',t'} \mathscr{D}\gamma \exp(iS(\pi\gamma)), \tag{4.3}$$

for any $y \in \pi^{-1}(x)$. Defining a propagator on C,

$$K^c(y',t';y,t) = \int_{y,t}^{y',t'} \mathscr{D}\gamma \exp(iS(\pi\gamma)), \tag{4.4}$$

then gives

$$K_b(x',t';x,t) = \sum_{y'\in\pi^{-1}(x')} g_b(y',y) K^c(y',t';y,t), \tag{4.5}$$

for any $y \in \pi^{-1}(x)$.

The free-particle Hamiltonian H_0, being a local operator on M, can be lifted[2] to an operator on C, which we still denote by H_0. Then K^c satisfies the free wave equation[2]

$$\left(i\frac{\partial}{\partial t'} - H_0\right) K^c(y',t';y,t) = 0, \quad t' > t, \tag{4.6}$$

on C, where H_0 acts on the first variable y'. By the Feynman sum form (4.4), of K^c, it is clear that

$$K^c(y',t;y,t) = \delta_c(y',y). \tag{4.7}$$

Equations (4.6) and (4.7) then tell us that[6]

$$K^c(y',t';y,t) = \sum_n \psi_n^*(y)\psi_n(y') e^{-iE_n(t'-t)}, \tag{4.8}$$

where the ψ_n are any orthonormal, complete set of eigenfunctions of H_0, on C, with eigenvalues E_n, and Σ_n denotes summation/integration over discrete/continuous indices n. By (4.5) and (4.8),

$$K_b(x',t';x,t) = \sum_{y'\in\pi^{-1}(x')} g_b(y',y) \sum_n \psi_n^*(y)\psi_n(y') e^{-iE_n(t'-t)}. \tag{4.9}$$

V. THE FUNDAMENTAL HOMOTOPY GROUP

Choose a particular $y'_0 \in \pi^{-1}(x')$. Then if γ is a path in C from y to y'_0, and ω_y is another path from y'_0 to $y' \in \pi^{-1}(x')$, then $\gamma\omega$ is a path from y to y'. So, from (4.2),

$$g_b(y',y) = \exp\left(\int_{\pi(\gamma\omega_y)} \mathbf{b}^k X_k \cdot d\mathbf{l}\right)$$
$$= \exp\left(\int_{(\pi\gamma)(\pi\omega_y)} \mathbf{b}^k X_k \cdot d\mathbf{l}\right), \tag{5.1}$$

and since our gauge phase factors are path ordered, (5.1) can be written

$$g_b(y',y) = \exp\left(\int_{\pi\gamma} \mathbf{b}^k X_k \cdot d\mathbf{l}\right) \exp\left(\int_{\pi\omega_y} \mathbf{b}^k X_k \cdot d\mathbf{l}\right). \tag{5.2}$$

Substituting into (4.5) yields

$$K_b(x',t';x,t) = \exp\left(\int_{\pi\gamma} \mathbf{b}^k X_k \cdot d\mathbf{l}\right) \times \sum_{y'\in\pi^{-1}(x')} \exp\left(\int_{\pi\omega_y} \mathbf{b}^k X_k \cdot d\mathbf{l}\right) K^c(y't';y,t). \tag{5.3}$$

Now if the particle is known to be at x at time t, the probability for it to be at x' at time t' is

$$|K_b(x',t';x,t)|^2$$
$$= \left| \sum_{y'\in\pi^{-1}(x')} \exp\left(\int_{\pi\omega_{y'}} \mathbf{b}^k X_k \cdot d\,\mathbf{l}\right) K^c(y',t';y,t) \right|^2. \tag{5.4}$$

Thus the whole AB effect is produced by the gauge phase factors

$$\left\{ \exp\left(\int_{\pi\omega_{y'}} \mathbf{b}^k X_k \cdot d\,\mathbf{l}\right) \middle| y'\in\pi^{-1}(x') \right\}.$$

Because these elements of the gauge group characterize any particular AB experiment being done, we proceed to investigate them.

First notice that by (3.3)

$$\exp\left(\int_{\pi\omega_{y'}} \mathbf{b}^k X_k \cdot d\,\mathbf{l}\right) = \exp\left(\int_{[\pi\omega_{y'}]} \mathbf{b}^k X_k \cdot d\,\mathbf{l}\right). \tag{5.5}$$

Here $\pi\omega_{y'}$ is a closed path in M from x' to itself. Therefore $[\pi\omega_{y'}]\in F(x',t';x,t)$. By the correspondence between paths in C and paths in M, as y' ranges over $\pi^{-1}(x')$, $[\pi\omega_{y'}]$ ranges over all of $F(x',t';x,t)$. So the subset of gauge phase factors characterizing the AB effect in (5.4) is

$$\left\{ \exp\left(\int_f \mathbf{b}^k X_k \cdot d\,\mathbf{l}\right) \middle| f\in F(x',t';x,t) \right\}.$$

A natural group structure can be put on $F(x',t';x,t)$, yielding $\mathscr{H}$, the *fundamental homotopy group*[7] of M. The map

$$f\in\mathscr{H}\mapsto \exp\left(\int_f \mathbf{b}^k X_k \cdot d\,\mathbf{l}\right)$$

is a group homomorphism[8] from the fundamental homotopy group into the gauge group G. The set of phase factors characterizing the AB effect in (5.4) is the image of $\mathscr{H}$ under the homomorphism, and so a subgroup of G,

$$\mathscr{H}^* = \left\{ \exp\left(\int_f \mathbf{b}^k X_k \cdot d\,\mathbf{l}\right) \middle| f\in\mathscr{H} \right\} \leqslant G. \tag{5.6}$$

Also $\mathscr{H}^*$ is just the holonomy group of the trivial vector bundle over M, with connection $\mathbf{b}^k X_k \cdot d\,\mathbf{l}$.

We use the group $\mathscr{H}^*$ to classify the various AB experiments. In particular we have two types:

(1) $\mathscr{H}^*\leqslant G$ (Abelian),
(2) $\mathscr{H}^*\leqslant G$ (non-Abelian).

The first type contains, for example, all experiments involving only electromagnetism, all subgroups of U(1) being Abelian. The experiment originally suggested by Wu and Yang[3] and performed by Zeilinger *et al.*,[9] of a nucleon outside a single tube of isotopic spin magnetic flux is also of the first type, since $\mathbb{R}^2$ with the region of flux removed has $\mathscr{H}$ isomorphic to the integers, an Abelian group. It follows that $\mathscr{H}^*$, the image of $\mathscr{H}$ under a homomorphism, is an Abelian subgroup of the non-Abelian gauge group SU(2). The second type consists of experiments which exhibit true non-Abelian AB effects, since they are characterized by non-Abelian subgroups of the gauge group.

VI. EXAMPLE A: U(1) AB EFFECT NEAR A THIN STRAIGHT MANGETIC FLUX

We check our theory by applying it to the simple example, first treated by Aharonov and Bohm,[1] of a charge excluded from an infinitely long straight cylinder of negligible radius, containing a magnetic flux $2\pi\alpha$. We are interested here only in the two dimensions perpendicular to the flux. So M is to be $\mathbb{R}^2$ with one point removed and the magnetic flux confined to the removed point, which we take as the origin of polar coordinates; C is $(0,\infty)\times\mathbb{R}$, with coordinates (r,θ). The projection π: $C\rightarrow M$ is given by

$$(r,\theta+2\pi m)\mapsto(r,\theta), \quad \text{for } 0\leqslant\theta<2\pi.$$

The vector potential is given by

$$i\mathbf{A}(r,\theta) = i(\alpha/r)\hat{\boldsymbol{\theta}}. \tag{6.1}$$

Let $(\rho(\lambda),\phi(\lambda))$ be a path in C from (r,θ) at t to (r',θ') at t'. By (4.2) and (6.1),

$$g_A((r',\theta')) = \exp\left(\int_{\pi(\rho,\phi)} i\mathbf{A}\cdot d\,\mathbf{l}\right)$$
$$= \exp(i\alpha(\theta'-\theta)). \tag{6.2}$$

The Schrödinger Hamiltonian is

$$H_A = -\frac{1}{2m}\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\left(\frac{\partial}{\partial\theta} - i\alpha\right)^2. \tag{6.3}$$

Notice H_0 is an operator on either M or C. A complete, orthonormal set of eigenfunctions in C, of H_0, is

$$\{J_{|\lambda|}(pr)e^{i\lambda\theta}\,|p\geqslant0,\lambda\in\mathbb{R}\},$$

with eigenvalues $p^2/2m$, where the $J_{|\lambda|}$ are Bessel functions. Note, we do not require $\lambda\in\mathbb{Z}$, since periodicity in θ is not required in the covering manifold C. By (4.9) and (6.2),

$$K_A((r',\theta'),t';(r,\theta),t) = \sum_{m\in\mathbb{Z}}\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_0^{\infty} J^*_{|\lambda|}(pr)J_{|\lambda|}(pr')\exp[i(\lambda+\alpha)(\theta'+2\pi m-\theta)]\exp[-i(p^2/2m)(t'-t)]\,d\lambda\,dp$$
$$= \sum_{l\in\mathbb{Z}}\frac{1}{2\pi}\int_0^{\infty} J^*_{|l-\alpha|}(pr)J_{|l-\alpha|}(pr)\exp[il(\theta'-\theta)]\exp[-i(p^2/2m)(t'-t)]\,dp, \tag{6.4}$$

since

$$\sum_{m\in\mathbb{Z}} e^{-2\pi i m(\lambda+\alpha)} = \sum_{l\in\mathbb{Z}}\delta(l-(\lambda+\alpha)).$$

Then (6.4) is just the expansion of the propagator in the complete orthonormal set of eigenfunctions on M of H_A,

$$\{J_{|l-\alpha|}(pr)e^{il\theta}\,|\,l\in\mathbb{Z},p\geqslant0\}.$$

Note that the eigenfunctions on M are periodic in θ, corresponding to the single valuedness of wave functions,[10] even though the eigenfunctions on C were not periodic in θ.

Berry[11] has considered this example using a method in which a single-valued wave function is given by a sum, in which the individual terms are not single values. Bernido and Inomata[12] and Gerry and Singh[13] also considered this example, evaluating Feynman sums explicitly. Schulman[14] was the first to take the approach that we have, but did not explicitly relate the resulting propagator with eigenfunctions in M of the full Hamiltonian H_A.

VII. EXAMPLE B: SU(3) AB EFFECT NEAR TWO THIN, STRAIGHT MAGNETIC FLUXES

In this example we will display an SU(3) gauge potential on a manifold, whose gauge curvature vanishes, but whose holonomy group $\mathscr{H}^*$ is a non-Abelian subgroup of SU(3). If an experiment could be constructed, described by this potential, for example, then it would produce a true non-Abelian (type 2) AB effect.

Once again ignoring the third spatial dimension, we consider the potential in the x-y plane, given in Cartesian coordinates by

$$\mathbf{b}^k X_k = \Big[\alpha_1 \nabla \tan^{-1}\Big(\frac{y}{x+1}\Big) - \alpha_1 \nabla\ s(x+1)\tan^{-1}\Big(\frac{y}{x+1}\Big)\ \Big] X_1 + \Big[\alpha_2 \nabla \tan^{-1}\Big(\frac{y}{x-1}\Big) - \alpha_2 \nabla\Big(s(1-x)\tan^{-1}\Big(\frac{y}{x-1}\Big)\Big)\Big] X_2, \tag{7.1}$$

where $s(x)$ is any smooth function that is zero in a neighborhood of $x=0$ and $s(x)=1$ for all $x \geqslant 1$; X_1 and X_2 are any pair of noncommuting generators of SU(3). The potential is singular at the points $(\pm 1,0)$ in the x-y plane. We take our manifold M to the x-y plane with these two points removed. The gauge curvature vanishes on M.

Let Γ be a path in M from the origin to itself that encircles $(-1,0)$ once and does not enter the region $x>0$. Now by the properties of s, $b^2 = 0$ in the region $x \leqslant 0$, to which Γ belongs. Therefore

$$\exp\Big(\int_\Gamma \mathbf{b}^k X_k \cdot d\mathbf{l}\Big) = \exp\Big(\int_\Gamma \mathbf{b}^1 X_1 \cdot d\mathbf{l}\Big). \tag{7.2}$$

Now $s(x+1)\tan^{-1}(y/(x+1))$ is defined on the whole x-y plane, for although $\tan^{-1}(y/(x+1))$ is singular at $(-1,0)$, $s(x+1)=0$ in a neighborhood of $(-1,0)$. Therefore,

$$\int_\Gamma \alpha_1 \nabla\Big(s(x+1)\tan^{-1}\Big(\frac{y}{x+1}\Big)\Big)\cdot d\mathbf{l} = 0. \tag{7.3}$$

So

$$\int_\Gamma \mathbf{b}^1 \cdot d\mathbf{l} = \int_\Gamma \alpha_1 \nabla \tan^{-1}\Big(\frac{y}{x+1}\Big)\cdot d\mathbf{l}. \tag{7.4}$$

This is most easily evaluated by putting polar coordinates on the plane with the origin at $(-1,0)$. Then

$$\alpha_1 \nabla \tan^{-1}(y/(x+1)) = \alpha_1 \nabla\theta = (\alpha_1/r)\hat{\boldsymbol{\theta}}, \tag{7.5}$$

$$\begin{aligned}\int_\Gamma \mathbf{b}^1 X_1 \cdot d\mathbf{l} &= \int_\Gamma \frac{\alpha_1 \hat{\theta}}{r}\cdot(dr\,\hat{r} + r\,d\theta\,\hat{\theta})X_1 \\ &= \int_\Gamma \alpha_1\, d\theta X_1 \\ &= 2\pi\alpha_1 X_1,\end{aligned} \tag{7.6}$$

and we see that the gauge potential describes a thin X_1 magnetic flux passing through the plane at $(-1,0)$, from which our particle is excluded. Similarly, it describes a thin X_2 magnetic flux at $(1,0)$.

We have shown that $\exp(2\pi\alpha_1 X_1)$ is an element of $\mathscr{H}^*$, the holonomy group. By similar means it can be shown that $\exp(2\pi\alpha_2 X_2)$ is an element of $\mathscr{H}^*$ as well. Since X_1 and X_2 do not commute, α_1 and α_2 can be chosen so that $\exp(2\pi\alpha_1 X_1)$ and $\exp(2\pi\alpha_2 X_2)$ do not commute, making $\mathscr{H}^*$ non-Abelian.

VIII. CONCLUDING REMARKS

We have seen that the AB effect of the gauge potential in a gauge curvature free region M, is described simply by multiplication of the homotopy propagators by certain gauge phase factors. The homotopy propagators are related to propagators on the universal covering manifold C, and can be expressed in terms of eigenfunctions of the free particle Hamiltonian on C. We saw that AB effects are closely related to the topology of M, and for a non-Abelian AB effect it is necessary that $\mathscr{H}$, the fundamental homotopy group of M, be non-Abelian. The SU(2) AB experiment proposed by Wu and Yang[3] could only test for the presence of an Abelian subgroup of the gauge group. A true non-Abelian AB effect is required to indicate a non-Abelian gauge group.

APPENDIX: PROOF THAT THE GAUGE PHASE FACTOR DEPENDS ON ONLY THE HOMOTOPY CLASS OF A PATH

(i) First notice that, if ω is any closed path, from a point z to itself, then, if

$$\exp\Big(\int_\omega \mathbf{b}^k X_k \cdot d\mathbf{l}\Big) = I \quad \text{the identity,}$$

we must have

$$\exp\Big(\int_{\omega'} \mathbf{b}^k X_k \cdot d\mathbf{l}\Big) = I,$$

where ω' is the result of changing the initial (and final) point of ω, z, to some other point of the path, z' (that is, ω' traverses the same points as ω, but in a rotated order). The reason for this is as follows. Let γ be a path from z' to z, which is also a segment of the path ω. Clearly, $\gamma\omega\gamma^{-1}$ and $\omega'\gamma\gamma^{-1}$ have the same trajectories. Keeping in mind that gauge phase factors are path ordered,

$$\begin{aligned}&\exp\Big(\int_{\omega'} \mathbf{b}^k X_k \cdot d\mathbf{l}\Big) \\ &\quad = \exp\Big(\int_{\omega'} \mathbf{b}^k X_k \cdot d\mathbf{l}\Big)\exp\Big(\int_\gamma \mathbf{b}^k X_k \cdot d\mathbf{l}\Big) \\ &\quad\quad \times \exp\Big(\int_{\gamma^{-1}} \mathbf{b}^k X_k \cdot d\mathbf{l}\Big)\end{aligned}$$

$$= \exp\left(\int_{\omega'\gamma\gamma^{-1}} \mathbf{b}^k X_k \cdot d\mathbf{l}\right)$$

$$= \exp\left(\int_{\gamma\omega\gamma^{-1}} \mathbf{b}^k X_k \cdot d\mathbf{l}\right)$$

$$= \exp\left(\int_{\gamma} \mathbf{b}^k X_k \cdot d\mathbf{l}\right) I \exp\left(\int_{\gamma^{-1}} \mathbf{b}^k X_k \cdot d\mathbf{l}\right)$$

$$= I. \tag{A1}$$

(ii) Let Γ and Γ' be homotopically equivalent paths in M. Therefore $\Gamma'\Gamma^{-1}$ is a closed path. The equivalence of the paths implies there is a simply connected surface in M whose boundary is $\Gamma'\Gamma^{-1}$. Divide this surface up into many infinitesimal parallelograms, with boundaries ω_i. For each of these,[4]

$$\exp\left(\int_{\omega_i} \mathbf{b}^k X_k \cdot d\mathbf{l}\right) = I + f^k_{\mu\nu} X_k \, dx_i^\mu \, dx_i'^\nu, \tag{A2}$$

where the ith parallelogram has sides dx_i and dx_i' and $f^k_{\mu\nu}$ is the gauge curvature at the ith parallelogram. But by (1.1), $f^k_{\mu\nu} = 0$ on M. Therefore

$$\exp\left(\int_{\omega_i} \mathbf{b}^k X_k \cdot d\mathbf{l}\right) = I. \tag{A3}$$

Now the phase factor for the boundary of two adjacent parallelograms can be easily written as the product of the phase factors for the individual parallelogram boundaries, giving I because of (A3). Using the result in (i), one can continue this process, progressively getting bigger regions on the surface and showing that the phase factors for their boundaries are just the identity, until one finally gets

$$\exp\left(\int_{\Gamma'\Gamma^{-1}} \mathbf{b}^k X_k \cdot d\mathbf{l}\right) = I. \tag{A4}$$

Path ordering of phase factors then implies

$$\exp\left(\int_{\Gamma'} \mathbf{b}^k X_k \cdot d\mathbf{l}\right) \exp\left(\int_{\Gamma^{-1}} \mathbf{b}^k X_k \cdot d\mathbf{l}\right) = I,$$

so

$$\exp\left(\int_{\Gamma'} \mathbf{b}^k X_k \cdot d\mathbf{l}\right) = \exp\left(\int_{\Gamma} \mathbf{b}^k X_k \cdot d\mathbf{l}\right). \tag{A5}$$

[1] Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485 (1959).

[2] L. S. Schulman, *Techniques and Applications of Path Integration* (Wiley, New York, 1981).

[3] T. T. Wu and C. N. Yang, Phys. Rev. D **12**, 3845 (1975).

[4] C. N. Yang, Phys. Rev. Lett. **33**, 445 (1974).

[5] B. Felsager, *Geometry, Particles and Fields* (Odense U. P., Odense, Denmark, 1981), Chap. 2.

[6] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965), Eq. (4.59).

[7] W. S. Massey, *Algebraic Topology: An Introduction* (Springer, New York, 1967), Chap. 2, Sec. 3.

[8] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Wiley, New York, 1963), Vol. 1, pp. 71, 72, 93.

[9] A. Zeilinger, M. A. Horne, and C. G. Shull, *Proceedings of the International Symposium on Foundations of Quantum Mechanics*, Tokyo, 1983 (Physical Society of Japan, Tokyo, 1984), p. 289.

[10] L. J. Tassie and M. Peshkin, Ann. Phys. NY **16**, 177 (1961).

[11] M. V. Berry, Eur. J. Phys. **1**, 240 (1980).

[12] C. C. Bernido and A. Inomata, J. Math. Phys. **22**, 715 (1981).

[13] G. C. Gerry and V. A. Singh, Phys. Rev. D **20**, 2550 (1979).

[14] L. S. Schulman, J. Math. Phys. **12**, 304 (1971).

Reeh–Schlieder-type density results in one- and n-body Schrödinger theory and the "unique continuation problem"

Manfred Requardt[a)]
A. Sloan Laboratory of Mathematics and Physics, California Institute of Technology, Pasadena, California 91125

(Received 1 August 1985; accepted for publication 8 January 1986)

A couple of Reeh–Schlieder-type density results are proved to hold in one- and n-body Schrödinger theory, that is, it is proved that states localized at time zero in an arbitrarily small open set of $\mathbb{R}^n$ are already total after an arbitrarily small time (which implies much more than the well-known acausal behavior of nonrelativistic theories). It is shown that there exists a close connection to the so-called "unique continuation property" of elliptic partial differential operators. Furthermore, a certain machinery of analytic continuation is developed and the notion of generalized propagation kernels is introduced, which also might be of use elsewhere (e.g., in scattering theory).

I. INTRODUCTION

Analytic continuation of physically interesting quantities, e.g., n-point functions or S-matrix elements, has proved to be of great importance in quantum field theory, in particular in the so-called "Wightman theory." The proofs of many of the central results in this field rely, sometimes almost exclusively, on the technique of extending objects, originally only defined over $\mathbb{R}$ or $\mathbb{R}^{(\nu+1)\cdot n}$ (where ν is the space dimension and $\nu+1$ is the dimension of space-time) into certain domains of $\mathbb{C}$ or $\mathbb{C}^{(\nu+1)\cdot n}$, $n \in N$.

This is evidently not so in Schrödinger theory, and it is probably a widespread belief that the applicability of powerful methods like these is typically restricted to the relativistic regime, where one has the so-called "spectrum condition," "locality," etc. We have shown previously, however, that several of the results carry over to the nonrelativistic regime.[1] In this paper, we will pursue two different strategies to prove a couple of results, which we think would be difficult to prove, without using the techniques developed below, in Schrödinger theory proper. One relies on an interplay of relatively deep results of functional analysis and a simple analytic continuation argument; the other employs exclusively spectral properties of energy momentum to develop a certain machinery of analytic continuation. Even in the simplest case, free motion of one particle, the results seem not to be easily accessible without using the methods described here.

We will proceed as follows. In the Sec. II we will exhibit the close connection between what we will call a Reeh–Schlieder property of an arbitrary domain $U \subset \mathbb{R}^\nu$ or $\mathbb{R}^{\nu\cdot n}$ (for the origin of the notion in Wightman theory see Ref. 2) and two seemingly different groups of concepts and ideas from the realm of classical functional analysis, one running under the catchword "unique continuation property," the other comprising the various notions of "generalized eigenfunction expansions" of Schrödinger operators. By "Reeh–Schlieder property" we mean roughly that the wave functions being localized in an arbitrarily small domain $U \subset \mathbb{R}^{\nu\cdot n}$ at time 0 are already total in the full Hilbert space $L^2(\mathbb{R}^{\nu\cdot n})$ after an arbitrarily short time interval. (Note that this implies much more than the feature, well-known at least for the free time evolution, that a wave function is more or less "everywhere" after an arbitrarily short time. However, without this property the stronger result could, of course, not hold.) These notions will be made more precise in Sec. II.

While in the Sec. II analyticity plays only a minor role, we have to rely heavily on it in the rest of the paper where we develop another sequence of ideas, pursuing more or less the goal of analytically extending both space and time translations into certain domains of $\mathbb{C}^{(\nu+1)\cdot n}$. That is, the first part carries a more functional analytic flavor while the latter draws more on procedures known to be successful in relativistic quantum field theory. Furthermore, we think the concepts we develop in the latter part, such as, e.g., "generalized propagation kernels," also will be of use elsewhere (for example, in a paper on a new general approach to scattering theory in energy-momentum space, which is in preparation).

To indicate some of the technical steps, we will start by rewriting n-body Schrödinger theory in the form of a certain time-dependent bilinear functional $W(X,Y;t)$ lying in $\mathscr{S}'(\mathbb{R}^{2\nu\cdot n})$, and acting on the wave functions at time 0 by using the nuclear theorem. These functionals contain the full physical information of the theory. We show that these "Wightman functions" of Schrödinger theory can be naturally viewed as restrictions of more complex functions lying in a bigger space, i.e.,

$$W(x_1,\dots,x_n, y_1,\dots, y_n;t) \to W(x_1t_1,\dots;y_1t_1',\dots)\,. \tag{1.1}$$

The distributional Fourier transform of this extended W has nice support properties in the energy-momentum variables $\{(\omega_i,k_i)\}$ corresponding to $\{(x_i,t_i)\}$, which allows us to make an analytic extension of the original $W(X,Y;t)$ $[X \simeq (x_1,\dots,x_n)$, etc.$]$ into a certain domain of $\mathbb{C}^{(\nu+1)\cdot n}$. The "values" $W(X,Y;t)$ then turn out to be the boundary values of an analytic function over $\mathbb{R}^{(\nu+1)\cdot n} + i\Gamma^n \subset \mathbb{C}^{(\nu+1)\cdot n}$ when we approach the real boundary $\mathbb{R}^{\nu\cdot n+1}$ (resp. $\mathbb{R}^{(\nu+1)\cdot n}$).

[a)] Permanent address: Institute for Theoretical Physics, Universität Göttingen, West Germany.

II. THE CONNECTION BETWEEN THE REEH-SCHLIEDER (RS) AND THE UNIQUE CONTINUATION PROPERTY

To begin with, we have to define exactly what the RS property is to mean. So let $\mathscr{H} = L^2(\mathbb{R}^m)$, $m = \nu \cdot n$, $H = -\Delta + V$, U be an open set $\subset \mathbb{R}^m$, and I be an arbitrary time interval about $t = 0$, then

$$S(U, I) := \{\exp(-itH)\Phi ; \operatorname{supp} \Phi \subset U, t \in I\} . \quad (2.1)$$

Definition 1: U has the RS property if $S(U, I)$ is total in $\mathscr{H}$ for some I.

We want to show that, in fact, for a large class of potentials and practically every arbitrarily small U and I, $S(U, I)$ has this property. To prove this we draw on two relatively advanced topics of elliptic partial differential operators, described by the catchwords (i) unique continuation property and (ii) generalized eigenfunction expansion. Both are topics of currently active research, and to prove them we need relatively advanced machinery. We do not want to go into detail in this paper concerning these two problems, but prefer to restrict ourselves to giving some definitions and references that show that both features are actually fulfilled for a sufficiently large class of potentials.

We need the unique continuation property in the following form.

Definition 2: Let Φ lie in the local Sobolev space $H^2_{\mathrm{loc}}(R^m)$, i.e., $u \cdot \Phi \in D(-\Delta)$ for all $u \in C_0^\infty(\mathbb{R}^m)$. Furthermore, let Φ satisfy the following differential inequality for some $\lambda \in \mathbb{R}$ and V the potential:

$$|\Delta\Phi(x)| \leqslant |(\lambda - V(x)) \cdot \Phi(x)| . \quad (2.2)$$

We say a *unique continuation* property holds if (2.2) implies the following: If we assume that Φ vanishes around a certain point x, then it vanishes everywhere (in the sense of L^2_{loc}).

Remark: The phenomenon mentioned above has a long history (see, e.g., the notes in Ref. 3 to the appendix to Chap. XIII.13 or Ref. 4). In recent years, the conditions imposed on the potential V have been more and more relaxed (see, e.g., Refs. 5–8 and further references given there).

The next property we shall need is the existence of an eigenfunction expansion of the Hamiltonian with the generalized eigenfunctions being "sufficiently nice." Also, here we do not aim at optimal results, but content ourselves with showing that something like this actually does exist for a sufficiently large class of potentials. In order not to struggle with perhaps nasty measure theoretic problems, we restrict ourselves to the class of so-called Agmon potentials (a treatment of long-range potentials can, e.g., be found in the book of Saito[9]). An approach more in the original spirit of Ikebe and Povzner can also be found in Ref. 10, see also Ref. 11, Sec. C 5. So, with V being an Agmon potential (cf., e.g., Ref. 12 for the necessary details) we have the following theorem.

Theorem 1: With V of short-range type in the sense of Agmon, we have a complete set of generalized continuum eigenfunctions $\phi(\cdot,k)$, labeled by $k \in \mathbb{R}^m$, lying in H^2_{loc} such that the following holds: With Φ, $\Psi \in L^2(\mathbb{R}^m)$, and g a bounded continuous function,

$$(\Psi | g(H)\Phi) = \sum_{n=1}^{N} g(\lambda_n)(\Psi | \phi_n)(\phi_n | \Phi) + \int d^m k\, g(k^2) \langle \Psi | \phi(\cdot,k)\rangle \times \langle \phi(\cdot,k) | \Phi\rangle , \quad (2.3)$$

where N may be infinite, the ϕ_n are the usual L^2-eigenfunctions with eigenvalue λ_n, and $\langle \phi(\cdot,k)|\Phi\rangle$ is an abbreviation for $\int \overline{\phi(x,k)} \cdot \Phi(x) \cdot d^m x$, defined in an appropriate sense, where the k^2 are the continuum eigenvalues.

Proof: For the proof see Ref. 12. Note in particular that there is no singular continuous spectrum in this case.

Now we are ready to prove the main result of this section. We assume throughout the paper that H is bounded below (which is of course fulfilled for, e.g., Agmon potentials). With U an arbitrary open set in $\mathbb{R}^m$ and I a fixed but arbitrarily small time interval about $t = 0$, we would like to show that $S(U, I)$ is total in $\mathscr{H} = L^2(\mathbb{R}^m)$. We assume the contrary, i.e., there exists a nonzero $\Psi \in L^2(\mathbb{R}^m)$ such that

$$(\Psi, e^{-itH}\Phi) \equiv 0, \quad \text{for all } \Phi \in S(U), \ t \in I \quad (2.4)$$

[where $S(U) \cong S(U, t = 0)$, i.e., the functions exactly localized in U]. Evidently,

$$(\Psi, e^{-itH}\Phi) = \int e^{-i\lambda t} d(\Psi, E_\lambda \Phi)$$

has an analytic continuation into the lower half plane, i.e.,

$$F(t - i\tau) := \int e^{-i\lambda(t - i\tau)} d(\Psi, E_\lambda \Phi) \quad (2.5)$$

is analytic for $\tau > 0$. Defining

$$G(t + i\tau) := \overline{F(t + i\tau)} , \quad (2.6)$$

G is analytic in the upper half plane. We have

$$G(I) \equiv F(I) \equiv 0 \quad (2.7)$$

by assumption, that is, on an open set of the common real boundary of G, F. Hence we see that G is the analytic continuation of F through the real open set I, where the analytic function $F \cup G$ is zero. This implies $F \equiv 0$ in the lower half plane, and by continuity F vanishes also on the real boundary, i.e., we have

$$(\Psi, e^{itH}\Phi) \equiv 0, \quad \text{for all } t . \quad (2.8)$$

(This simple reasoning, in fact quite common in Wightman theory, was also exploited in Ref. 13 in order to study the localization properties in quantum theory.)

By uniqueness of the Fourier transform the measure $d(\Psi, E_\lambda \Phi)$ is zero. Assuming now that H has a generalized eigenfunction expansion according to Theorem 1 we can conclude

$$\int_{k^2 = \lambda} dS_{m-1} \langle \Psi, \phi(\cdot,k)\rangle \cdot \langle \phi(\cdot,k), \Phi\rangle + \sum_{\lambda_n = \lambda} (\Psi, \phi_n) \cdot (\phi_n, \Phi) = 0 , \quad (2.9)$$

for a.e. $\lambda \in \mathbb{R}$ (with respect to the measure $[\theta(\lambda) \cdot d\lambda + \Sigma_n \delta(\lambda - \lambda_n) d\lambda\,]$!).

Remark: The above restricted integration over the sphere $k^2 = \lambda$ is well defined, cf. Ref. 12, Theorem 5.1.

Abbreviating now $\langle \Psi, \phi(\cdot,k)\rangle$ [resp. (Ψ, ϕ_n)] by $c(k)$ (resp. c_n) we can write this as

$$\left\langle \int_{k^2=\lambda} dS_{m-1}\, c(k)\, \phi(\cdot,k) + \sum_{\lambda_n=\lambda} c_n\, \phi_n, \Phi \right\rangle = 0, \quad \text{a.e.,} \tag{2.10}$$

$$u_\lambda := \left(\int_{k^2=\lambda} dS_{m-1}\, c(k)\, \phi(\cdot,k) + \sum_{\lambda_n=\lambda} c_n\, \phi_n \right) \tag{2.11}$$

is again a solution of $(H-\lambda)u=0$, $u \in H^2_{\text{loc}}$.

It is our aim to show that (2.9) implies, in fact, that $c_n = 0$, $c(k)=0$ for all n (resp. almost all k). This then would yield that $\Psi = 0$ in L^2, i.e., that $S(U,I)$ is total. To show this we need the unique continuation property of Definition 2; u_λ of (2.11) fulfills the hypotheses of Definition 2, furthermore,

$$\langle u_\lambda, \Phi\rangle = 0,$$
$$\text{for all} \quad \Phi \in S(u) \Rightarrow u_\lambda \equiv 0 \quad \text{on} \quad U. \tag{2.12}$$

By the unique continuation property this entails that $u_\lambda \equiv 0$ a.e. (in L^2_{loc}). But the expansion with respect to $\{\phi(\cdot,k), \phi_n\}$ is "orthogonal," that is, $u_\lambda \equiv 0$ for almost all λ implies that $\{c(k), c_n\} \equiv 0$ a.e., hence $\Psi = 0$ in L^2. This proves the first part of the following theorem.

Theorem 2: Assuming that H has the unique continuation property and a generalized eigenfunction expansion in the sense of Theorem 1, the following statements hold.

(i) Given an arbitrary open set $U \subset \mathbb{R}^m$ and an arbitrarily small time interval I around $t=0$, then $S(U,I)$ is already total in the full $L^2(\mathbb{R}^m)$.

(ii) Conversely, if a RS property holds for every open $U \subset \mathbb{R}^m$, then the generalized eigenfunctions have the unique continuation property in the sense of Definition 2.

Proof: The proof of (ii) is easy. Assuming that u is a nontrivial generalized eigenfunction vanishing, e.g., on a certain open $U \subset \mathbb{R}^m$, we have for all $\Phi \in S(U)$ and all $t \in \mathbb{R}$

$$0 = e^{-i\lambda t} \langle u_\lambda, \Phi\rangle = \langle e^{iHt} u_\lambda, \Phi\rangle = \langle u_\lambda, e^{-itH}\Phi\rangle . \tag{2.13}$$

But, by assumption, $\{e^{itH}\phi\}$ is total in L^2, in particular in every L^2_Ω, Ω compact in $\mathbb{R}^m$. Hence u_λ (which lies in L^2_{loc}) vanishes in L^2_{loc}.

We would like to mention that the above result is much stronger than that implied by the well-known feature of nonrelativistic quantum theory, namely, that wave functions have the tendency to spread out to infinity almost instantaneously. The latter says only that the wave function cannot be orthogonal to certain functions that have their support concentrated in possibly very small neighborhoods of points $x \in \mathbb{R}^m$. Theorem 2 says that even arbitrary extended and oscillating functions cannot be orthogonal to $S(U,I)$. The physics behind the result is perhaps even more striking. Theorem 2 tells us that the physical content of the theory is already contained in an arbitrarily small space-time neighborhood of an arbitrary point.

III. ANOTHER APPROACH AND THE NOTION OF GENERALIZED PROPAGATION KERNELS

In the rest of the paper we will develop a different approach to the problem with slightly different results and completely different methods, the whole approach being more in the spirit of the original relativistic context. Stated somewhat sloppily, it consists of extending time and space translations together into certain domains of $\mathbb{C}^{(\nu+1)\cdot n}$. This analytical continuation is, however, a little bit subtle since momentum is not bounded below. It would be tempting to use a scale of auxiliary spaces (which are possibly no longer Hilbert spaces) to give meaning to expressions like $\{\exp i(a+ib)P\}\,\phi$, $\phi \in L^2(\mathbb{R}^m)$, where P is the momentum operator. We postpone this approach to the future, however, and choose another strategy in this paper.

In a first step we want to present n-body Schrödinger theory in a way slightly different from the conventional approach, i.e., by means of so-called propagation kernels. It will turn out that these distributions contain all the physics of the theory and are amenable to a certain embedding of ordinary Schrödinger theory into a larger theory, in which certain spectral and analyticity properties of the objects of interest can be visualized more easily.

We start from the expression

$$(\phi, e^{-itH}\psi), \quad \phi, \psi \in L^2(\mathbb{R}^{\nu\cdot n}) . \tag{3.1}$$

Evidently this defines a sesquilinear functional over $L^2(\mathbb{R}^{\nu\cdot n}) \times L^2(\mathbb{R}^{\nu\cdot n})$ which depends on t. We will restrict this functional to the dense subset $\mathscr{S}(\mathbb{R}^{\nu\cdot n})$. In contrast to, e.g., L^2, $\mathscr{S}$ is a *nuclear space*, which has a far-reaching consequence. In this case the kernel or *nuclear theorem* holds, which allows us to prove the following theorem.

Theorem 3: With ϕ, $\psi \in \mathscr{S}(\mathbb{R}^{\nu\cdot n})$ there exists a time-dependent tempered distribution $W \in \mathscr{S}'(\mathbb{R}^{\nu n} \times \mathbb{R}^{\nu n})$ such that

$$\text{(i)}\ (\phi, e^{itH}\psi) = \int \bar{\phi}(X_n)\cdot \psi(Y_n) \cdot W(X_n, Y_n; t)\, dX_n\, dY_n ,$$

$$\text{(ii)} \int W(X_n, Y_n; t)\cdot \psi(Y_n)\, dY_n = \psi_t = e^{-itH}\psi \quad \text{in } L^2\text{-sense, and}$$

(iii) $W(X_n, Y_n; 0) = \delta(X_n - Y_n)$ and $W(X_n, Y_n; t)$ "solves" the Schrödinger equation with respect to $X_n = (x_1,\ldots,x_n)$.

Proof: Continuity in $\mathscr{S}$ implies continuity in L^2 such that (ϕ, ψ_t) is separately continuous in $\mathscr{S}\times\mathscr{S}$ for every fixed t. The simple proof goes as follows:

$$f_n \to f \text{ in } \mathscr{S} \Rightarrow \sup_x\, (1+|X|)^k \cdot |f - f_n|^2 \to 0 , \tag{3.2}$$

with $|X| := \Sigma^n_{i=1} |x_i|$ and every $k \in \mathbb{N}$. Thus we have

$$\int |f - f_n|^2\, dX = \int |f - f_n|^2 \cdot (1+|X|)^k (1+|X|)^{-k}\, dX \leqslant \sup_x \{(1+|X|)^k |f - f_n|^2\} \cdot \int (1+|X|)^{-k}\, dX \to 0 , \tag{3.3}$$

for k large enough and $n \to \infty$. So the kernel theorem can be applied. By varying ϕ over a dense set in $L^2 \cap \mathscr{S}$, we see that (ii) must hold, and (iii) is evident.

Remarks: (i) That the kernel theorem yields nontrivial information can be seen by the following observation. The scalar product (ϕ,ψ) is also a sesquilinear, even jointly continuous, functional over $L^2 \times L^2$. If a kernel theorem would hold in this situation, the W under discussion would be an element from $L^2(\mathbb{R}^{\nu n} \times \mathbb{R}^{\nu n})$! But we already know what W looks like, namely for $\phi, \psi \in \mathscr{S}$, $W = \delta(X_n - Y_n)$, which is not in L^2 but in $\mathscr{S}'$.

(ii) The above distribution W has special additional continuity properties such that we can hope to be able to restrict the general structure of W further. The structure of distributions lying in $\mathscr{S}'$ is well known (see, e.g., Refs. 14 and 15):

$$W(X_n,Y_n) = D^{\alpha}(X_n,Y_n) m(X_n,Y_n) , \qquad (3.4)$$

with D a certain differential operator of degree α, $m(X_n,Y_n)$ a measure or a function on $\mathbb{R}^{\nu n} \times \mathbb{R}^{\nu n}$. In the simplest case, e.g., H_0, the free Hamiltonian on $L^2(\mathbb{R}^\nu)$, the free propagator has the well-known form

$$P_0(x,y;t) = (4\pi i t)^{-\nu/2} \exp(i|x-y|^2/4t) . \qquad (3.5)$$

Other kernel representations are known for, e.g., $(H-E)^{-1}$ in the context of eigenfunction expansions and the Lippmann–Schwinger equation. In the restricted case of potential scattering we have, for example,

$$((H-E)^{-1} \cdot \psi)(x) = \int G(x,y;E)\, \psi(y) dy , \qquad (3.6)$$

$G \in L^1 \cap L^2$ for a.e. fixed x with respect to y and certain classes of potentials (see, e.g., Ref. 10, Chap. XI.6). We must, however, emphasize that the situation in (3.6) is considerably simpler since one makes heavy use of the fact that certain operators are Hilbert–Schmidt (which is typical for the case of resolvents).

(iii) Note that the distribution W of Theorem 3 occurs also as a path integral in the Feynman–Kac theory.

The observation above motivates the name propagation kernel or Wightman function of n-particle Schrödinger theory. The W defined in Theorem 3 is of the form

$$W(x_1 t,\ldots,x_n t; y_1,\ldots,y_n) , \qquad (3.7)$$

i.e., all time coordinates are equal. In the next step we want to make a natural extension to a more general distribution, depending on $t_1,\ldots,t_n$. To this end we will assume (whereas this is not strictly necessary) that the potential occurring in the n-particle Hamiltonian is a sum of pair potentials, i.e.,

$$\begin{aligned} H &= H_0 + V \\ &= \frac{1}{2} \sum_{i=1}^{n} -\Delta_i + \sum_{i<j} V_{ij}(x_i - x_j) \end{aligned} \qquad (3.8)$$

(for simplicity all masses are normalized to 1). This makes V translation invariant, more precisely, invariant under overall translations. Thinking now of the coordinates as ordered n-tuples $(x_n,\ldots,x_1)$ and correspondingly $L^2(\mathbb{R}^{\nu \cdot n}) = L^2(\mathbb{R}^\nu_n) \otimes \cdots \otimes L^2(\mathbb{R}^\nu_i)$, we can define individual time evolutions in each subspace

$$\mathscr{H}_1 := L^2(\mathbb{R}^\nu_1), \quad \mathscr{H}_2 := L^2(\mathbb{R}^\nu_2) \otimes L^2(\mathbb{R}^\nu_1), \ldots$$

and

$$\begin{aligned} H_1 &:= -\tfrac{1}{2}\Delta_1 , \\ H_2 &:= -\frac{1}{2} \sum_{i=1}^{2} \Delta_i + V_{12}(x_1 - x_2) , \\ &\vdots \\ H_n &:= -\frac{1}{2} \sum_{i=1}^{n} \Delta_i + \sum_{i<j} V_{ij}(x_i - x_j) . \end{aligned} \qquad (3.9)$$

For notational simplicity we assume the overall wave function at time zero, $\phi(x_n,\ldots,x_1)$, to be given as a product (the general case is analogous):

$$\phi(x_n,\ldots,x_1) := \phi_n(x_n) \cdot \cdots \cdot \phi_1(x_1) . \qquad (3.10)$$

We can then extend $\phi(x_n,\ldots,x_1;t)$ in the following way: We let $\exp -i(t_1 - t_2) \cdot H_1$ operate on $\phi(x_1)$, $\exp -i(t_2 - t_3) \cdot H_2$ operate on the product $\{\phi_2(x_2) \cdot \phi_1(x_1;t_1 - t_2)\}$, etc., that is we get the following definition.

Definition 3: Let $\phi(x_n,\ldots,x_1) = \phi_n(x_n) \cdot \cdots \cdot \phi_1(x_1)$ with $\phi_i \in L^2(\mathbb{R}^\nu)$. We define $\phi(x_n t_n,\ldots,x_1 t_1)$ by

$$\begin{aligned} &\phi(x_n t_n,\ldots,x_1 t_1) \\ &\quad := e^{-it_n H_n}[\phi_n(x_n) \cdot \exp(-i(t_{n-1} - t_n)H_{n-1})] \\ &\qquad \times [\phi_{n-1}(x_{n-1}) \exp(-i(t_{n-2} - t_{n-1})H_{n-2})] \\ &\qquad \cdots [\cdots [\phi_2(x_2) \cdot e^{-i(t_1 - t_2)H_1} \phi_1(x_1)] \cdots] . \end{aligned} \qquad (3.11)$$

We have the simple following corollary, which shows that this actually defines an embedding.

Corollary 1: $\phi(x_n,\ldots,x_1;t)$ is recovered by setting

$$t_n = t_{n-1} = \cdots = t_1 = t .$$

This procedure can be extended in an evident manner to every function of $L^2(\mathbb{R}^{\nu \cdot n})$ since by the Fubini–Tonelli theorem, every function of $L^2(\mathbb{R}^{\nu \cdot n})$ is an L^2-function in, e.g., $\{x_k,\ldots,x_1\}$ for almost all coordinates $\{x_n,\ldots,x_{k+1}\}$ being held fixed.

In a completely analogous way we can extend space translations.

Definition 4: With ϕ of Definition 3, we define

$$\begin{aligned} &\phi(x_n + a_n; t_n, x_{n-1} + a_{n-1}; t_{n-1},\ldots,x_1 + a_1; t_1) \\ &\quad := e^{-it_n H_n} \cdot e^{ia_n P_n} \\ &\qquad \cdot [\phi_n(x^n) \cdot \exp(-i(t_{n-1} - t_n)H_{n-1}) \\ &\qquad \cdot \exp(i(a_{n-1} - a_n)P_{n-1}) [\phi_{n-1}(x_{n-1}) \cdots] \cdots] , \end{aligned} \qquad (3.12)$$

with

$$P_k := \sum_{j<k} p_j, \quad p_j = -i \cdot \partial_{x_j}, \quad a_j \in \mathbb{R}^\nu .$$

Remark: Again we see that we get an overall translation by setting $a_n = a_{n-1} = \cdots = a_1 = a$. Note furthermore that H_k and P_k commute on $\mathscr{H}_k$ and that, in fact, the above-defined extended space translations shift the individual coordinates $\{x_k\}$ by vectors $\{a_k\}$.

Employing the above definitions we can make corresponding extensions of the propagation kernels.

Proposition 1: By

$$\int W(x_n + a_n;t_n,\ldots,x_1 + a_1,t_1 | y_n,\ldots, y_1)$$

$$\times \phi(y_n,\ldots,y_1)dY_n$$

$$:= \int W(x_n,\ldots,x_1 | y_n,\ldots,y_1)$$

$$\times \phi(y_n + a_n,t_n;\ldots,y_1 + a_1,t_1)dY_n\ , \qquad (3.13)$$

a *generalized propagation kernel* is defined, which is again a distribution in $\mathscr{S}'$ depending on the parameters $\{t_n,a_n;\ldots;t_1,a_1\}$.

Proof: The continuity properties with respect to X_n and Y_n in the $\mathscr{S}$-topology can be shown as in (3.2).

The next natural step is to make a Fourier transformation with respect to the variables $\{t_na_n,\ldots,t_1a_1\}$ and employ certain support properties of the Fourier transform $\hat{W}$ in the variables $\{\omega_nk_n,\ldots,\omega_1k_1\}$. While we do not intend to talk about scattering theory in this paper we would nevertheless like to make a short aside about the possible use of the above extension method in this field. The above approach offers the possibility of dealing with scattering phenomena on a very broad scale, that is, in quantum field theory, Schrödinger theory, temperature states, ground states, etc. The particular advantage is that one need not worry (at least openly) about reference dynamics, range of interactions, spectral gaps around the mass shells, etc. This is replaced by a study of the spectral properties of the $\hat{W}$'s along certain submanifolds in the support of $\hat{W}$ on which, in the limit $t \to \pm\infty$, the "scattering states" will live. For temperature states this has been done in a recent paper.[16] By means of certain geometric arguments we can show when and why scattering states exist and exactly which properties, in the neighborhood of the mass shells of ingoing and outgoing clusters, particles in the spectral support of $\hat{W}$ yield a nontrivial S-matrix. The whole approach is more in the spirit of Buchholz's treatment of scattering of massless particles and will be developed in a forthcoming paper.

IV. SUPPORT PROPERTIES OF THE FOURIER TRANSFORMS OF $\phi(t_na_n,\ldots,t_1a_1)$ AND $W(t_na_n,\ldots,t_1a_1)$

We will now investigate the support properties of the Fourier transforms of $\phi(t_na_n,\ldots,t_1a_1)$ [resp. $W(t_na_n,\ldots,t_1a_1)$] (where we dropped the remaining coordinates X_n,Y_n) with respect to $(t_n,a_n,\ldots,t_1a_1)$. This is the same as investigating the support of the joint energy-momentum spectrum of (H_m,P_m) in each subspace $\mathscr{H}_m$ [cf. (3.9)].

Proposition 2: Let all V_{ij} be infinitesimally H_0-bounded, either in operator sense or form sense, or, slightly weaker, each $V^{(m)}$ in $\mathscr{H}_m$ relatively bounded with all relative bounds $\{a_m\}$ smaller than 1. Then the joint (H_m,P_m)-spectrum, a set in $\mathbb{R}^4$, can be bounded from below by the hypersurface $(a_m < 1)$

$$\omega = c_m\cdot(1 - a_m)\cdot k^2 - b_m\ , \qquad (4.1)$$

with c_m, b_m certain constants and (ω,k) the energy-momentum variables corresponding to (H_m,P_m) in the subspace $\mathscr{H}_m$ of the particles $(1),\ldots,(m)$.

Proof: An analogous proof for the more general case of general nonrelativistic quantum field theory can be found in Ref. 1, so we give only a sketch of the proof. In the first step we show that the joint spectrum of $H_0^{(m)} := -\frac{1}{2}\Sigma_{i=1}^m \Delta_i$ and $P_m := \Sigma_{i=1}^m p_i$ can be bounded below by a parabolic hypersurface (bounding, e.g., Σp_i^2 by $c\cdot|\Sigma p_i|$). In the second step we exploit the relative boundness to show that the interaction results only in a *finite* overall shift of this paraboloid.

Corollary 2: (i) The joint spectrum of (H_m,P_m) can be embedded in a domain $K_m \cup \Gamma_c \subset \mathbb{R}^4$, where K_m is a ball around $(0,0) \in \mathbb{R}^4$ with sufficiently large radius and Γ_c the so-called forward cone $\{(\omega,k);\ \omega > c\cdot|k|,\ c > 0\}$.

(ii) By choosing the radius of K_m large enough, c also can be chosen arbitrarily large.

Proof: Each paraboloid of the form (4.1) intersects every cone Γ_c for sufficiently large $|k|$. So there exists always a finite ω_0, depending on a_m, b_m, and c such that all (ω,k) lying above the surface (4.1) are ultimately contained in Γ_c for $\omega \geqslant \omega_0$.

Now we take a $\phi(x_n + a_n;t_n,\ldots,x_1 + a_1;t_1)$ defined in Definition 4 and observe that each $\exp(-i(t_{k-1} - t_k)H_{k-1})\cdot\exp(i(a_{k-1} - a_k)\cdot P_{k-1})$ standing between the functions ϕ_k and $[\phi_{k-1}\cdots]$ acts on a Hilbert space of particle number $k-1$. We know from the above discussion that the joint (H_{k-1}, P_{k-1})-spectrum is contained in some $K_{k-1} \cup \Gamma_c \subset \mathbb{R}^4$. Inserting the spectral resolution for each of the above operators, we get

$$\phi(a_n;t_n,\ldots,a_1;t_1) = (2\pi)^{-2n} \int \exp[-i(t_n\omega - a_nk_n)]\exp[-i(t_{n-1} - t_n)\omega_{n-1} - (a_{n-1} - a_n)k_{n-1}]$$

$$\cdot\cdots\cdot\{E^{(n)}(d\omega_n\, dk_n)\cdot \phi_n E^{(n-1)}(d\omega_{n-1}\, dk_{n-1})\phi_{n-1}\cdot\cdots\cdot E^{(1)}(d\omega_1\, dk_1)\, \phi_1\} \qquad (4.2)$$

$$= (2\pi)^{-2n} \int \exp[-i(t_n\omega_n - a_n\, k_n)]\cdot\cdots\cdot\hat{W}(\omega_nk_n,\ldots,\omega_1k_1|X_n,Y_n)\phi(y_n\cdots y_1)dY_n\ . \qquad (4.3)$$

It should be noted that, whereas the various $E\ (d\omega\ dk)$ occurring in (4.2) are spectral measures, the curly bracket is a measure only in each of the coordinates (ω_ik_i) with the remaining variables (ω_jk_j) being integrated over (with appropriate test functions). Taken as a whole, the curly bracket in (4.2) is a vector valued distribution in the variables $(\omega_nk_n,\ldots,\omega_1k_1)$ with $\exp[-i(t_n\omega_n - a_n\, k_n)\cdot\cdots]$ lying in the domain of definition. By the same token $\hat{W}$ is a distribution with respect to $\{(\omega_ik_i)\}$. The support properties of the joint spectrum of $\{(H_k\, P_k)\}$ now entail that the curly bracket on the right-hand side (rhs) of (4.2) (resp. $\hat{W}$) are distributions with support:

$$\operatorname{supp} \hat{W} \quad [\text{resp. } \operatorname{supp} \hat{\phi} \subset \{(\omega_n k_n,...,\omega_1 k_1) \text{ such that each } (\omega_i\ k_i) \in K_i \cup \Gamma_c\}] , \qquad (4.4)$$

that is, we have proved the following theorem.

Theorem 4: $\phi(t_n a_n,...,t_1 a_1)$ is the Fourier transform [with respect to $(t_n a_n,...,t_1 a_1)$] of a vector valued distribution $\hat{\phi}$, given as the curly bracket in (4.2), with (ω_i,k_i)-support contained in $K_i \cup \Gamma_c$. The same support properties hold for $\hat{W}$ the Fourier transform of the generalized propagation kernel.

V. THE ANALYTIC CONTINUATION OF ϕ $(t_n a_n,...,t_1 a_1)$ AND $W(t_n a_n,...,t_1 a_1)$

We will now exploit the special support properties of ϕ, $\hat{W}$ to show that ϕ, W can be analytically continued into a domain of $\mathbb{C}^{(\nu+1)\cdot n}$. We have seen that the support of $\hat{\phi}$, $\hat{W}$ is restricted by $(\omega_i\ k_i) \in K_i \cup \Gamma_c$. We observe now that the rhs of, e.g., (4.2) still exists if the exponents

$$(t_n \omega_n - a_n k_n) ,$$
$$((t_{n-1} - t_n)\cdot\omega_{n-1} - (a_{n-1} - a_n)k_{n-1}),...$$

are replaced by

$$(z_n \cdot \omega_n - \zeta_n \cdot k_n) ,$$
$$((z_{n-1} - z_n)\cdot\omega_{n-1} - (\zeta_{n-1} - \zeta_n)\cdot k_{n-1}),..., \qquad (5.1)$$

with

$$z_n := t_n - i\tau_n, \quad \zeta_n := a_n - ib_n,..., \qquad (5.2)$$

where (τ_n,b_n),..., fulfill the support condition

$$(\tau_n \omega_n - b_n\ k_n) > 0 ,$$
$$((\tau_{n-1} - \tau_n)\omega_{n-1} - (b_{n-1} - b_n)k_{n-1}) > 0,..., \qquad (5.3)$$

for $(\omega_n,k_n) \in \Gamma_c$, $(\omega_{n-1},k_{n-1}) \in \Gamma_c$,... .

This can be seen as follows. We can split the support of $\hat{W}$ with respect to each (ω_i,k_i) into the sets $K_i \cup \Gamma_c \setminus \Gamma_c$ and Γ_c. Since K_i has a finite diameter the analytic continuation of

$$\int_{K_i u \Gamma_c \setminus \Gamma_c} (\cdots) d\omega_i\ dk_i \quad \text{with respect to } (z_i, \zeta_i) \qquad (5.4)$$

always exists. To continue the integral over Γ_c, $\int_{\Gamma_c} (\cdots)\, d\omega_i \times dk_i$, we need the special support properties of the (τ_i, ζ_i) mentioned in (5.3). These properties guarantee that

$$\exp[-(\tau_n \omega_n - b_n\ k_n)],$$
$$\exp[-((\tau_{n-1} - \tau_n)\omega_{n-1} - (b_{n-1} - b_n)k_{n-1})],... \qquad (5.5)$$

are globally bounded on the domain, Γ_c, of integration.

This observation allows us to prove the following theorem.

Theorem 5: $\phi(t_n a_n,...,t_1 a_1)$, $W(t_n a_n,...,t_1 a_1)$, defined in Sec. III can be analytically continued into the domain $T^n \subset \mathbb{C}^{(\nu+1)\cdot n}$, given by

$$(z_n, \zeta_n), \quad ((z_{n-1} - z_n),(\zeta_{n-1} - \zeta_n)),...$$
$$\in \mathbb{R}^{\nu+1} + i\{(\tau,b);\ b \in \mathbb{R}^\nu,\ \tau > 0\} := T , \qquad (5.6)$$

with $z_n = t_n - i\tau_n$, $\zeta_n = a_n - ib_n$,... .

Proof: The set of pairs (τ_n,b_n),... given by (5.3) span the interior of the so-called dual cone $\tilde{\Gamma}_c$ of Γ_c. We showed in Sec. IV that the (ω_i,k_i)-support of $\hat{\phi}$, $\hat{W}$ is bounded below by a paraboloid that intersects eventually every cone Γ_c for arbitrarily large c. This implies that ϕ, W can be analytically continued into $\mathbb{R}^{(\nu+1)} + i\,\tilde{\Gamma}_c$ in each of the variables given in (5.2) for $c \to \infty$. For $c \to \infty$, $\tilde{\Gamma}_c$ becomes the whole half space $(b \in \mathbb{R}^\nu, \tau > 0)$.

It is perhaps instructive to apply the above machinery to the simplest example we can think of, the free time evolution in $\mathbb{R}^3$. In this case we have $(H_0 := -\Delta)$

$$(e^{-itH_0}\phi)(a) = (2\pi)^{-3/2} \cdot \int e^{-itp^2} \cdot e^{ipa} \hat{\phi}(p) d^3p . \qquad (5.7)$$

The energy spectrum is supported on the hypersurface $\omega = p^2$, that is,

$$\hat{W}(\omega,k) = \delta(\omega - p^2) , \qquad (5.8)$$

and we see that

$$(2\pi)^{-3/2} \int e^{-i(t-i\tau)p^2} \cdot e^{i(a-ib)p} \hat{\phi}(p) d^3p \qquad (5.9)$$

exists provided that $\tau > 0$ and is analytic in the domain

$$\{(z, \zeta);\ z = t - i\tau, \quad \zeta = a - ib, \quad \tau > 0\} \qquad (5.10)$$

(since p^2 wins out against $|p|$ for $|p| \to \infty$). The generalized propagation kernel is

$$W(a - ib, y; t - i\tau) := (4\pi i(t - i\tau))^{-3/2} \cdot \exp(i\cdot(a - y - ib)^2/4(t - i\tau)) , \qquad (5.11)$$

and we see again that

$$\int W(a - ib, y, t - i\tau)\phi(y) dy$$

is complex differentiable with respect to $\{(a - ib), (t - i\tau)\}$ in L^2 as long as $\tau > 0$ since this provides us with a term $\sim \exp(-|a - y^2|/4\tau)$.

VI. ANOTHER VERSION OF THE REEH–SCHLIEDER THEOREM

We now prove another version of the RS theorem, which is more in the spirit of the original version proved in Wightman theory.

Theorem 6: With a Hamiltonian H and generalized states $\phi(t_n\ a_n,...,t_1 a_1)$ as given in Sec. III, the following holds: The set

$$S'(U_n, I) := \{\phi(x_n t_n,...,x_1 t_1) , \quad t_i \in I, \quad \operatorname{supp} \phi(x_n,...,x_1) \subset U_n \subset \mathbb{R}^{\nu\cdot n}\}$$

is already total in $\mathscr{H}_n = L^2(\mathbb{R}^{\nu\cdot n})$, where U_n is an arbitrarily small open set in $\mathbb{R}^{\nu\cdot n}$ and I is an arbitrarily small time interval around $t = 0$.

Remark: It is already sufficient to choose the wave functions of the form

$$\phi(x_n,...,x_1) = \phi_n(x_n) \cdot \cdots \cdot \phi_1(x_1) .$$

Proof: We assume the contrary, i.e., there exists a wave function $\psi(x_n,...,x_1)$ such that

$$(\psi,\phi) = 0, \quad \text{for all } \phi \in S'(U_n, I) . \qquad (6.1)$$

We proved in Sec. V that the function

$$F(t_n a_n,\ldots,t_1 a_1) := (\psi, \phi(t_n a_n,\ldots,t_1 a_1)) \tag{6.2}$$

can be analytically continued, with respect to $\{(t_i a_i)\}$, into an open domain T^n of $\mathbb{C}^{(\nu+1)\cdot n}$ and that $F(t_n a_n,\ldots,t_1 a_1)$ [resp. $F(a_n,\ldots,a_1)$] are the boundary values for $\{(\operatorname{Im} z_i, \operatorname{Im} \zeta_i)\} \to 0$.

Choosing the ϕ's to have their supports in a subset $U'_n \subset U_n$ such that for $\{a_i\}$ sufficiently small their space translates, $\phi(a_n,\ldots,a_1)$ have their support still contained in U_n, we see that we can arrange matters such that

$$F(a_n t_n,\ldots,a_1 t_1) \equiv 0\,,$$

for an open set $\mathscr{O}$ of

$$\{(t_i a_i)\} \subset \mathbb{R}^{(\nu+1)\cdot n}\,. \tag{6.3}$$

This set $\mathscr{O}$ is part of the boundary of the analytic continuation of F into T^n. Proceeding as in Sec. II, we define

$$G(z_n \zeta_n,\ldots,z_1 \zeta_1) := \overline{F(\bar{z}_n \bar{\zeta}_n,\ldots,\bar{z}_1 \bar{\zeta}_1)}\,, \tag{6.4}$$

where G is now analytic in $\overline{T^n}$ and $\overline{T^n}$, T^n having a common real boundary set $\mathscr{O}$, where

$$F \equiv G \equiv 0 \tag{6.5}$$

holds.

Again we conclude that $F \equiv 0$ in T^n (by using the "edge of the wedge" theorem, see Ref. 2) which, by continuity, holds also for the real boundary, i.e.,

$$F(t_n a_n,\ldots,t_1 a_1) \equiv 0 \tag{6.6}$$

on $\mathbb{R}^{(\nu+1)\cdot n}$. Now we can set all time coordinates $\{t_i\}$ equal to zero and vary the a_i's independently in $\mathbb{R}$ yielding

$$\int \bar{\psi}(x_n,\ldots,x_1) \prod_{i=1}^{n} \phi_i(x_i + a_i)\, dX_n \equiv 0\,, \tag{6.7}$$

for all $\{a_1\}$. The ϕ_i's can be chosen to be arbitrary functions as long as $\Pi_{i=1}^{n} \phi_i$ has its support contained in $U'_n \subset U_n$. This, together with (6.7), implies that $\psi = 0$ in $L^2(\mathbb{R}^{\nu \cdot n})$, which proves the theorem.

ACKNOWLEDGMENTS

This paper was written while I was a guest at the California Institute of Technology. I would like to thank Professor W. A. Luxemburg and Professor B. Simon for their kind hospitality.

I would also like to thank the Deutsche Forschungsgemeinschaft for financial support.

[1] M. Requardt, J. Phys. A **15**, 3715 (1982).
[2] R. Streater and A. S. Wightman, *PCT, Spin and Statistics and All That* (Benjamin, New York, 1964).
[3] M. Reed and B. Simon, *Analysis of Operators* (Academic, New York, 1978).
[4] L. Hormander, *Linear Partial Differential Operators* (Springer, New York, 1964).
[5] W. O. Amrein, A. M. Berthier and V. Georgescue, Ann. Inst. Fourier (Grenoble) **XXXI**, 153 (1981).
[6] M. Schechter and B. Simon, J. Math. Anal. Appl. **77**, 482 (1980).
[7] L. Hormander, Commun. Partial Differential Equations **8**, 21 1983.
[8] D. Jerison and C. E. Kenig, University of Minnesota preprint, 1984.
[9] Y. Saito, "Spectral representations for Schrödinger operators," in *Lecture Notes in Mathematics*, Vol. 727 (Springer, New York, 1979).
[10] M. Reed and B. Simon, *Scattering Theory* (Academic, New York, 1979).
[11] B. Simon, Bull. Am. Math. Soc. **7**, 447 (1982).
[12] S. Agmon, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **2**, 151 (1975).
[13] G. Hegerfeldt and S. Ruigsenaars, Phys. Rev. D **22**, 377 (1980).
[14] L. Schwartz, *Theorie des Distributions* (Herman, Paris, 1966).
[15] M. Reed and B. Simon, *Fourier Analysis* (Academic, New York, 1975).
[16] H. Narnhofer, M. Requardt, and W. Thirring, Commun. Math. Phys. **92**, 247 (1983).

Cosmologies with a noninteracting mixture of dust and radiation

Edmond Weber
Université Catholique de Louvain, Institut de Physique Théorique, B-1348 Louvain-La-Neuve, Belgium and Université du Burundi, B.P. 2700 Bujumbura, Burundi[a]

(Received 30 December 1985; accepted for publication 21 February 1986)

In this paper a particular class of anisotropic cosmologies, the Kantowski–Sachs models, is considered. It is assumed that the matter content of the models consists of a noninteracting mixture of ordinary matter ("dust") and thermal radiation. A qualitative study by means of a three-dimensional autonomous system is carried out, giving us the global behavior of the "dust" density, the radiation density, and the shear anisotropy during the models' evolution. All the models have past and future cosmological singularities where both the dust density and the radiation density diverge. A particular interesting result is a set of solutions of three-measure zero, which is radiation dominated at one (past) singularity (of the "point" type) and evolving to a (future) singularity, where ordinary matter and thermal radiation become negligible.

I. INTRODUCTION

For a long time cosmologists have used the most simple solutions of Einstein's general relativity as applied to cosmology and have developed the so-called "standard model."[1] In spite of that, they have obtained remarkable results for the prediction of the present cosmic helium abundance, calculated on the assumption that our universe was very hot in the past, which in turn is based on the most prominent relic of this hot past, namely the 2.7 °K microwave background radiation.[2] This period is known as the "radiation dominated universe" when the pressure p_r of the thermal radiation was one third of its energy density ρ_r. Later, after one million years from the "big bang" onwards, matter and radiation decoupled and our universe became matter dominated: We still live in this universe where the matter density largely exceeds that of cosmic radiation.

The standard model is based on a very strong assumption: our universe is spatially homogeneous in general and isotropic around us. Although this is true on large scales nowadays, there is no reason to maintain this "cosmological principle"[3] for the very early history of our universe. For almost 20 years cosmologists have studied under the very effective stimulus of Misner's "chaotic cosmology,"[4] spatially homogeneous but anisotropic cosmologies belonging to the Bianchi class.[5–8] Part of this work was done by using qualitative techniques of plane autonomous systems,[9–14] because the Einstein field equations can be transformed quite easily into such a system when the above symmetries are assumed. An exceptional case to the Bianchi cosmologies was discovered by Kantowski and Sachs.[15] The resulting Kantowski–Sachs (KS) cosmologies have been analyzed extensively by Collins[13] using such global qualitative techniques and assuming a perfect fluid as matter content but for a vanishing cosmological term Λ. We have generalized[16–18] this work by allowing Λ to be nonzero, which leads us to consider this time a three-dimensional autonomous system. Interesting results emerged like the isotropy of some of these models when $\Lambda > 0$ and the cosmic time tends to infinity.

In this paper we suppose Λ to be zero but assume a much richer matter content for the KS models studied herein, i.e., a noninteracting mixture of ordinary matter ("dust") and thermal radiation. We also obtain in this case a three-dimensional autonomous system, but one for which the variables are now dust density ρ_d, the radiation density ρ_r, and the shear anisotropy σ. One interesting result is the existence of cosmological models of three-measure zero, which are radiation dominated at the beginning of their evolution and evolve to a singularity where matter and radiation are insignificant.

Our paper is organized as follows: In Sec. II we derive the three-dimensional system for the KS metric with the assumed content of our fluid and study this system by global qualitative techniques in Sec. III. The conclusion is drawn in Sec. IV.

II. QUALITATIVE ANALYSIS

We consider the KS metric[8,15]

$$ds^2 = dt^2 - \exp(-2\Omega)$$
$$\times\{e^{2\beta}\,dr^2 + e^{-\beta}(d\theta^2 + \sin^2\theta\, d\phi^2)\}$$

in the coordinates (t,r,θ,ϕ), where t is the cosmic time coordinate, r a radial coordinate, and θ,ϕ the usual spherical coordinates, and $\Omega = \Omega(t)$ and $\beta = \beta(t)$ are two unknown functions of t. As the fluid we will take, as mentioned above, a noninteracting mixture of dust and radiation, which means that the total energy density $\rho_{\text{total}} = \rho_d + \rho_r$ and the pressure $p_{\text{total}} = p_r = \rho_r/3$. Changing notations, we shall write $\rho = \mu + \epsilon$, $p = \epsilon/3$. Einstein's field equations with a nonzero Λ can be written as follows:

$$3\dot\Omega^2 - \tfrac{3}{4}\dot\beta^2 - \Lambda + e^{2\Omega+\beta} = \mu + \epsilon, \qquad (2.1)$$

$$6\ddot\Omega + 3\ddot\beta - 9\dot\Omega^2 - 9\dot\Omega\dot\beta - \tfrac{9}{4}\dot\beta^2 + 3\Lambda - 3e^{2\Omega+\beta} = \epsilon, \qquad (2.2)$$

$$6\ddot\Omega - \tfrac{3}{2}\ddot\beta - 9\dot\Omega^2 + \tfrac{9}{2}\dot\Omega\dot\beta - \tfrac{9}{4}\dot\beta^2 + 3\Lambda = \epsilon. \qquad (2.3)$$

In the following we will use Ω as a time variable and a prime will denote differentation with respect to Ω. The mean expansion rate $\theta = -3\dot\Omega$ and the shear $\sigma = (\sqrt{3}/2)\,\dot\beta$ enable us to define $\beta' = -2\sqrt{3}\,\sigma/\theta$ measuring the dynamic importance of shear. We introduce the quantities $x_1 = \mu/\dot\Omega^2$

[a] Present address.

and $x_2 = \epsilon/\dot{\Omega}^2$, measuring, respectively, the dynamic importance of dust and radiation. Finally we will make use of a fourth quantity $z = 3x_1/\mu = 3x_2/\epsilon = \dot{\Omega}^{-2}$.

These definitions allow us to reexpress the field equations (2.1)–(2.3) together with the two conservation equations

$$\mu' = 3\mu, \tag{2.4}$$

$$\epsilon' = 4\epsilon, \tag{2.5}$$

in the form of a four-dimensional autonomous system and a constraint equation with four dependent variables x_1, β', x_2, z and with Ω as the independent variable.

Equation (2.1) becomes

$$\beta'^2 - 4 + \tfrac{4}{3}\Lambda z + 4x_1 + 4x_2 = (4/3\dot{\Omega}^2)\exp(2\Omega + \beta). \tag{2.6}$$

By eliminating $\ddot{\Omega}$ from (2.2) and (2.3) and substituting the expression of the right-hand term $\exp(2\Omega + \beta)$ obtained from (2.6) we shall have

$$\beta'' = \tfrac{1}{2}\beta'\{4 - \beta'^2 + (2\Lambda/3)z - x_1 - 2x_2\}$$
$$- \tfrac{1}{2}\{4 - \beta'^2 - (4\Lambda/3)z - 4x_1 - 4x_2\}. \tag{2.7}$$

We obtain

$$6\ddot{\Omega} - 9\dot{\Omega}^2 - \tfrac{9}{4}\dot{\beta}^2 + 3\Lambda = \epsilon + \exp(2\Omega + \beta) \tag{2.8}$$

when we eliminate $\ddot{\beta}$ in (2.2) and (2.3).

By using (2.4) and (2.6) we get

$$x_1' = x_1\{1 - \beta'^2 + (2\Lambda/3)z - x_1 - 2x_2\}. \tag{2.9}$$

Analogously we get

$$x_2' = x_2\{2 - \beta'^2 + (2\Lambda/3)z - x_1 - 2x_2\}. \tag{2.10}$$

Differentiation of z with respect to Ω yields

$$z' = -2z\{1 + \beta'^2/2 - \Lambda z/3 + x_1/2 + x_2\}. \tag{2.11}$$

Equations (2.7) and (2.9)–(2.11) form a four-dimensional autonomous system of ordinary differential equations. No attempt has been made as yet to solve such a system globally by qualitative methods.

Unfortunately there do not exist theorems to describe the topological behavior around singular points in this (four-dimensional) case, as we have done for a three-dimensional system in a previous paper[16] (hereafter referred to as Paper I). The qualitative behavior of the solutions would be drawn in the (x_1, β', x_2, z) phase space and the region of physical interest would be given by $(x_1 > 0,\ x_2 > 0,\ z > 0,\ \beta'^2 - 4 + (4\Lambda/3)z + 4x_1 + 4x_2 > 0)$. Let us point out that there is a one-to-one correspondence between the solutions of the original system (2.1)–(2.5) and the transformed one (2.6), (2.7), and (2.9)–(2.11), but that the advantage of transforming the original system lies clearly in the fact that as a result we get an autonomous system. In the following we shall limit ourselves to particular three-dimensional cases.

When we set $x_2 = 0$ in (2.7) and (2.9)–(2.11) we obtain a three-dimensional autonomous system in the variables (x_1, β', z): it is the particular case $\gamma = 1$ of the system studied in Paper I.

Similarly by setting $x_1 = 0$ we obtain another three-dimensional system that corresponds to the particular case $\gamma = \tfrac{4}{3}$ in Paper I.

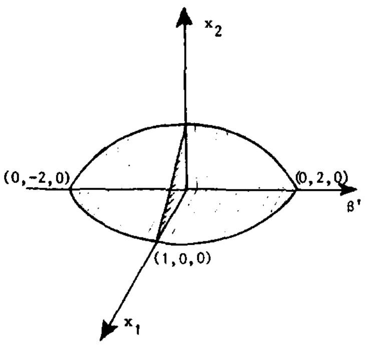

FIG. 1. The region of physical interest for the three-dimensional autonomous system as well as the singular points at finite distance are shown in the case of the KS models. Striped parts are outside the region of physical interest.

When the cosmological term vanishes we get a new three-dimensional autonomous system

$$\beta'' = \tfrac{1}{2}\beta'(4 - \beta'^2 - x_1 - 2x_2)$$
$$- \tfrac{1}{2}(4 - \beta'^2 - 4x_1 - 4x_2), \tag{2.12}$$

$$x_1' = x_1(1 - \beta'^2 - x_1 - 2x_2), \tag{2.13}$$

$$x_2' = x_2(2 - \beta'^2 - x_1 - 2x_2). \tag{2.14}$$

It describes the KS models in the presence of a noninteracting mixture of dust and radiation. It is a combination into a three-dimensional system of two particular cases ($\gamma = 1$ and $\gamma = \tfrac{4}{3}$) of a plane autonomous system studied by Collins[13] (hereafter referred to as Paper II). The region of physical interest is given by $(x_1 > 0, x_2 > 0, \beta'^2 - 4 + 4x_1 + 4x_2 > 0)$ and the singular points at finite distance are $(x_1 = 0, \beta' = \pm 2, x_2 = 0)$, $(x_1 = 0, \beta' = 0, x_2 = 1)$, and $(x_1 = 1, \beta' = 0, x_2 = 0)$ (see Fig. 1).

Let us remark that the system (2.12)–(2.14) describes also the orthogonal Bianchi models of type III, with the same content of matter and radiation and a vanishing cosmological term. The only difference with the KS case lies in the region of physical interest, which is now given by $(x_1 > 0, x_2 = 0, \beta'^2 - 4 + 4x_1 + 4x_2 < 0)$. We have in addition to the singular points given above the point $(x_1 = 0,\ \beta' = 1,\ x_2 = 0)$. The two plane systems obtained for $x_1 = 0$ and $x_2 = 0$ have been studied by Collins[10] for the two cases $\gamma = 1$ and $\gamma = \tfrac{4}{3}$.

III. GLOBAL QUALITATIVE STUDY

In order to study qualitatively the three-dimensional system in the KS case we examine first the behavior of integral curves in the neighborhood of the critical points at finite distance as well as the infinity. We do this by applying the theorems indicated in the Appendix of Paper I. By analyzing the three surfaces $dx_1/d\Omega = 0, d\beta'/d\Omega = 0, dx_2/d\Omega = 0$ we obtain then a global picture of the orbits.

All the critical points at finite distance are simple. The point (0,2,0) is a stable node with negative characteristic roots: $\lambda = -3, \mu = -2, \nu = -2$. The orbits starting at a sphere centered at (0,2,0) tend to this point for $\Omega \to \infty$. The characteristic vector $e_1 = (1, -1, 0)$ corresponds to the λ root; we have an infinity of characteristic vectors in the plane (β', x_2) [Fig. 2(a)]. There is a double infinity of physically

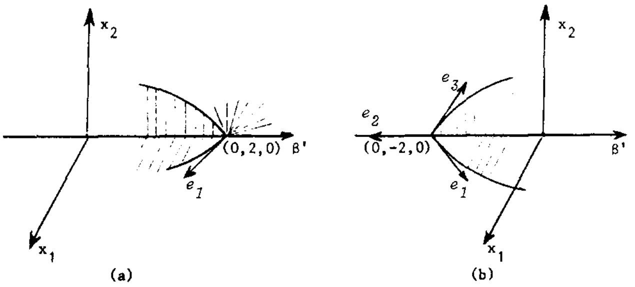

FIG. 2. The region of physical interest that is outside the striped parts in the neighborhoods of the two points $(0, +2,0)$ and $(0, -2,0)$ is depicted. e_1,e_2,e_3 are the three characteristic vectors.

interesting orbits starting at the sphere and tending to $(0,2,0)$ along the plane (β',x_2); there is only one orbit along the vector e_1 given by $\beta'^2 - 4 + 4x_1 = 0$.

The point $(0, -2,0)$ is also a stable node whose roots are $\lambda = -3, \mu = -6, \nu = -2$. The corresponding characteristic vectors are $e_1 = (1,1,0)$, $e_2 = (0,1,0)$, and $e_3 = (0,1,1)$, respectively [Fig. 2(b)]. A double infinity of orbits tends to $(0, -2,0)$ alongside the vector e_3. There is a simple infinity of orbits tending to $(0, -2,0)$ alongside e_1; these orbits correspond to those tending to $(x = 0, \beta' = -2)$ in the particular plane case $\gamma = 1$ studied in Paper II.

The singular point $(0,0,1)$ is an (unstable) saddle point with $\lambda = -1, \mu = 1$, and $\nu = -2$, and corresponding vectors $e_1 = (1,0,-1)$, $e_2 = (0,1,0)$, and $e_3 = (0,-1,\frac{3}{2})$. We find a simple infinity of orbits tending to $(0,0,1)$ along e_1, with $\Omega \to \infty$. There is only one orbit along e_2 given by $\beta'^2 - 4 + 4x_2 = 0$, and only one along e_3, corresponding to the one tending to $(x = 1, \beta' = 0)$ in the plane case $\gamma = \frac{4}{3}$ (Paper II).

In the plane (x_1, β') we have the singular point $(1,0,0)$, which is a saddle point with $\lambda = -1, \mu = \frac{3}{2}$, and $\nu = 1$. The corresponding vectors are $e_1 = (1, -\frac{4}{5},0)$, $e_2 = (0,1,0)$, and $e_3 = (-1,0,1)$. There is only one orbit along e_1 (as for the plane case $\gamma = 1$ in Paper II), only one along e_2 given by the equation $\beta'^2 - 4 + 4x_1 = 0$, and a simple infinity along the vector e_3 given by the equation $\beta'^2 - 4 + 4x_1 + 4x_2 = 0$.

The Poincaré transformations[16,19] $x_1 = s^{-1}, \beta' = us^{-1}$, and $x_2 = vs^{-1}$ enable us to study the critical points at infinity that are not in the plane (β',x_2): these are the double singular points $(s = 0, u = 0, v \geqslant 0)$ (Fig. 3). The general expression of the directions of approach (see the Appendix in Paper I), not in the plane $(s = 0)$, of the singular points $(s = 0, u = 0, v = v_0)$ is given by

$$\{\omega_i\} = \left\{\frac{1}{2(1+2v_0)}, \frac{1}{(\frac{1}{2}+v_0)(1+2v_0)}, \frac{v_0}{2(1+2v_0)^2}\right\}.$$

One should notice that $\tan\psi = \omega_1/\omega_2$ reduces to $\frac{1}{4}$ when $v_0 = 0$ and $\frac{1}{2}$ when $v_0 \to \infty$, which agrees with the results of the two plane cases when $\gamma = 1$ and $\gamma = \frac{4}{3}$ in Paper II.

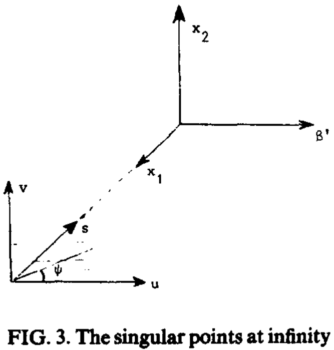

FIG. 3. The singular points at infinity obtained by the Poincaré transformations $x_1 = s^{-1}, \beta' = us^{-1}, x_2 = vs^{-1}$ are $(s = 0, u = 0, v > 0)$. The directions of approach in the plane $(s = 0)$ are depicted by the parallel lines. The line in the plane (u,s) with $\psi = \arctan\frac{1}{4}$ shows a particular direction of approach not in the plane $(s = 0)$.

By studying the three surfaces $dx_1/d\Omega = 0$, $d\beta'/d\Omega = 0$, $dx_2/d\Omega = 0$, we obtain a global picture of the orbits. We have a double infinity of orbits starting at $(0,2,0)$ and tending to $(0, -2,0)$, becoming tangent to the singular line at infinity with $\tan\psi = 0$. There is a time-symmetric surface of orbits starting at $(0,2,0)$ and approaching the singular points at infinity with $\psi = \arctan \omega_1/\omega_2$. We have further a double infinity or orbits starting at $(0,2,0)$ extending to infinity with $\psi = \pi$ and coming back to the same singular point. There is finally a simple infinity of orbits starting at $(0,0,1)$ extending to infinity with $\psi = \pi$ and tending to $(0,2,0)$. The time reverses of all these models are also feasible.

The general features of the singularities are as follows. There is one "cigar" singularity $(X \to \infty, Y \to 0)$ represented by the point $(0,2,0)$, and one "pancake" singularity $(X \to 0, Y \to \text{const}$, where const is a positive constant) represented by the point $(0, -2,0)$. The singular point $(0,0,1)$ is a "point" singularity $(X \to 0, Y \to 0)$. The Raychaudhuri equation

$$\dot{\theta} + \theta^2/3 + 2\sigma^2 + \tfrac{1}{2}(\mu + 2\epsilon) = 0$$

tells us that the dust density μ is insignificant at all these singularities, and that the radiation density ϵ is dominant only around the singular point $(0,0,1)$. The expansion rate θ is dominant at all three points whereas the fluid shear σ is only important around the points $(0, \pm 2,0)$. All these singularities are of a cosmological nature, i.e., it takes a finite cosmic time t to get there $(\Omega \to \infty : t \to 0_\pm)$. The two variables μ and ϵ diverge at all the singularities. There are particle and event horizons in the sense defined by Rindler[20] in all directions except in the $\partial/\partial r$ direction around the singularity $(0, -2,0)$ for which these horizons are removed. We have summarized all the information about the asymptotic behavior of the models around the singularities in Table I.

IV. CONCLUSION

We have carried out a detailed analysis of KS models containing a noninteracting mixture of ordinary matter and thermal radiation. Let us be reminded that a very interesting set of solutions was found: models that are radiation dominated at the past singularity and evolving to another future singularity where matter and radiation are insignificant. The two singularities are of a cosmological nature and accordingly the average length scale l vanishes there. The past singularity is of the point type and the future one is a cigar singularity. Particle and event horizons exist around these singularities in all directions.

Although the KS cosmologies are of a very special kind, the qualitative study by means of a three-dimensional auton-

TABLE I. In this table the different singularity types are indicated as well as the asymptotic behaviors on the physically relevant variables: the average length scale l, the length scales X and Y, the dust density μ, the radiation density ϵ, the fluid expansion θ, the fluid shear σ, and the integrated shear β. The notation const stands for a nonzero finite limit. The upper sign corresponds to a past singularity and the lower one to a future singularity.

(x_1, β', x_2)	Singularity type		Particle and event horizon	$l = e^{-\Omega}$	Cosmological singularity	X	Y	μ	ϵ	θ	σ	β	Dominant terms in Raychaudhuri equation
(0,2,0)	cigar	$X \to \infty$ $Y \to 0$	in all directions	$(\pm t)^{1/3}$	$\Omega \to \infty : t \to 0_{\pm}$	$(\pm t)^{-1/3}$	$(\pm t)^{2/3}$	$(\pm t)^{-1}$	$(\pm t)^{-4/3}$	t^{-1}	$-t^{-1}$	$-\ln\|t\|$	$\sigma^2, \theta^2 \sim t^{-2}$
(0, − 2,0)	pancake	$X \to 0$ $Y \to$const	removed in $\partial/\partial r$ direction	$(\pm t)^{1/3}$	$\Omega \to \infty : t \to 0_{\pm}$	$\pm t$	const	$(\pm t)^{-1}$	$(\pm t)^{-4/3}$	t^{-1}	t^{-1}	$\ln\|t\|$	$\sigma^2, \theta^2 \sim t^{-2}$
(0,0,1)	point	$X \to 0$ $Y \to 0$	in all directions	$(\pm t)^{1/2}$	$\Omega \to \infty : t \to 0_{\pm}$	$(\pm t)^{1/2}$	$(\pm t)^{1/2}$	$(\pm t)^{-3/2}$	$(\pm t)^{-2}$	t^{-1}	t^{-1}	$\ln\|t\|$	$\epsilon, \theta^2 \sim t^{-2}$

omous system of these models gives us an important tool to investigate more "realistic" models in comparison with the real universe in its entire evolution, invoking eventually quantum processes, and gives hope for other interesting results when applied to the Bianchi class of cosmologies.

ACKNOWLEDGMENTS

Professor Dr. J. Ehlers is warmly thanked for reading the manuscript and giving very useful comments. I am also grateful to the Institute of Theoretical Physics at Louvain-la-Neuve (Belgium) for their technical assistance in preparing this paper.

[1] S. Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1972).
[2] A. A. Penzias and R. W. Wilson, Astrophys. J. **142**, 419 (1965).
[3] E. Milne, Z. Astrophys. **6**, 1 (1933).
[4] C. W. Misner, Astrophys. J. **151**, 431 (1968).
[5] L. Bianchi, *Lezioni sulla teoria dei gruppi continui finiti di transformazioni* (Spoerri, Pisa, 1918).
[6] F. B. Estabrook, H. D. Wahlquist, and C. G. Behr, J. Math. Phys. **9**, 497 (1968).
[7] G. F. R. Ellis and M. A. H. MacCallum, Commun. Math. Phys. **12**, 108 (1969).
[8] M. P. Ryan and L. C. Shepley, *Homogeneous Relativistic Cosmologies* (Princeton U.P., Princeton, NJ, 1975).
[9] C. B. Collins and J. M. Stewart, Mon. Not. R. Astron. Soc. **153**, 419 (1971).
[10] C. B. Collins, Commun. Math. Phys. **23**, 137 (1971).
[11] C. B. Collins, Commun. Math. Phys. **27**, 37 (1972).
[12] C. B. Collins, Commun. Math. Phys. **39**, 131 (1974).
[13] C. B. Collins, J. Math. Phys. **18**, 2116 (1977).
[14] C. B. Collins and G. F. R. Ellis, Phys. Rep. **56**, 65 (1979).
[15] R. Kantowski and R. K. Sachs, J. Math. Phys. **7**, 443 (1966).
[16] E. Weber, J. Math. Phys. **25**, 3279 (1984).
[17] E. Weber, J. Math. Phys. **26**, 1308 (1985).
[18] E. Weber, "*Global qualitative study of Bianchi Universes in presence of a cosmological constant*," Preprint UCL-IPT-84-24, submitted to J. Math. Phys.
[19] A. A. Andronov, E. A. Leontovich, I. I. Gordon, and A. G. Maier, *Qualitative Theory of Second-order Dynamic Systems* (Wiley, New York, 1973).
[20] W. Rindler, Mon. Not. R. Astron. Soc. **116**, 662 (1956).

Killing spinors and gravitational perturbations

G. F. Torres del Castillo

Departamento de Física Matemática, Instituto de Ciencias de la Universidad Autónoma de Puebla, Puebla, 72000 Mexico and Departamento de Física, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Apartado Postal 14-740, 07000 Mexico, D.F., Mexico

(Received 29 August 1985; accepted for publication 12 February 1986)

It is shown that in a vacuum space-time, possibly with a nonzero cosmological constant, which admits a $D(1,0)$ Killing spinor, one component of the perturbed Weyl spinor that satisfies a decoupled equation, when multiplied by an appropriate factor made out of the components of the Killing spinor, constitutes a Debye potential that generates metric perturbations of the considered background. It is also shown that in the case where the background is of type N, there is an operation that relates the gravitational perturbations and the zero-rest-mass fields of spin-0, -$\frac{1}{2}$, and -1.

I. INTRODUCTION

In a recent paper[1] it was shown that the existence of a $D(1,0)$ Killing spinor L_{AB} in an otherwise arbitrary space-time, implies that one can construct a solution of the spin-$\frac{1}{2}$ and spin-1 zero-rest-mass field equations from a given solution of the corresponding field equations. Moreover, the new solution constructed in this way can be given in terms of a single component of the old one, obtained by the full contraction of the spinorial components of the field with a principal spinor of L_{AB}. In the case of the electromagnetic perturbations of a vacuum type D space-time, the fact that one component of the electromagnetic field can be used to generate another full solution was previously noted by Wald.[2]

Even though there is no known geometric interpretation for a $D(1,0)$ Killing spinor, it seems that the existence of such a spinor field is associated with the separability of the spin-$\frac{1}{2}$ and spin-1 zero-rest-mass field equations (see, e.g., Refs. 3 and 4).

The aim of this paper is to show that in the case of a vacuum space-time (with or without cosmological term), the existence of a $D(1,0)$ Killing spinor permits the construction of a full metric perturbation of the background space-time from a single component of the perturbed Weyl spinor, which satisfies a decoupled equation. In the specific case of the Kerr metric, which admits a $D(1,0)$ Killing spinor, this fact implies that one can obtain full metric perturbations from the separated Teukolsky functions (cf. Refs. 5 and 6).

Most investigations on Killing spinors have been restricted to the case of algebraically general Killing spinors of valence 2; and therefore to type D or conformally flat space-times (see, e.g., Refs. 4, 7, and 8). In the present paper the two possible algebraic types of the Killing spinors of valence 2 are considered. Some results previously known valid in the case of type D vacuum space-times, derived by other means, are included; pointing out, however, their origin in the existence of a Killing spinor.

In Sec. II a brief description of the Debye potentials for gravitational perturbations of algebraically special space-times is given. In Sec. III we show how a $D(1,0)$ Killing spinor relates the component of the perturbed Weyl spinor along a repeated principal null direction with the gravitational Debye potentials. In the case where the unperturbed Weyl spinor is of type N, we show that there is a connection between the gravitational perturbations and the zero-rest-mass fields of spin-0, -$\frac{1}{2}$, and -1.

II. GRAVITATIONAL DEBYE POTENTIALS

As was originally conjectured by Chrzanowski,[5] the solutions of a certain linear second-order differential equation for a scalar potential lead, by differentiation, to metric perturbations of a given arbitrary algebraically special vacuum space-time. Independent proofs of the validity of this conjecture have been given by Kegeles and Cohen,[9] through a very lengthy computation, and by Wald,[2] who, based on a more general result, devised a very simple derivation of the expressions proposed by Chrzanowski.[10]

Wald's derivation depends on the fact that the linearized Einstein operator is self-adjoint and that, for an algebraically special vacuum space-time, there is a decoupled equation for a component of the perturbed Weyl spinor. In a spinor frame such that the components of the unperturbed Weyl spinor satisfy $\Psi_0 = \Psi_1 = 0$, the component $\dot{\Psi}_0$ of the perturbed Weyl spinor obeys the decoupled equation[11]

$$
\begin{aligned}
&[(D - 4\rho - 3\epsilon - \rho^* + \epsilon^*)(\Delta - 4\gamma + \mu) \\
&\quad - (\delta - 4\tau - 3\beta - \alpha^* + \pi^*)(\delta^* - 4\alpha + \pi) \\
&\quad - 3\Psi_2]\dot{\Psi}_0 = 0. \qquad (1)
\end{aligned}
$$

Solutions to the adjoint equation of (1),

$$
\begin{aligned}
&[(\Delta + 3\gamma + \mu^* - \gamma^*)(D + 4\epsilon + 3\rho) \\
&\quad - (\delta^* + 3\alpha + \beta^* - \tau^*)(\delta + 4\beta + 3\tau) \\
&\quad - 3\Psi_2]\psi = 0, \qquad (2)
\end{aligned}
$$

generate metric perturbations of the given background space-time according to

$$
\begin{aligned}
h_{\mu\nu} = &- \{l_\mu l_\nu [(\delta + \alpha^* + 3\beta - \tau)(\delta + 4\beta + 3\tau) - \lambda^*(D + 4\epsilon + 3\rho)] + m_\mu m_\nu (D - \epsilon^* + 3\epsilon - \rho)(D + 4\epsilon + 3\rho) \\
&- l_{(\mu} m_{\nu)} [(D + \epsilon^* + 3\epsilon + \rho^* - \rho)(\delta + 4\beta + 3\tau) + (\delta - \alpha^* + 3\beta - \pi^* - \tau)(D + 4\epsilon + 3\rho)]\}\psi + \text{c.c.}, \qquad (3)
\end{aligned}
$$

0022-2488/86/061583-03$02.50

where l_μ and m_μ are Newman–Penrose tetrad vectors.[2,5,9,10]

Teukolsky's equation (1), which follows from the Bianchi identities, also applies when there exists a nonzero cosmological constant and, therefore, Eqs. (1)–(3) are valid for all algebraically special vacuum space-times with cosmological constant.

III. GENERATION OF PERTURBATIONS

A $D(1,0)$ Killing spinor[7] is a symmetric spinor field L_{AB} that satisfies

$$\nabla_{(A}{}^{\dot{R}}L_{BC)}=0 \tag{4}$$

$(A, B, C,... = 1,2; \dot{R},... = \dot{1},\dot{2})$. If the space-time is not conformally flat, then the existence of a nontrivial solution of (4) implies that the conformal curvature is of type D or N, i.e., the space-time is algebraically special, and the solution of (4) is unique modulo a constant factor.[12]

In the forthcoming we will assume that the space-time admits a $D(1,0)$ Killing spinor and that the traceless part of the Ricci tensor vanishes ($\Phi_{ij}=0$, in the Newman–Penrose notation), i.e., the Einstein vacuum field equations, possibly with a nonzero cosmological constant, are satisfied. Then

$$K_{A\dot{R}}=\nabla^{B}{}_{\dot{R}}L_{AB} \tag{5}$$

is a (complex) Killing vector field.[8]

It will be shown that the decoupled components of the perturbed Weyl spinor and the gravitational Debye potentials are related through the components of the Killing spinor L_{AB}. In the case where the conformal curvature is of type D, this relation has been previously given.[2] However, for the sake of completeness, we shall also treat this case briefly.

A. Type *D*

When the conformal curvature is of type D then, in a frame such that $\Psi_0=\Psi_1=\Psi_3=\Psi_4=0$, the only nonvanishing components of L_{AB}, $L_{12}=L_{21}$, are given by $L_{12}=\text{const}(\Psi_2)^{-1/3}$ (see Ref. 7). Due to the existence of two repeated principal null directions of the conformal curvature, the components $\dot{\Psi}_0$ and $\dot{\Psi}_4$ of the perturbed Weyl spinor satisfy decoupled equations; $\dot{\Psi}_0$ satisfies Eq. (1) and $\dot{\Psi}_4$ obeys[11]

$$\begin{aligned}&[(\Delta+4\mu+3\gamma+\mu^*-\gamma^*)(D+4\epsilon-\rho)\\&\quad-(\delta^*+4\pi+3\alpha+\beta^*-\tau^*)(\delta+4\beta-\tau)\\&\quad-3\Psi_2]\dot{\Psi}_4=0.\end{aligned} \tag{6}$$

By using Eq. (4), it is easy to see that if $\dot{\Psi}_4$ is a solution of (6), then

$$\psi=(L_{12})^4\dot{\Psi}_4 \tag{7}$$

is a solution of (2), and conversely. This result, with L_{12} expressed in terms of Ψ_2, was previously given by Wald[2]; without realizing, however, the role played by the existing Killing spinor.

In a similar manner, if $\dot{\Psi}_0$ is a solution of (1), then $\psi'=(L_{12})^4\dot{\Psi}_0$ is a solution of the adjoint equation of (6), and conversely. The gravitational Debye potential ψ' also yields metric perturbations of the background space-time by an expression analogous to (3).

B. Type *N*

When the conformal curvature is of type N, then, in a frame such that $\Psi_0=\Psi_1=\Psi_2=\Psi_3=0$, the only nonvanishing component of L_{AB}, L_{22} is determined by

$$\begin{aligned}&D\ln L_{22}=-2(\rho+\epsilon),\quad \delta\ln L_{22}=-2(\tau+\beta),\\&\delta^*\ln L_{22}=-2\alpha,\quad \Delta\ln L_{22}=-2\gamma.\end{aligned} \tag{8}$$

A straightforward but somewhat lengthy computation, using Eqs. (8) and the commutation relations for the tetrad vectors, shows that if $\dot{\Psi}_0$ satisfies Eq. (1), then

$$\psi=(L_{22})^4\dot{\Psi}_0 \tag{9}$$

satisfies Eq. (2), and conversely. In the present case, in contrast with the type D case, conditions $\Phi_{ij}=0$ do not imply the existence of a Killing spinor L_{AB}. In fact, there are type N vacuum space-times where Eqs. (8) are not integrable.

By using Eqs. (8), one finds that the Killing vector (5) is given by

$$K=L_{22}(\tau D-\rho\delta), \tag{10}$$

modulo a constant factor.

When the conformal curvature is of type N, the solution L_{AB}, of Eq. (4) must be algebraically special and it can be written as $L_{AB}=L_AL_B$, where L_A is a $D(\frac{1}{2},0)$ Killing spinor,[12] i.e., L_A satisfies

$$\nabla_{(A}{}^{\dot{R}}L_{B)}=0. \tag{11}$$

In fact, Eq. (11) is integrable only if the conformal curvature is of type N or equal to zero.

If ϕ_{AB} is a solution of the source-free Maxwell equations, $\nabla^{A\dot{R}}\phi_{AB}=0$, then, by virtue of (11), $\phi_A=\phi_{AB}L^B$ satisfies the Weyl neutrino equation $\nabla^{A\dot{R}}\phi_A=0$ (cf. Ref. 13). Likewise, $\chi=\phi_AL^A$ satisfies the massless field equation for spin-0

$$(\nabla^\mu\nabla_\mu+R/6)\chi=0, \tag{12}$$

where R denotes the scalar curvature, provided that ϕ_A obeys the Weyl neutrino equation. Similarly, from the Bianchi identities for a vacuum (those that do not involve Ψ_4), with $\Psi_0=\Psi_1=\Psi_2=\Psi_3=0$ and $\kappa=\sigma=0$, it follows that

$$\begin{aligned}&(\delta^*-4\alpha+\pi)\dot{\Psi}_0-(D-4\rho-2\epsilon)\dot{\Psi}_1=0,\\&(\Delta-4\gamma+\mu)\dot{\Psi}_0-(\delta-4\tau-2\beta)\dot{\Psi}_1=0,\\&(\delta^*-2\alpha+2\pi)\dot{\Psi}_1-(D-3\rho)\dot{\Psi}_2=\lambda\dot{\Psi}_0,\\&(\Delta-2\gamma+2\mu)\dot{\Psi}_1-(\delta-3\tau)\dot{\Psi}_2=\nu\dot{\Psi}_0,\end{aligned} \tag{13}$$

then, using Eqs. (8), one finds that $\phi_0\equiv L_{22}\dot{\Psi}_0$, $\phi_1\equiv L_{22}\dot{\Psi}_1$, and $\phi_2\equiv L_{22}\dot{\Psi}_2$ satisfy Maxwell's equations; i.e., even though the perturbed Weyl spinor *does not* satisfy the equation $\nabla^{A\dot{R}}\dot{\Psi}_{ABCD}=0$ when $\dot{\Psi}_4\neq 0$, $\phi_{AB}\equiv\dot{\Psi}_{ABCD}L^{CD}$ does satisfy Maxwell's equations (cf. Ref. 4), and therefore,

$$\phi_A=\phi_{AB}L^B=\dot{\Psi}_{ABCD}L^BL^CL^D$$

is a solution of the Weyl neutrino equation and $\chi=\dot{\Psi}_{ABCD}L^AL^BL^CL^D$ satisfies Eq. (12).

Since the existence of a Killing spinor L_{AB} establishes a connection between the Debye potentials and the field components in the cases of the gravitational perturbations of vacuum space-times and of the spin-$\frac{1}{2}$ and spin-1 zero-rest-mass fields,[1] one could also expect some relationship between the Debye potentials for these fields. Indeed, by using Eqs. (8), it is clear that if ψ_N is a Debye potential for the Weyl neutrino field,[9,14] i.e.,

$$[(\Delta+\mu^*-\gamma^*)(D+\epsilon) \\ \quad -(\delta^*+\beta^*-\tau^*)(\delta+\beta)]\psi_N=0, \qquad (14)$$

then $\psi_E=L_2\psi_N=(L_{22})^{1/2}\psi_N$ is a Debye potential for the Maxwell field,

$$[(\Delta+\gamma+\mu^*-\gamma^*)(D+2\epsilon+\rho) \\ \quad -(\delta^*+\alpha+\beta^*-\tau^*)(\delta+2\beta+\tau)]\psi_E=0, \qquad (15)$$

and conversely. As is shown in Refs. 14 and 15, under the present assumptions ($\Phi_{ij}=0$ and algebraic degeneracy of the Weyl spinor), the most general solution of Weyl's neutrino equation or Maxwell's equations is given locally in terms of a scalar potential that satisfies Eq. (14) or (15), respectively. The field components are given by

$$\phi_{\dot A}=\phi^{-1}\nabla_{B\dot A}(\phi\psi_N l^B) \qquad (16)$$

and

$$\phi_{\dot A\dot B}=\nabla^C{}_{(\dot A}\phi^{-2}\nabla^D{}_{\dot B)}[\phi^2\psi_E l_C l_D], \qquad (17)$$

where l_A is the repeated principal direction of the Weyl spinor and ϕ is a function such that

$$l^B\nabla_{A\dot C}l_B=l_A l^B\partial_{B\dot C}\ln\phi \qquad (18)$$

(i.e., $\rho=D\ln\phi, \tau=\delta\ln\phi$). In a similar way, one finds that the Debye potentials ψ_N and ψ_E are related with the solutions of Eqs. (2) and (12) by

$$\psi=(L_2)^2\psi_E=(L_2)^3\psi_N=(L_2)^4\chi.$$

Computation of the components of the Maxwell field generated by the Debye potential $\psi_E=L_2\psi_N$ according to (17), using that $L_A=L_2 l_A$, yields

$$\phi_{\dot A\dot B}=\nabla^C{}_{(\dot A}\phi^{-2}\nabla^D{}_{\dot B)}[\phi^2\psi_N L_C l_D] \\ =\nabla^C{}_{(\dot A}[\phi^{-1}L_C\nabla^D{}_{\dot B)}(\phi\psi_N l_D)+\psi_N l_D\nabla^D{}_{\dot B)}L_C \\ +\psi_N l_D L_C\,\partial^D{}_{\dot B}\ln\phi],$$

then, using Eq. (16) and

$$l_D\nabla^D{}_{\dot B}L_C=-l_D\nabla_{C\dot B}L^D=L^D\nabla_{C\dot B}l_D=L_C l^D\partial_{D\dot B}\ln\phi$$

[see Eq. (18)], one obtains

$$\phi_{\dot A\dot B}=\nabla_{C(\dot A}L^C\phi_{\dot B)}=L^C\nabla_{C\dot A}\phi_{\dot B}-\phi_{(\dot A}\nabla^C{}_{\dot B)}L_C, \qquad (19)$$

where $\phi_{\dot A}$ are the components of the Weyl neutrino field generated by ψ_N. Similarly, the Weyl neutrino field generated by $\psi_N=L_2\chi$, where χ is a solution of Eq. (12), is given by

$$\phi_{\dot A}=\phi^{-1}\nabla_{B\dot A}(\phi\chi L^B) \\ =L^B\partial_{B\dot A}\chi+\chi L^B\partial_{B\dot A}\ln\phi+\chi\nabla_{B\dot A}L^B.$$

Then, using the fact that $\nabla^B{}_{\dot A}L_B=2L^B\partial_{B\dot A}\ln\phi$ [which can be deduced from (18) and $\nabla^{A\dot R}\phi^3L_AL_B=0$ (see Ref. 1)], one gets

$$\phi_{\dot A}=L^B\partial_{B\dot A}\chi-\tfrac{1}{2}\chi\nabla^B{}_{\dot A}L_B. \qquad (20)$$

Since the most general solution $\phi_{\dot A\dot B}$ of the Maxwell equation is given locally by Eq. (17) in terms of the solution ψ_E of Eq. (15), $\phi_{\dot A\dot B}$ is given also by Eq. (19) in terms of the Weyl neutrino field generated by $\psi_N=\psi_E/L_2$. Similarly, it follows that the most general solution of Weyl's neutrino equation is given, locally, by Eq. (20) in terms of the solutions of Eq. (12). Expressions (19) and (20) together with the relations $\phi_A=\phi_{AB}L^B$ and $\chi=\phi_A L^A$ coincide with those found in the case of flat space-time, which lead to the interpretation of a twistor as a helicity-raising operator for massless fields [see Ref. 13, Eqs. (4.38) and (4.37)]. [It may be remarked that Eqs. (19) and (20) apply without any explicit restriction on the Ricci tensor, provided that L_A satisfies Eq. (11).]

IV. CONCLUDING REMARKS

We have shown that, in a background space-time that admits a $D(1,0)$ Killing spinor, an appropriate component of a Weyl spinor perturbation generates a full metric perturbation such that the corresponding perturbed Weyl spinor is, in general, different from the starting one. It seems reasonable to conjecture that the existence of a $D(1,0)$ Killing spinor is associated with the separability of the decoupled equations for the Weyl spinor perturbations. In fact, Dudley and Finley[16] have shown that for all the type D vacuum space-times, which admit a $D(1,0)$ Killing spinor, the equations for the radiative components of the gravitational perturbations are separable.

ACKNOWLEDGMENTS

The author wishes to thank Professor J. F. Plebański and Professor J. D. Finley, III for helpful comments.

The author also acknowledges support from the Sistema Nacional de Investigadores (Mexico).

[1] G. F. Torres del Castillo, Proc. R. Soc. London Ser. A **400**, 119 (1985).
[2] R. M. Wald, Phys. Rev. Lett. **41**, 203 (1978).
[3] N. Kamran and R. G. McLenaghan, Lett. Math. Phys. **7**, 381 (1983).
[4] B. P. Jeffryes, Proc. R. Soc. London Ser. A **392**, 323 (1984).
[5] P. L. Chrzanowski, Phys. Rev. D **11**, 2042 (1975).
[6] S. Chandrasekhar, in *General Relativity: An Einstein Centenary Survey,* edited by S. W. Hawking and W. Israel (Cambridge U. P., Cambridge, England, 1979).
[7] M. Walker and R. Penrose, Commun. Math. Phys. **18**, 265 (1970).
[8] L. P. Hughston and P. Sommers, Commun. Math. Phys. **33**, 129 (1973).
[9] L. S. Kegeles and J. M. Cohen, Phys. Rev. D **19**, 1641 (1979).
[10] For another derivation see G. F. Torres del Castillo, J. Math. Phys. **27**, 1586 (1986).
[11] S. A. Teukolsky, Astrophys. J. **185**, 635 (1973).
[12] G. F. Torres del Castillo, Commun. Math. Phys. **92**, 485 (1984).
[13] R. Penrose, in *Quantum Gravity: An Oxford Symposium,* edited by C. J. Isham, R. Penrose, and D. W. Sciama (Clarendon, Oxford, 1975).
[14] G. F. Torres del Castillo, J. Math. Phys. **25**, 342 (1984).
[15] G. F. Torres del Castillo, J. Math. Phys. **24**, 590 (1983).
[16] A. L. Dudley and J. D. Finley, III, J. Math. Phys. **20**, 311 (1979).

Gravitational perturbations of algebraically special space-times via the $\mathscr{H}\mathscr{H}$ equation

G. F. Torres del Castillo
Departamento de Física Matemática, Instituto de Ciencias de la Universidad Autónoma de Puebla, Puebla, Mexico and Departamento de Física, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Apartado Postal 14-740, 07000 México, D. F., Mexico

(Received 10 October 1985; accepted for publication 12 February 1986)

By using the approach to the Einstein field equations based on the existence of a congruence of null two-dimensional surfaces, it is shown that a scalar potential that satisfies a second-order linear partial differential equation generates gravitational perturbations of a given algebraically special solution of the Einstein vacuum field equations with cosmological constant. Generalizations of this result to the case of simultaneous perturbations of the gravitational and the matter fields are also indicated.

I. INTRODUCTION

A gravitational perturbation of a solution of the Einstein vacuum field equations, g, is a "small variation," $\delta g = h_{\mu\nu}\, dx^\mu \otimes dx^\nu$, of the metric such that $g + \delta g$ is also a solution of the Einstein equations to first order in δg. The components $h_{\mu\nu}$ are thus restricted by a set of ten coupled second-order *linear* partial differential equations.

Various methods have been applied to find the gravitational perturbations for some specific curved background metrics. In particular, the perturbations of the Kerr metric have been studied in detail, leading to the discovery of some interesting physical effects and some remarkable mathematical properties of the solutions. (See, e.g., Ref. 1.) The complete gravitational perturbations, in the case of the Kerr metric, were obtained by Chandrasekhar[2] by solving the linearized equations and by Chrzanowski,[3] who, by postulating a certain factorized form for the Green's function of the gravitational perturbations, obtained the result that the metric perturbations can be expressed in terms of second derivatives of a scalar (Debye) potential.

By analogy with the corresponding expressions for the electromagnetic perturbations, Chrzanowski's formulas for the metric perturbations were generalized by himself and by Cohen and Kegeles[4] to all vacuum algebraically special space-times. The proof that such expressions are indeed solutions of the linearized Einstein equations was given by Kegeles and Cohen[5] by a direct substitution that, even in the spinor formalism, involves a very lengthy computation. An equivalent result was obtained by Wald[6] by a very simple derivation based on a more general theorem. The metric perturbations found by these authors are given in terms of the second derivatives of a complex scalar potential that satisfies a second-order linear partial differential equation.

The purpose of this paper is to present a derivation of the expressions for the metric perturbations mentioned above, including the presence of the cosmological constant, by using the description of the algebraically special vacuum space-times based on the complex two-dimensional totally null foliation that the complex extensions of these space-times possess. This last formulation has been developed with the objective of finding the algebraically special solutions of the (nonlinear) Einstein equations for a vacuum or coupled to a suitably restricted matter field[7–11]; however, it is also very useful for the integration of other field equations on this class of space-times, e.g., those of the Killing vector fields,[12] Killing spinors,[13] massless spinor fields,[9,10,14] and Yang–Mills fields.[11] In the derivation given here we make use of the result that the Einstein field equations reduce, in the case of an algebraically special vacuum space-time with cosmological constant, to a single nonlinear second-order partial differential equation for a scalar function—the so-called hyperheavenly, or $\mathscr{H}\mathscr{H}$, equation—[Eqs. (3) and (5) below]. The similarity between the $\mathscr{H}\mathscr{H}$ equation and the equation for the gravitational Debye potential given in Refs. 3–5 was pointed out by McIntosh and Hickman.[15]

II. THE $\mathscr{H}\mathscr{H}$ EQUATION

If g denotes the metric of an algebraically special solution of the Einstein vacuum field equations with cosmological constant λ, then there exist, locally, complex coordinates $q^{\dot A}, p^{\dot A}$ $(\dot A = \dot 1, \dot 2)$ such that[7,10]

$$g = 2\phi^{-2}\, dq^{\dot A} \underset{s}{\otimes} (dp_{\dot A} + Q_{\dot A \dot B}\, dq^{\dot B})\,, \qquad (1)$$

where ϕ and $Q_{\dot A \dot B}$ are complex-valued functions, with $Q_{\dot A \dot B} = Q_{\dot B \dot A}$, and the indices are raised and lowered according to $\psi_{\dot A} = \epsilon_{\dot A \dot B}\psi^{\dot B}$, $\psi^{\dot B} = \epsilon^{\dot A \dot B}\psi_{\dot A}$. The complex two-dimensional surfaces, given by $q^{\dot A} = \text{const}$, constitute the foliation mentioned previously.

Einstein's field equations imply that $\phi = J_{\dot A} p^{\dot A} + \kappa$, where $J_{\dot A}$ and κ depend on $q^{\dot B}$ only. By choosing a set of coordinates $q^{\dot A}, p^{\dot A}$ such that $J_{\dot A}$ and κ are constant, it follows that, if $J_{\dot A} = 0$,

$$Q_{\dot A \dot B} = -\partial_{\dot A}\, \partial_{\dot B}\Theta - \tfrac{2}{3}\kappa^2 L_{(\dot A} p_{\dot B)} + (\lambda/3)\kappa^{-2} p_{\dot A} p_{\dot B}\,, \qquad (2)$$

where $\partial_{\dot A} = \partial/\partial p^{\dot A}$, $L_{\dot A} = L_{\dot A}(q^{\dot B})$, and Θ is a solution of

0022-2488/86/061586-03$02.50

$$\frac{1}{2}(\partial^{\dot{A}}\partial^{\dot{B}}\Theta)\partial_{\dot{A}}\,\partial_{\dot{B}}\Theta-\partial^{\dot{A}}\frac{\partial\Theta}{\partial q^{\dot{A}}}-\kappa^2L^{\dot{A}}\,\partial_{\dot{A}}\Theta+\frac{1}{18}(\kappa^2L_{\dot{A}}p^{\dot{A}})^2+\frac{2}{3}\kappa^2L^{\dot{A}}p^{\dot{B}}\,\partial_{\dot{A}}\,\partial_{\dot{B}}\Theta$$
$$-\frac{1}{6}\kappa^2\frac{\partial L_{\dot{A}}}{\partial q^{\dot{B}}}p^{\dot{A}}p^{\dot{B}}-\left(\frac{\lambda}{3}\right)\kappa^{-2}p^{\dot{A}}p^{\dot{B}}\,\partial_{\dot{A}}\,\partial_{\dot{B}}\,\Theta+\lambda\kappa^{-2}(p^{\dot{A}}\partial_{\dot{A}}\Theta-\Theta)=\kappa^2(N_{\dot{A}}p^{\dot{A}}+\gamma)\,, \tag{3}$$

where $N_{\dot{A}}$ and γ are functions of $q^{\dot{B}}$ only, and, in the case where $J_{\dot{A}}\neq 0$,

$$Q_{\dot{A}\dot{B}}=-\phi^3\,\partial_{(\dot{A}}\phi^{-2}\partial_{\dot{B})}\,W+(\mu\phi^3+(\lambda/6))K_{\dot{A}}K_{\dot{B}}\,, \tag{4}$$

where $K_{\dot{A}}$ is a constant spinor such that $K^{\dot{A}}J_{\dot{A}}=1, \mu=\mu(q^{\dot{B}})$, and W is a solution of

$$\frac{1}{2}\phi^4(\partial^{\dot{A}}\phi^{-2}\,\partial^{\dot{B}}W)\partial_{\dot{A}}\phi^{-2}\,\partial_{\dot{B}}W-\phi^{-1}\partial^{\dot{A}}\frac{\partial W}{\partial q^{\dot{A}}}-\mu\phi^4K^{\dot{A}}\,\partial_{\dot{A}}\phi^{-1}K^{\dot{B}}\,\partial_{\dot{B}}\phi^{-1}W-\left(\frac{\lambda}{6}\right)\phi^{-1}K^{\dot{A}}\,\partial_{\dot{A}}K^{\dot{B}}\,\partial_{\dot{B}}W$$
$$+\frac{1}{2}K^{\dot{A}}p_{\dot{A}}\,[K^{\dot{B}}p_{\dot{B}}J^{\dot{C}}-(\phi+\kappa)K^{\dot{C}}]\,\frac{\partial\mu}{\partial q^{\dot{C}}}=N_{\dot{A}}p^{\dot{A}}+\gamma\,. \tag{5}$$

With respect to the null tetrad

$$\partial_{A\dot{B}}=\sqrt{2}\left\{\delta^1_A\,\partial_{\dot{B}}+\phi^2\delta^2_A\left(\frac{\partial}{\partial q^{\dot{B}}}+Q_{\dot{B}\dot{C}}\,\partial^{\dot{C}}\right)\right\}, \tag{6}$$

the spinor components of the Weyl tensor satisfy $C_{1111}=C_{1112}=0$ (which corresponds to the assumed algebraic degeneracy of the curvature) and, in the case where $J_{\dot{A}}=0$,

$$C^{(3)}\equiv 2C_{1122}=-\tfrac{2}{3}\lambda\,, \tag{7a}$$

while in the case where $J_{\dot{A}}\neq 0$,

$$C^{(3)}=-2\mu\phi^3\,. \tag{7b}$$

The dotted components are given in the case $J_{\dot{A}}=0$ by

$$C_{\dot{A}\dot{B}\dot{C}\dot{D}}=\kappa^2\,\partial_{\dot{A}}\,\partial_{\dot{B}}\,\partial_{\dot{C}}\,\partial_{\dot{D}}\,\Theta\,, \tag{8a}$$

and in the case $J_{\dot{A}}\neq 0$ by

$$C_{\dot{A}\dot{B}\dot{C}\dot{D}}=\phi^3\,\partial_{\dot{A}}\,\partial_{\dot{B}}\,\partial_{\dot{C}}\,\partial_{\dot{D}}\,W-6\mu\phi^3J_{(\dot{A}}J_{\dot{B}}K_{\dot{C}}K_{\dot{D})}\,. \tag{8b}$$

If Θ or W is a solution of Eq. (3) or (5), respectively, corresponding to a real solution of Einstein's equations, then $C_{\dot{A}\dot{B}\dot{C}\dot{D}}$ is of the same algebraic type as C_{ABCD} and hence algebraically special; however, not every solution of Eq. (3) or (5) corresponds to a real metric (with Lorentzian signature) and, in general, $C_{\dot{A}\dot{B}\dot{C}\dot{D}}$ given by Eqs. (8) will not be algebraically special. Actually, Eqs. (1)–(5) follow from the (complex) Einstein vacuum field equations assuming the algebraic degeneracy of C_{ABCD} only, with $C_{\dot{A}\dot{B}\dot{C}\dot{D}}$ arbitrary. This point is essential in what follows.

III. GRAVITATIONAL PERTURBATIONS

Assuming now that Θ or W is a solution of Eq. (3) or (5), respectively, corresponding to a given algebraically special solution of Einstein's equations and that $\Theta+\delta\Theta$ or $W+\delta W$ also satisfies, to first order in $\delta\Theta$ or δW, Eq. (3) or (5), respectively, then, using Eqs. (2), (4), and (7), a straightforward computation gives

$$\left[\frac{\partial}{\partial q^{\dot{A}}}+Q_{\dot{A}\dot{B}}\,\partial^{\dot{B}}-(\partial^{\dot{B}}\,Q_{\dot{A}\dot{B}})\right]\partial^{\dot{A}}\chi-\frac{3}{2}\phi^{-2}C^{(3)}\chi=0\,, \tag{9}$$

where $\chi=\kappa^{-1}\delta\Theta$ in the case where $J_{\dot{A}}=0$ and $\chi=\delta W$ when $J_{\dot{A}}\neq 0$.

From Eqs. (1), (2), and (4) it follows that the metric perturbation δg generated by a solution of (9) is given by

$$\delta g=-2\phi\,\partial_{(\dot{A}}\,\phi^{-2}\,\partial_{\dot{B})}\,\chi\,dq^{\dot{A}}\otimes dq^{\dot{B}}\,. \tag{10}$$

This metric perturbation will be complex in general. In fact, according to Eqs. (8) the perturbation of the dotted spinor components of the Weyl tensor is

$$\delta C_{\dot{A}\dot{B}\dot{C}\dot{D}}=\phi^3\,\partial_{\dot{A}}\,\partial_{\dot{B}}\,\partial_{\dot{C}}\,\partial_{\dot{D}}\,\chi\,, \tag{11}$$

while, in the case $J_{\dot{A}}=0$, $\delta C_{ABCD}=0$ and if $J_{\dot{A}}\neq 0$, only δC_{2222} can be different from zero. However, since we are dealing with linearized equations, the real and the imaginary parts of δg are real solutions of the linearized Einstein equations and the perturbation of the Weyl tensor corresponding to them is not necessarily algebraically special.

It turns out that Eqs. (9)–(11) are valid in any coordinate system $q^{\dot{A}}, p^{\dot{A}}$ in which the metric has the form (1), even though $J_{\dot{A}}$ and κ may not be constant. This can be readily verified by checking that Eqs. (9)–(11) are form-invariant under any coordinate transformation that maintains the metric (1) form-invariant.

By using the procedure given in Ref. 14, the expressions derived above can be written in a covariant way, which gives for the (real) metric perturbations

$$h_{CD\dot{R}\dot{S}}=-2\nabla^{A}{}_{(\dot{R}}\phi^{-4}\nabla^{B}{}_{\dot{S})}\phi^4\pi_{ABCD}+\text{H.c.}\,, \tag{12}$$

where the covariant derivatives are with respect to the background connection and π_{ABCD} is a null Hertz potential

$$\pi_{ABCD}=\psi l_A l_B l_C l_D\,, \tag{13}$$

where l_A denotes the multiple Debever–Penrose spinor of the background curvature and ψ is a complex function that takes the place of χ. (In the notation of Ref. 14, $\psi=\phi^{-5}\Delta^{-2}\chi$.) Equation (9) then amounts to

$$\nabla_{\dot{R}(A}\phi^{-4}\nabla^{S\dot{R}}\phi^4\pi_{BCD)S}-6C_{(AB}{}^{RS}\pi_{CD)RS}=0\,. \tag{14}$$

Equations (12) and (14) are equivalent to those postulated in Ref. 5 for the case of vacuum, taking into account the fact that[10]

$$l^A\nabla_{B\dot{C}}l_A=l_Bl^A\,\partial_{A\dot{C}}\ln\phi\,. \tag{15}$$

Written in terms of the Newman–Penrose spin coefficients, using (13) and (15), Eqs. (12) and (14) correspond to the expressions given in Refs. 3–6.

It may be pointed out that in all type D vacuum backgrounds with cosmological constant, Eq. (14) admits separable solutions.[16]

IV. CONCLUSION

The derivation outlined above exemplifies the power of the approach to the dynamic equations of general relativity based on the existence of complex two-dimensional null foliations. It is noteworthy that the complex scalar potential for the gravitational perturbations found previously by other authors, through quite different approaches, corresponds precisely to the first-order variation of the key function, Θ or W, governed by the $\mathscr{H}\mathscr{H}$ equation.

It should be remarked that the arbitrariness in the dotted components of the Weyl spinor, corresponding to a solution of the $\mathscr{H}\mathscr{H}$ equation, allows the possibility of having an algebraically general perturbed metric.

By a procedure similar to that followed here, one can consider the simultaneous perturbations of the gravitational field and of a coupled electromagnetic, neutrino or Yang–Mills field, provided that the (complex extension of the) background space-time admits a complex two-dimensional null foliation and that the background matter field is suitably aligned to that foliation, by using the reduced form of the Einstein equations coupled to one of these fields.[9–11] As in the case of vacuum, the perturbed metric can be algebraically general and the perturbed matter field will not be necessarily aligned to the foliation of the background space-time. These cases will be treated elsewhere.

ACKNOWLEDGMENTS

The author is pleased to thank Professor J. F. Plebánski and Claudia Arámburo for their interest in this work.

The author also acknowledges support from the Sistema Nacional de Investigadores (Mexico).

[1] S. Chandrasekhar, in *General Relativity: An Einstein Centenary Survey*, edited by S. W. Hawking and W. Israel (Cambridge U. P., Cambridge, England, 1979).

[2] S. Chandrasekhar, Proc. R. Soc. London Ser. A **358**, 421, 441 (1978).

[3] P. L. Chrzanowski, Phys. Rev. D **11**, 2042 (1975).

[4] J. M. Cohen and L. S. Kegeles, Phys. Lett. A **54**, 5 (1975).

[5] L. S. Kegeles and J. M. Cohen, Phys. Rev. D **19**, 1641 (1979).

[6] R. M. Wald, Phys. Rev. Lett. **41**, 203 (1978).

[7] J. F. Plebański and I. Robinson, Phys. Rev. Lett. **37**, 493 (1976). The details can be found in *Asymptotic Structure of Space-Time*, edited by F. P. Esposito and L. Witten (Plenum, New York, 1977).

[8] M. S. Hickman and C. B. G. McIntosh, Gen. Relativ. Gravit. **18**, 107 (1986).

[9] A. García, J. F. Plebański, and I. Robinson, Gen. Relativ. Gravit. **8**, 841 (1977); J. D. Finley, III and J. F. Plebański, J. Math. Phys. **18**, 1662 (1977).

[10] G. F. Torres del Castillo, J. Math. Phys. **24**, 590 (1983).

[11] G. F. Torres del Castillo, J. Math. Phys. **26**, 836 (1985).

[12] J. D. Finley, III and J. F. Plebański, J. Math. Phys. **19**, 760 (1978); S. A. Sonnleitner and J. D. Finley, III, J. Math. Phys. **23**, 116 (1982); G. F. Torres del Castillo, J. Math. Phys. **25**, 1980 (1984).

[13] G. F. Torres del Castillo, Commun. Math. Phys. **92**, 485 (1984).

[14] G. F. Torres del Castillo, J. Math. Phys. **25**, 342 (1984).

[15] C. B. G. McIntosh and M. S. Hickman, Gen. Relativ. Gravit. **17**, 111 (1985).

[16] G. F. Torres del Castillo, J. Math. Phys. **27**, 1583 (1986).

Parallel-propagated frame along the geodesics of the metrics admitting a Killing–Yano tensor

Niky Kamran
Centre de Recherches Mathematiques, Université de Montréal, Case Postale 6128, Succ. A, Montréal, Quebec, Canada H3C3J7

Jean-Alain Marck
Groupe d'Astrophysique Relativiste, Observatoire de Paris, F-92195 Meudon Cedex, France

(Received 8 October 1985; accepted for publication 30 January 1986)

It is shown that the equations for a parallel-propagated frame along geodesics can be solved explicitly by separation of variables assuming the existence of a valence-2 Killing–Yano tensor that is indecomposable and such that the associated Killing tensor has no constant eigenvalue.

I. INTRODUCTION

In two previous papers,[1,2] it has been shown by one of us that the separability properties of the Hamilton–Jacobi equation for the geodesics of the Kerr solution and the existence of a valence-2 Killing–Yano tensor therein allow the explicit construction through separation of variables of a parallel-propagated frame along an arbitrary geodesic in Kerr geometry. This result proved to be useful in the study of tidal effects near a black hole[3,4] as it allowed the explicit computation of the tidal tensor. On the other hand, it is known[5] that the existence of a valence-2 Killing–Yano tensor $(f_{\mu\nu})$ that is indecomposable and such that the associated Killing tensor $(K_{\mu\nu}:=f_{\mu\rho}f_{\nu}{}^{\rho})$ has no constant eigenvalue is enough by itself to ensure the existence of separable coordinates for the Hamilton–Jacobi equation.

One is thus naturally led to ask the question of whether the method used for Kerr could be adapted to construct a parallel-propagated frame under the *only* assumption of the existence of a Killing–Yano tensor having the above properties, and in particular with no reference to any field equations. In the present paper, we show that this is indeed the case, and in particular we extend the results of Refs. 1 and 2 to all the "nonaccelerating" Petrov type D vacuum solutions, that is, the Carter solutions[6] and the Debever–McLenaghan null orbit solution.[7]

In Sec. II we briefly recall some known facts about the metrics admitting a Killing–Yano tensor $(f_{\mu\nu})$ having the above properties. In Sec. III we construct our parallel-propagated frame following a procedure inspired by that of Ref. 1.

Throughout this paper, all indices run from 0 to 3. We denote the components of any tensor with respect to the coordinate basis with Greek indices and with respect to an orthonormal frame with Latin indices in round brackets.

II. THE METRICS

By a valence-2 Killing–Yano tensor, we mean a skew-symmetric tensor $(f_{\mu\nu})$ satisfying the equation[8]

$$\nabla_\rho f_{\mu\nu}+\nabla_\nu f_{\mu\rho}=0. \tag{2.1}$$

It follows from Eq. (2.1) that, if (u^α) is the unit tangent vector to a geodesic C, then the vector (L^α) defined by

$$L^\alpha=f^\alpha{}_\beta u^\beta \tag{2.2a}$$

is parallel-propagated along C, that is

$$u^\alpha\nabla_\alpha L^\beta=0. \tag{2.2b}$$

We shall consider only the case of valence-2 Killing–Yano tensors $(f_{\mu\nu})$ that are *indecomposable*, in other words such that the two-form $\mathscr{F}:=\frac{1}{2}f_{\mu\nu}\,dx^\mu\wedge dx^\nu$ satisfies

$$\mathscr{F}\wedge\mathscr{F}\neq 0, \tag{2.3}$$

and such the associated Killing tensor $(K_{\mu\nu}:=f_{\mu\rho}f_\nu{}^\rho)$ has *no constant eigenvalue*. Indeed, for the metrics with a Killing–Yano tensor possessing a constant eigenvalue, the existence of more than one Killing vector never follows automatically from that of the Killing–Yano tensor.[9] The existence of separable coordinates for the Hamilton–Jacobi equation is thus not ensured without further assumptions. We then have the following result.[5,9]

Theorem: For every metric admitting a valence-2 indecomposable real-valued Killing–Yano tensor $(f_{\mu\nu})$ such that the associated Killing tensor $(K_{\mu\nu}:=f_{\mu\rho}f_\nu{}^\rho)$ has no constant eigenvalue, there exist local coordinates $(x^\mu)=(u,v,w,x)$ and an orthonormal frame $(\omega^{(a)})=(\omega^{(0)},\omega^{(1)},\omega^{(2)},\omega^{(3)})$ in which

$$ds^2=\eta_{(a)(b)}\,\omega^{(a)}\omega^{(b)}, \tag{2.4a}$$

$$\tfrac{1}{2}(f_{\mu\nu}-i{}^*f_{\mu\nu})dx^\mu\wedge dx^\nu=\alpha_0(x\omega^{(1)}\wedge\omega^{(0)}+w\omega^{(3)}\wedge\omega^{(2)}), \tag{2.4b}$$

where

$$\omega^{(0)}:=\frac{1}{2}Z^{1/2}\left\{(1+f)\frac{W}{Z}(du-x^2\,dv)+\frac{2}{1+f^2}(1-f)\frac{dw}{W}\right\}, \tag{2.5a}$$

$$\omega^{(1)}:=\frac{1}{2}Z^{1/2}\left\{(f-1)\frac{W}{Z}(du-x^2\,dv)+\frac{2}{1+f^2}(1+f)\frac{dw}{W}\right\}, \tag{2.5b}$$

$$\omega^{(2)}:=-(Z^{1/2}/X)dx,$$
$$\omega^{(3)}:=(X/Z^{1/2})(du+w^2\,dv), \tag{2.5c}$$

$$Z:=w^2+x^2,\quad \eta_{(a)(b)}:=\mathrm{diag}(-1,+1,+1,+1), \tag{2.5d}$$

where W and X are arbitrary functions of w and x, respectively, and where α_0 and f are arbitrary real constants.

0022-2488/86/061589-03$02.50

The metric given by Eqs. (2.4) and (2.5) manifestly possesses a two-parameter Abelian group of isometries that acts on non-null or null orbits according to whether $f \neq 0$ or $f = 0$. In the former case, the isometry group is invertible and hence orthogonally transitive,[10] while in the latter case it is orthogonally transitive but not invertible. We make the important observation that when the Einstein vacuum field equations are imposed on the metrics of the above theorem, one obtains[11] Carter's $[A]$ solutions, which contain the Kerr solution as a special case, and the null orbit solution of Debever and McLenaghan. Also, the orthonormal frame $(\boldsymbol{\omega}^{(a)})$ is identical when $f^2 = 1$ to Carter's *symmetric frame*,[6] which has proven itself to be the simplest for treating the separability properties of these metrics.[12]

III. THE PARALLEL-PROPAGATED FRAME

In this section, we shall construct a (locally defined) orthonormal frame $(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3)$, which is parallel-propagated along an arbitrary timelike geodesic C of the metric given by Eqs. (2.4) and (2.5). The procedure we follow is modeled on the one used for the Kerr solution in Ref. 1.

We choose $\boldsymbol{\lambda}_0$ to be the unit tangent vector to C, which is obviously parallel-propagated. Using the separability property of the Hamilton–Jacobi equation for the geodesics, we easily obtain that

$$\lambda_0{}^{(0)} = \frac{1+f}{1+f^2}\frac{1}{WZ^{1/2}}(\beta - w^2\alpha) - \left(\frac{1-f}{2}\right)\frac{W}{Z^{1/2}}p_w, \tag{3.1a}$$

$$\lambda_0{}^{(1)} = \frac{1-f}{1+f^2}\frac{1}{WZ^{1/2}}(\beta - w^2\alpha) + \left(\frac{1+f}{2}\right)\frac{W}{Z^{1/2}}p_w, \tag{3.1b}$$

$$\lambda_0{}^{(2)} = -(X/Z^{1,2})p_x, \tag{3.1c}$$

$$\lambda_0{}^{(3)} = (1/Z^{1/2}X)(\beta + x^2\alpha), \tag{3.1d}$$

where the p_μ are the momenta canonically conjugate to the velocities $\dot{x}^\mu$. They are determined by the following relations:

$$X^2p_x^2 = K - x^2 - (1/X^2)(\beta + x^2\alpha)^2, \tag{3.2a}$$

$$fW^2p_w^2 + 2\left(\frac{1-f^2}{1+f^2}\right)(\beta - w^2\alpha)p_w = -K - w^2 + \left(\frac{2}{1+f^2}\right)^2 f\frac{1}{W^2}(\beta - w^2\alpha)^2, \tag{3.2b}$$

$$p_u = \alpha, \tag{3.2c}$$

$$p_v = \beta, \tag{3.2d}$$

where α, β, and K are separation constants. It should be noted that K can be expressed as

$$K = K_{\mu\nu}\lambda_0{}^\mu\lambda_0{}^\nu, \tag{3.2e}$$

where $(K_{\mu\nu})$ is the Killing tensor associated to the Killing–Yano tensor $(f_{\mu\nu})$ given by Eq. (2.4b), and reduces to Carter's "fourth constant of motion" in the special case of the Kerr solution. Now, from Eqs. (2.2a), (2.2b), and (2.4b), where we have chosen for convenience $\alpha_0 = 1$, we know that the unit vector $\boldsymbol{\lambda}_2$ defined by

$$\lambda_2{}^\mu := (1/K^{1/2})f^\mu{}_\nu\lambda_0{}^\nu, \tag{3.3a}$$

whose components in the symmetric frame $(\boldsymbol{\omega}^{(a)})$ are explicitly given by

$$\lambda_2{}^{(0)} = (1/K^{1/2})x\lambda_0{}^{(1)}, \tag{3.3b}$$

$$\lambda_2{}^{(1)} = (1/K^{1/2})x\lambda_0{}^{(0)}, \tag{3.3c}$$

$$\lambda_2{}^{(2)} = -(1/K^{1/2})w\lambda_0{}^{(3)}, \tag{3.3d}$$

$$\lambda_2{}^{(3)} = (1/K^{1/2})w\lambda_0{}^{(2)}, \tag{3.3e}$$

is parallel-propagated along C and orthogonal to $\boldsymbol{\lambda}_0$:

$$\lambda_0{}^\mu\nabla_\mu\lambda_2{}^\nu = 0, \quad \lambda_{0\mu}\lambda_2{}^\mu = 0. \tag{3.3f}$$

To obtain the remaining vectors $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_3$ of our parallel-propagated frame, we make a natural choice of two vectors $\tilde{\boldsymbol{\lambda}}_1$ and $\tilde{\boldsymbol{\lambda}}_3$, which form an orthonormal frame when taken with $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_2$. They are defined by

$$\tilde{\lambda}_3{}^{(0)} = \sigma\lambda_0{}^{(0)}, \quad \tilde{\lambda}_1{}^{(0)} = (1/K^{1/2})\sigma w\lambda_0^{(1)}, \tag{3.4a}$$

$$\tilde{\lambda}_3{}^{(1)} = \sigma\lambda_0{}^{(1)}, \quad \tilde{\lambda}_1{}^{(1)} = (1/K^{1/2})\sigma w\lambda_0{}^{(0)}, \tag{3.4b}$$

$$\tilde{\lambda}_3{}^{(2)} = (1/\sigma)\lambda_0{}^{(2)}, \quad \tilde{\lambda}_1{}^{(2)} = (1/K^{1/2})(x/\sigma)\lambda_0{}^{(3)}, \tag{3.4c}$$

$$\tilde{\lambda}_3{}^{(3)} = (1/\sigma)\lambda_0{}^{(3)}, \quad \tilde{\lambda}_1{}^{(3)} = -(1/K^{1/2})(x/\sigma)\lambda_0{}^{(2)}, \tag{3.4d}$$

where

$$\sigma := ((K - x^2)/(K + w^2))^{1/2}. \tag{3.4e}$$

Now, as was the case for the Kerr solution, the vectors $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_3$ are obtained by performing an appropriate spatial rotation on $\tilde{\boldsymbol{\lambda}}_1$ and $\tilde{\boldsymbol{\lambda}}_3$, which can be expressed in term of a parameter Ψ by

$$\boldsymbol{\lambda}_1 = \tilde{\boldsymbol{\lambda}}_1 \cos\Psi - \tilde{\boldsymbol{\lambda}}_3 \sin\Psi, \tag{3.5a}$$

$$\boldsymbol{\lambda}_3 = \tilde{\boldsymbol{\lambda}}_1 \sin\Psi + \tilde{\boldsymbol{\lambda}}_3 \cos\Psi. \tag{3.5b}$$

The conditions for parallel transport of $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_3$ along C,

$$\lambda_0{}^\mu\nabla_\mu\lambda_1{}^\nu = 0, \quad \lambda_0{}^\mu\nabla_\mu\lambda_3{}^\nu = 0, \tag{3.6a}$$

are now equivalent to

$$\frac{d\Psi}{d\tau} = \frac{K^{1/2}}{Z}\left\{\frac{\beta + x^2\alpha}{K - x^2} + \frac{w^2\alpha - \beta}{K + w^2}\right\}, \tag{3.6b}$$

where τ denotes an affine parameter associated to the unit tangent vector $\boldsymbol{\lambda}_0$. It then follows using Eqs. (2.5), (3.1), and (3.2) that Ψ takes the separable form

$$\Psi(w,x) = F(w) + G(x), \tag{3.7a}$$

where

$$F(w) := K^{1/2}\int\frac{w^2\alpha - \beta}{K + w^2}\frac{dw}{\sqrt{(\beta - w^2\alpha)^2 - fW^2(K + w^2)}}, \tag{3.7b}$$

$$G(x) := K^{1/2}\int\frac{\beta + x^2\alpha}{K - x^2}\frac{dx}{\sqrt{X^2(K - x^2) - (\alpha x^2 + \beta)^2}}. \tag{3.7c}$$

It is easily checked that the expression for Ψ given by Eqs. (3.7) reduces to the one given in Ref. 1 for the Kerr solution.

Finally, it should be noted that the above method would fail to produce a parallel-propagated frame along the null geodesics of the metrics admitting a *conformal* valence-2 Killing–Yano tensor, that is a skew-symetric tensor $(D_{\mu\nu})$, which satisfies

$$\nabla_\gamma D_{\alpha\beta} + \nabla_\beta D_{\alpha\gamma} = -\tfrac{2}{3}\nabla_\delta D^\delta{}_\alpha g_{\beta\gamma} + \tfrac{1}{3}\nabla_\delta D^\delta{}_\beta g_{\gamma\alpha} + \tfrac{1}{3}\nabla_\delta D^\delta{}_\gamma g_{\beta\alpha}. \tag{3.8}$$

Indeed, it has been shown by Jeffryes[13] that there exist metrics admitting a valence-2 indecomposable conformal Killing–Yano tensor ($D_{\mu\nu}$), such that the associated conformal Killing tensor ($B_{\mu\nu} = D_{\mu\rho}D_{\nu}{}^{\rho}$) has no constant eigenvalue, that are conformally equivalent to a metric of the form given by Eqs. (2.4) and (2.5) with a conformal factor that is independent of u and v. (Although it was explicitly shown that there exist metrics admitting conformal valence-2 Killing–Yano tensors that are not conformal to ones admitting Killing–Yano tensors.) On the other hand, it is well known that the equations of parallel transport along a geodesic are not conformally invariant, even in the case of null geodesics.

IV. CONCLUSION

We have shown that the equations for a parallel-propagated frame can be solved explicitly by separation of variables assuming only the existence of a valence-2 Killing–Yano tensor that is indecomposable and such that the associated Killing tensor has no constant eigenvalue. The result obtained above for timelike geodesics can be easily extended to spacelike geodesics and null geodesics using the method of Ref. 2.

ACKNOWLEDGMENTS

It is a pleasure to thank B. Carter and R. G. McLenaghan for helpful discussions and encouragements.

One of us (N.K.) thanks the members of the "Group d'Astrophysique Relativiste" for their kind hospitality and gratefully acknowledges financial support through a postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada.

[1] J. A. Marck, Proc. R. Soc. London Ser. A **385**, 431 (1983).
[2] J. A. Marck, Phys. Lett A **97**, 140 (1983).
[3] J. P. Luminet and J. A. Marck, Mon. Not. R. Astron. Soc. **212**, 57 (1985).
[4] J. P. Luminet and J. A. Marck, in *Proceedings of the General Relativity and Gravitation Meeting GR10*, edited by B. Bertotti and F. de Felice (North-Holland, Amsterdam, 1983).
[5] N. Kamran and R. G. McLenaghan, Bull. Cl. Sci. Acad. R. Belg. **LXX**, 596 (1984).
[6] B. Carter, Commun. Math. Phys. **10**, 280 (1968).
[7] R. Debever and R. G. McLenaghan, J. Math. Phys. **22**, 1711 (1981).
[8] R. Penrose, Ann. N. Y. Acad. Sci. **224**, 125 (1973).
[9] W. Dietz and R. Rudiger, Proc. R. Soc. London Ser. A **375**, 361 (1981).
[10] B. Carter, in *Black Holes*, edited by B. De Witt and C. De Witt (Gordon and Breach, New York, 1973).
[11] R. Debever, N. Kamran, and R. G. McLenaghan, J. Math. Phys. **25**, 1955 (1984).
[12] N. Kamran and R. G. McLenaghan, J. Math. Phys. **25**, 1019 (1984).
[13] B. Jeffryes, Proc. R. Soc. London Ser. A **392**, 323 (1984).

A method for generating exact Bianchi type II cosmological models

J. Hajj-Boutros
Department of Physics, Lebanese University, Mansourieh, P. O. Box 72, Lebanon

(Received 25 February 1985; accepted for publication 3 January 1986)

A method for generating exact Bianchi type II cosmological models with a perfect fluid distribution of matter is presented. Two new classes of Bianchi type II solutions have been generated from Lorenz's solution [D. Lorenz, Phys. Lett. A **79**, 19 (1980)]. A detailed study of physical and kinematic properties of one of them has been carried out.

I. INTRODUCTION

In cosmology, the Friedmann–Robertson–Walker (FRW) models play an essential role. It is not believed that these models truly represent the universe, but in some sense they are good global approximations of the present universe. FRW models are characterized by (i) the universe being the same at all points in space (spatial homogeneity) and (ii) all spatial directions at a point being equivalent (isotropy).

In recent years, experimental studies of the isotropy of the cosmic microwave radiation and speculation about the amount of helium formed at early stages and many other effects have stimulated theoretical interest in anisotropic cosmological models.

The spatially homogeneous and anisotropic Bianchi models present a medium way between FRW models and completely inhomogeneous and anisotropic universes and thus play an important role in current modern cosmology. A spatially homogeneous Bianchi model necessarily has a three-dimensional group, which acts simply transitively on spacelike three-dimensional orbits.

Here we confine ourselves to a locally rotationally symmetric (LRS) model of Bianchi type II. This model is characterized by three metric functions $R_1(t)$, $R_2(t)$, and $R_3(t)$ such that $R_1 = R_2 \neq R_3$. The metric functions are functions of time only. (For non-LRS Bianchi metrics we have $R_1 \neq R_2 \neq R_3$.) For LRS Bianchi type II metric, Einstein's field equations reduce, in the case of perfect fluid distribution of matter, to three nonlinear differential equations.

If we restrict ourselves to a barytropic equation of state such as

$$P = \rho \qquad \text{(stiff matter)},$$

we are able to reduce the field equations to a Riccati equation, and by the same procedure as that used in Refs. 1 and 2, we generate several new exact solutions of Bianchi type II; the known solution used here is that given by Lorenz.[3]

The geometric and kinematic properties of one of them has been studied in some detail. The nature of singularity has been clarified. It then appears that the new solutions have a barrel singularity.

II. FIELD EQUATION AND GENERATION TECHNIQUE

In an orthonormal frame, the metric for Bianchi type II in the LRS case is given by[3]

$$ds^2 = \eta_{ij}\sigma^i\sigma^j, \qquad \eta_{ij} = \text{diag}(-1, 1, 1, 1), \tag{2.1}$$

where the Cartan bases σ^i are given by

$$\sigma^0 = dt, \qquad \sigma^1 = S(t)\omega^1, \qquad \sigma^2 = R(t)\omega^2, \qquad \sigma^3 = R(t)\omega^3, \tag{2.2}$$

where $R(t)$ and $S(t)$ are the metric functions. The time independent differential one-forms ω^i are given by

$$\omega^1 = dy + x\,dz, \qquad \omega^2 = dz, \qquad \text{and} \qquad \omega^3 = dz\,. \tag{2.3}$$

The field equations, in the case of perfect fluid, are

$$G_{ij} = 8\pi T_{ij}\,, \tag{2.4}$$

with

$$T_{ij} = (\rho + p)u_i u_j + \eta_{ij}P, \tag{2.5}$$

G_{ij} being the Einstein tensor and ρ and p being, respectively, the energy density and the pressure of the fluid, read

$$2\frac{\dot R}{R}\frac{\dot S}{S} + \frac{\dot R^2}{R^2} - \frac{1}{4}\frac{S^2}{R^4} = 8\pi\rho\,, \tag{2.6}$$

$$2\frac{\ddot R}{R} + \frac{\dot R^2}{R^2} - \frac{3}{4}\frac{R^2}{S^4} = -8\pi p\,, \tag{2.7}$$

$$\frac{\ddot S}{S} + \frac{\ddot R}{R} + \frac{\dot S}{S}\frac{\dot R}{R} + \frac{1}{4}\frac{S^2}{R^4} = -8\pi p\,, \tag{2.8}$$

where "$\cdot$" represents derivation with respect to t. We consider here the special case of the stiff matter; Eq. (2.6) and (2.8) imply

$$\frac{\ddot S}{S} + \frac{\ddot R}{R} + 3\frac{\dot S}{S}\frac{\dot R}{R} + \frac{\dot R^2}{R^2} = 0\,. \tag{2.9}$$

By performing the substitution

$$r = \dot R/R\,, \tag{2.10}$$

Eq. (2.9) becomes

$$\dot r + 2r^2 + 3r(\dot S/S) + \ddot S/S = 0\,, \tag{2.11}$$

which is a Riccati equation in r; it can be linearized by means of change of function

$$r = r_0 + 1/B\,. \tag{2.12}$$

So we obtain

$$\dot B + B(-3\dot S/S - 4r_0) = 2\,, \tag{2.13}$$

r_0 being a particular solution for (2.11).

Equation (2.13) is a linear first-order differential equation. The solution for (2.13) is then

0022-2488/86/061592-03$02.50

$$R(t) = R_0(t)[J(t, C)]^{1/2}/C_1 , \tag{2.14}$$

where

$$J(t, C) = \int \frac{2}{S^3}\frac{dt}{R^4} + C , \tag{2.15}$$

and C and C_1 are two arbitrary constants.

Note that Eq. (2.9) can be considered as a Riccati equation in $\dot S/S$ and, by a similar method to that given in the case (2.11), we obtain the solution

$$S(t) = S_0(t) \times [M(t, C_2)/C_3] , \tag{2.16}$$

where

$$M(t, C_2) = \int \frac{dt}{S^2R^3} + C_2 , \tag{2.17}$$

and C_2 and C_3 are two arbitrary constants.

We notice here, in the case of (2.14), the metric function $R_0(t)$ is a known solution of (2.11). Hence, from the couple of metric functions $[R_0(t),\ S(t)]$, our generation technique allows us to obtain the new one $[R(t),\ S(t)]$. A similar statement holds in the case of (2.16).

III. GENERATION OF NEW EXACT SOLUTIONS

Using the Bianchi type II exact solution elaborated by Lorenz[4] as a particular solution to Eq. (2.12), we are able to linearize the Riccati equation (2.11). The Lorenz solution reads

$$S^2(\tau) = 2(q^2 - b^2)\lambda^{-1}(\tau) , \tag{3.1}$$

$$R^2(\tau) = [e^{2(q\tau + \phi)}/2(q^2 - b^2)]\lambda(\tau) , \tag{3.2}$$

where

$$\lambda(\tau) \equiv (q^2 - b^2)\cosh[2(q^2 - b^2)^{1/2}\tau + \psi] , \tag{3.3}$$

and q, b, ψ, and ϕ are arbitrary constants, with $|q| > |b|$. The pressure and density are given by

$$P = \rho = b^2/8\pi S^2R^4 . \tag{3.4}$$

The temporal variable τ is related to the old familiar one t by the relation

$$dt = SR^2\, d\tau . \tag{3.5}$$

Inserting now the values of $S(t)$ and $R(t)$ into formulas (2.15) and (2.14), we obtain the new class of solutions

$$S^2(t) = 2(q^2 - b^2)\lambda^{-1}(\tau) , \tag{3.6}$$

$$R^2(t) = \frac{1}{2(q^2 - b^2)C_1}\left\{\frac{-1}{q} + Ce^{2(q\tau + \phi)}\right\}\lambda(\tau) , \tag{3.7}$$

where C and C_1 are two arbitrary constants. We call this model E1.

Applying again formulas (2.15) and (2.14) (with E1 as a particular solution), we get the new class of solutions

$$S^2(t) = 2(q^2 - b^2)\lambda^{-1}(\tau) ,$$

$$\begin{aligned} R^2(t) = {}& \frac{\lambda(\tau)}{2(q^2 - b^2)C_4}\left\{\frac{-1}{q} + Ce^{2(q\tau + \phi)}\right\} \\ & \times\left\{qC_1\left[2q\tau - \ln\left(-\frac{1}{q} + Ce^{2(q\tau + \phi)}\right)\right] + C_3\right\} . \end{aligned} \tag{3.8}$$

We call this model E2.

IV. GEOMETRIC AND KINEMATIC PROPERTIES OF CLASS OF SOLUTIONS OBTAINED

In this section we discuss the kinematic and geometric properties of our solutions. We first find the values of the various kinematic quantities and then look more closely at these values.

The projection tensor $h_{ab} = g_{ab} + u_a u_b$ is used for splitting the covariant derivative of the four vector velocity u_i as follows:

$$u_{a;b} = -\dot u_a u_b + \omega_{ab} + \sigma_{ab} + \tfrac13\Theta h_{ab} , \tag{4.1}$$

where $\dot u_\theta$, ω_{ab}, Θ, and σ_{ab} are, accordingly, called acceleration, rotation, scalar expansion, and shear, respectively.[5] The expansion tensor $\theta_{\rho\nu}$ is defined by the relation

$$\Theta = g^{\alpha\beta}\theta_{\alpha\beta} . \tag{4.2}$$

The shear σ is given by

$$\sigma^2 = \tfrac12\sigma_{\alpha\beta}\sigma^{\alpha\beta} . \tag{4.3}$$

Two other important geometric quantities are also introduced; they measure, respectively, the dynamic importance of the shear and the dynamic importance of the fluid. They are, accordingly,

$$\beta = -2\sqrt3(\sigma/\Theta) , \tag{4.4}$$

$$\Omega = 3\rho/\Theta^2 . \tag{4.5}$$

Using metric (2.2) and straightforward calculation, we obtain

$$\Theta = 2\dot R/R + \dot S/S , \tag{4.6}$$

$$\sigma^2 = \tfrac13(\dot R/R - \dot S/S) . \tag{4.7}$$

The acceleration and the rotation of our models are zero.

Let us study now our universe E1 (a similar study holds for the universe E2). The density ρ and the pressure P are given by formula (3.4). We can then immediately conclude that the strong energy conditions of Hawking and Penrose,[6] which require that $\rho + p \geqslant 0$ and $\rho + 3p \geqslant 0$, are always verified. Using formulas (4.6) and (4.7), we find

$$\begin{aligned} \Theta = {}& (1/SR^2)[2(q^2 - b^2)^{1/2} \\ & \times\tanh(2(q^2 - b^2)^{1/2}\tau + \psi) \\ & + 2Cqe^{2(q\tau + \phi)}/(-1/2q) + Ce^{2(q\tau + \phi)}] , \end{aligned} \tag{4.8}$$

$$\begin{aligned} \sigma^2 = {}& (1/3S^2R^4)\{2(q^2 - b^2)^{1/2} \\ & \times\tanh[2(q^2 - b^2)^{1/2}\tau + \psi] \\ & + Cqe^{2(q\tau + \phi)}/(-1/2q) + Ce^{2(q\tau + \phi)}\} . \end{aligned} \tag{4.9}$$

The energy density ρ is given by (3.4) so we have

$$\rho = (1/8\pi)(b^2/S^2R^4) , \tag{4.10}$$

where S^2 and R^2 are given by formulas (3.6) and (3.7). For

$$SR^2 = 0 , \tag{4.11}$$

we have a singularity ($\rho \to \infty$, $\Theta \to \infty$, $\sigma \to \infty$). Equation (4.11) is satisfied for

$$Ce^{2(q\tau + \phi)} = 1/2q . \tag{4.12}$$

The above equation is verified for

$$\tau = \tau_0 \equiv \frac{1}{2q} \ln\left(\frac{1}{2qC}\right) - \frac{\phi}{q}. \qquad (4.13)$$

This equation can always be satisfied (q and C are two arbitrary constants). Hence, for $\tau = \tau_0$, E1 expands without bound ($\Theta \to \infty$); this expansion is highly anisotropic. Since

$$S^2 = S_0^2 = 2(q^2 - b^2)\lambda^{-1}(\tau_0),$$

and $R^2 = 0$ for $\tau \to \tau_0$, we have a "barrel" singularity.[7]

We can now evaluate β and obtain

$$\beta \to 1,$$

Ω goes to zero when $\tau \to \tau_0$. Hence, matter is dynamically negligible near the singularity; this agrees with a result already given by Collins.[8]

V. CONCLUDING REMARKS

As mentioned by Lorenz himself, his solution reduces in a particular case to that of Taub[9] and in other case to that of Maartens and Nel.[10] Thus model E1, which generalizes the Lorenz solution, constitutes a larger class of solution than those given by Taub and Maartens and Nel.

ACKNOWLEDGMENT

I would like to thank Dr. B. Leaute for a number of comments and suggestions concerning this paper.

[1] J. Hajj-Boutros, "Gravitation, geometry, and relativistic physics," in *Lecture Notes in Physics*, Vol. 212 (Springer, Berlin, 1984), p. 51.

[2] J. Hajj-Boutros, Math. Phys. **26**, 2297 (1985).

[3] D. Lorenz, Phys. Lett. A **79**, 19 (1980).

[4] Taking the constant $a = 0$ in formula 12 in Ref. 3, we obtain a perfect fluid distribution of matter; the electric and magnetic field are then zero.

[5] D. Kramer, H. Stephanie, E. Herlt, and M. MacCullum, *Exact Solutions of Einstein's Field Equations* (Deutscher Verlag, Berlin, 1980), p. 71.

[6] S. W. Hawking and R. Penrose, Proc. R. Soc. London Ser. A **314**, 529 (1970).

[7] M. A. H. MacCallum, Commun. Math. Phys. **20**, 57 (1971).

[8] C. B. Collins, J. Math. Phys. **18**, 2116 (1977).

[9] A. H. Taub, Ann. Math. **53**, 472 (1951).

[10] R. Maartens and S. D. Nel, Commun. Math. Phys. **59**, 273 (1978).

Invariant operators for the *n*-dimensional super-Poincaré algebra and the decomposition of the scalar superfield

R. Finkelstein and M. Villasante
Department of Physics, University of California, Los Angeles, California 90024

(Received 26 June 1985; accepted for publication 5 February 1986)

An analysis of supersymmetric Kaluza–Klein theories is begun by obtaining the Casimir operators for the super-Poincaré algebra in any number of dimensions. The knowledge of these operators is used to decompose the general scalar superfield in 11 dimensions into its irreducible parts. The irreducible superfields are expressed as products of Grassmann–Hermite functions and Grassmann–Bargmann–Wigner multispinor fields. Some Lagrangians for these superfields are written down. The formulation is off shell but global.

I. INTRODUCTION

We shall understand by the Kaluza–Klein hypothesis the conjecture that the physical continuum has more than the four dimensions that are usually ascribed to space-time and inferred from our relatively low energy physics. There are many ways of implementing this hypothesis that differ either in their formulations on the expanded physical continuum or in their rules for dimensional reduction. Here we shall make two basic assumptions only.

First, to describe the new manifold in the small let us assume an extended Poincaré or rotation-translation group in *n* dimensions with no additional timelike dimensions. Second, let us assume that the Poincaré algebra is also extended by the supersymmetry generators. The first assumption implies that the new manifold is homogeneous and isotropic in the small and the second introduces a particular physical hypothesis, the Fermi–Bose symmetry. These assumptions are too general to lead to a well-defined theory but provide conditions that hold for a large class of theories. In this paper we shall not discuss dimensional reduction or field content but it is of course known that local supersymmetry implies the Einstein gravitational field and some version of supergravity.

Off-shell formulations of extended supergravity in four dimensions have been discovered only for $N=1$ and $N=2$. Since the potentially most interesting theory corresponds to $N=8$ (maximal extension), it would be an important advance if a method could be devised to find the auxiliary fields needed to complete its off-shell algebra. One method that has been suggested requires the introduction of central charges. This possibility may be investigated in the context of a Kaluza–Klein theory: The central charges may then be identified with the components of the momentum associated with the additional dimensions, since these components commute with the usual space-time part of the angular momentum. Then these generators belong to the center of a four-dimensional supersymmetrically extended Poincaré algebra.

Since $N=8$ supergravity in four dimensions can be reformulated as simple ($N=1$) supergravity in 11 dimensions we shall emphasize the 11-dimensional Kaluza–Klein space. This space is especially attractive because it spontaneously compactifies into a ground state that is the product of space-time and a potentially realistic internal space.

Nevertheless there are many difficulties standing in the way of a realistic physical scenario for any of the models of supergravity that have been proposed so far.[1] For this reason the present paper is devoted to an essentially mathematical investigation.

II. SUPERALGEBRA

The Poincaré algebra in *n* dimensions (P_n) is given by the commutation relations

$$[P_A,P_B]=0, \tag{2.1}$$

$$[J_{AB},J^{CD}]=-4i\,\delta_{[A}{}^{[C}J_{B\,]}{}^{D]}, \tag{2.2}$$

$$[P^A,J_{BC}]=2i\,\delta^A{}_{[B}P_{C\,]} \tag{2.3}$$

where the indices run from 0 to $n-1$ and are raised and lowered with the Minkowski tensor

$$\eta^{AB}=\eta_{AB}=\mathrm{diag}(+\ -\ \cdots\ -).$$

The components P_A, where $A=5,\dots,n-1$, may be regarded as central charges in four dimensions.

The Dirac algebra in *n* dimensions consists of matrices of dimension $\Lambda=2^{[n/2]}$. In 2, 3, 4, 8, 9 mod 8 dimensions one can define Majorana spinors. Introduce fermionic generators into the algebra:

$$Q^\alpha,\quad \alpha=1,\dots,\Lambda.$$

This satisfies the Majorana condition and the commutation relations

$$\{Q^\alpha,Q^\beta\}=(\not{P}C^{-1})^{\alpha\beta}, \tag{2.4}$$

$$[Q^\alpha,P_A]=0, \tag{2.5}$$

$$[J_{AB},Q^\alpha]=-(i/2)\Gamma_{AB}{}^\alpha{}_\beta Q^\beta. \tag{2.6}$$

The commutation relations (2.1)–(2.6) define the super-Poincaré algebra in *n* dimensions SP_n. In the rest frame we have the "little group" $\mathrm{SO}(n-1)$.

In an odd number of dimensions Q corresponds to the representation

$$[\underbrace{\tfrac12\cdots\tfrac12}_{\nu}],$$

$\nu=[n/2]$, of $\mathrm{SO}(n-1,1)$. Upon restriction to $\mathrm{SO}(n-1)$, Q corresponds to

$$[\underbrace{\tfrac12\cdots\tfrac12}_{\nu}]+[\underbrace{\tfrac12\cdots\tfrac12\ -\tfrac12}_{\nu}]\,.$$

0022-2488/86/061595-12$02.50

These two parts are separated by the operators $\frac{1}{2}(I+\Gamma_0)$ and $\frac{1}{2}(I-\Gamma_0)$, which are the rest frame forms of the covariant operators in (2.7) below. (In an even number of dimensions Q corresponds to the reducible representation

$$[\underbrace{\tfrac{1}{2}\tfrac{1}{2}\cdots\tfrac{1}{2}}_{\nu}]+[\underbrace{\tfrac{1}{2}\cdots\tfrac{1}{2}\ -\tfrac{1}{2}}_{\nu}],$$

$\nu=[n/2]$, of $\mathrm{SO}(n-1,1)$. These two parts are separated by the operators $\frac{1}{2}(I+\Gamma_{n+1})$ and $\frac{1}{2}(I-\Gamma_{n+1})$, where Γ_{n+1} is the analog of Dirac's γ_5 in four dimensions. Each representation becomes $[\underbrace{\tfrac{1}{2}\cdots\tfrac{1}{2}}_{\nu-1}]$ when $\mathrm{SO}(n-1,1)$ is reduced to $\mathrm{SO}(n-1)$.)

$P^2=P_AP^A$ is an obvious Casimir operator of the superalgebra. Its eigenvalues M^2 characterize the representations and can be used to define positive and negative "energy" projection operators

$$\Lambda_\pm=(1/2M)(M\pm\not{P}). \tag{2.7}$$

They satisfy

$$\Lambda_\pm\Lambda_\mp=0,\quad \Lambda_\pm\Lambda_\pm=\Lambda_\pm,\quad \Lambda_++\Lambda_-=I, \tag{2.8}$$

$$C\Lambda_\pm C^{-1}=\Lambda_\mp{}^T, \tag{2.9}$$

$$\Lambda_\pm\not{P}=\pm M\Lambda_\pm. \tag{2.10}$$

We can then split the Majorana spinor Q into two pieces:

$$Q_\pm=\Lambda_\pm Q, \tag{2.11}$$

with the anticommutation relations

$$\{Q_\pm{}^\alpha,Q_\pm{}^\beta\}=0, \tag{2.12}$$

$$\{Q_+{}^\alpha,Q_-{}^\beta\}=(\Lambda_-\not{P}C^{-1})^{\beta\alpha}=M(\Lambda_+C^{-1})^{\alpha\beta}. \tag{2.13}$$

So we have raising and lowering operators, which we can use to construct the irreducible representations of the superalgebra. We should notice at this point that Q_+ and Q_- each have only $\frac{1}{2}\Lambda$ components.

We shall consider the action of this algebra on the space of superfields. For example, let Φ be a scalar superfield. Then

$$\Phi(\theta,x)=\sum_{\mathscr{A}_1\cdots}\Theta^{\mathscr{A}_1\cdots}(\theta)F_{\mathscr{A}_1\cdots}(x). \tag{2.14}$$

Here the $\mathscr{A}_{1\cdots}$ are spinor indices in the n-dimensional space, the x are the space-time coordinates, and the θ are the anticommuting coordinates of superspace, while $\Theta^{\mathscr{A}_1\cdots}$ and $F_{\mathscr{A}_1\cdots}$ are contravariant and covariant multispinors in Kaluza–Klein space. Let X be a generator of the superalgebra. Then the action of X on the superfield will induce corresponding transformation on the tensor coefficients $F_{\mathscr{A}_1\cdots}$ as follows:

$$\delta\Phi=X\Phi, \tag{2.15a}$$

$$\delta F_{\mathscr{A}_1\cdots}=\sum_{\mathscr{B}_1\cdots}D_{\mathscr{A}_1\cdots}{}^{\mathscr{B}_1\cdots}F_{\mathscr{B}_1\cdots}. \tag{2.15b}$$

In this way the space of tensor fields like $F_{\mathscr{A}_1\cdots}(x)$ becomes the basis for a representation of the superalgebra. The basis $\{F_{\mathscr{A}_1\cdots}\}$ is very large but reducible. Our aim is to project out of the total superfield the parts that are irreducible under the superalgebra. These irreducible parts contain both the physical fields and the auxiliary fields that are needed to complete the off-shell algebra. To determine these irreducible fields we next discuss the irreducible representations.

III. CASIMIRS OF SUPERALGEBRA

To find the full set of Casimirs one may try to generalize the corresponding operators for the Poincaré group in four dimensions. For this purpose one needs a generalized spin operator U_{AB} with the properties

$$[U_{AB},P_C]=0, \tag{3.1}$$

$$[U_{AB},Q^\alpha]=0. \tag{3.2}$$

If (3.1) and (3.2) are satisfied then Casimirs may be constructed as the traces

$$\mathrm{tr}\,U^m.$$

Although these traces will be Casimirs for all values of m, the number of independent ones is limited and equals the rank of the group. In addition since the traces of the odd powers may be expressed in terms of the traces of the even powers, we shall choose

$$C_m=\mathrm{Tr}\,U^{2m}. \tag{3.3}$$

To obtain an explicit form of U_{AB} we impose

$$U_{AB}=-U_{BA}, \tag{3.4}$$

$$P^AU_{AB}=0. \tag{3.5}$$

Equation (3.5) is imposed so that the only nonvanishing components of U_{AB} in the proper frame $[P^A=(P^0,0)]$ are U_{jk}, where $j,k=1,\ldots,n-1$. Then $U_{AB}U^{AB}$ will describe the generalized $(\mathrm{spin})^2$ and in this respect it resembles the square of the four-dimensional Pauli–Lubanski vector. Since U_{AB} is an angular momentum satisfying (3.4) and (3.5), let us put (Γ-tensors are defined in Appendix A)

$$U_{AB}=J_{AB}+aP^E(J_{EA}P_B-J_{EB}P_A)+bP^E\overline{Q}\Gamma_{EAB}Q. \tag{3.6}$$

Then (3.4) holds by construction. To satisfy (3.5) choose $a=1/P^2$; then (3.1) is also satisfied. Finally (3.2) requires that $b=-i/4P^2$. Therefore

$$U_{AB}=J_{AB}+\frac{1}{P^2}P^E(J_{EA}P_B-J_{EB}P_A)-\frac{i}{4}\frac{P^E}{P^2}\overline{Q}\Gamma_{EAB}Q \tag{3.7}$$

is the object from which we can construct the Casimirs, either by (3.3) or by forming scalars out of the generalized Pauli–Lubanski tensors:

$$W^{A_1\cdots A_k}=\epsilon^{A_1\cdots A_kB_1\cdots B_{n-k}}P_{B_1}U_{B_2B_3}\cdots U_{B_{n-k-1}B_{n-k}}.$$

The commutation rules for this generalized spin, or superspin, are

$$\begin{aligned}[U_{AB},U^{CD}]={}&-4i\delta_{[A}{}^{[C}U_{B]}{}^{D]}\\&-(4i/P^2)P_{[A}U_{B]}{}^{[C}P^{D]},\end{aligned} \tag{3.8}$$

$$A,B,C,D=0,\ldots,n-1.$$

In the rest frame one finds

$$[U_{ij},U^{kl}]=-4i\delta_{[i}{}^{[k}U_{j]}{}^{l]},\quad i,j=1,\ldots,n-1. \tag{3.9}$$

Then the U_{ij} satisfy the algebra of the rotation group $\mathrm{SO}(n-1)$; that is, $\mathrm{SO}(n-1)$ is the "little group."

The Casimir operators

$$C_m = \mathrm{Tr}\, U^{2m} = U_{B_1}{}^{B_2} U_{B_2}{}^{B_3} \cdots U_{B_{2m}}{}^{B_1} \tag{3.10}$$

become in the rest frame

$$C_m = U_{i_1}{}^{i_2} U_{i_2}{}^{i_3} \cdots U_{i_{2m}}{}^{i_1}. \tag{3.11}$$

When $n-1$ is even $(n-1=2\nu)$, the operators in (3.11) do not suffice to characterize completely the representations of $\mathrm{SO}(n-1)$. In that case one needs to make use of the Pfaffian of the matrix $U_i{}^j$,

$$C'_\nu = \mathrm{Pf}(U) = \epsilon^{i_1 \cdots i_{2\nu}} U_{i_1 i_2} \cdots U_{i_{2\nu-1} i_{2\nu}}, \tag{3.12}$$

and a complete set of independent Casimir operators is $C_1, C_2, \ldots, C_{\nu-1}, C'_\nu$.

When $n-1$ is odd $(n=2\nu)$, the traces $(C_1, C_2, \ldots, C_{\nu-1})$ are sufficient to form a complete set. Therefore we can choose, as the complete set of Casimir operators for SP_n,

$$\begin{aligned} &P^2, C_m, \quad m = 1,\ldots,\nu-1 \quad \text{if} \quad n = 2\nu, \\ &P^2, C_m, C'_\nu, \quad m = 1,\ldots,\nu-1 \quad \text{if} \quad n = 2\nu+1, \end{aligned} \tag{3.13}$$

where

$$C'_\nu = (1/M)\epsilon^{A_1 \cdots A_n} P_{A_1} U_{A_2 A_3} \cdots U_{A_{n-1} A_n},$$

where M^2 is the eigenvalue of P^2. Equation (3.13) reduces to (3.12) in the rest frame.

Since the operators C'_ν and C_m are "Lorentz" invariants, they can be evaluated in the rest frame, and, therefore, the problem of finding the eigenvalues for the operators (3.10) and (3.11) of SP_n reduces to the problem of finding the eigenvalues for the operators (3.11) and (3.12) of $\mathrm{SO}(n-1)$.

IV. SUPERSPACE

The algebra SP_n can be represented in the superspace[2] with coordinates

$$(x^A, \theta^\alpha), \quad A = 0,1,\ldots,n-1, \quad \alpha = 1,2,\ldots,\Lambda,$$

where the θ are anticommuting coordinates that are arranged in a Majorana spinor. Superfields are arbitrary functions of these coordinates that are at most polynomials of degree Λ in the θ coordinates:

$$\Phi_J(x,\theta) = \sum_{j=0}^{\Lambda} \theta^{\alpha_1} \cdots \theta^{\alpha_j} X_{J\alpha_1 \cdots \alpha_j}(x). \tag{4.1}$$

Here J represents a collection of Kaluza–Klein indices. These general superfields provide a basis for an enlarged super-Poincaré algebra that includes the covariant derivatives D^α. These obey the commutation relations

$$\begin{aligned} &\{Q^\alpha, D^\beta\} = 0, \\ &\{D^\alpha, D^\beta\} = -(\not{P} C^{-1})^{\alpha\beta}, \\ &[P^A, D^\alpha] = 0, \\ &[J_{AB}, D^\alpha] = -(i/2)\Gamma_{AB}{}^\alpha{}_\beta D^\beta. \end{aligned} \tag{4.2}$$

Q^α and D^α may be represented as differential operators:

$$Q^\alpha = i\left(\frac{\partial}{\partial\bar{\theta}_\alpha} + \frac{1}{2}\not{P}^\alpha{}_\beta \theta^\beta\right), \tag{4.3}$$

$$D^\alpha = i\left(\frac{\partial}{\partial\bar{\theta}_\alpha} - \frac{1}{2}\not{P}^\alpha{}_\beta \theta^\beta\right). \tag{4.4}$$

We can decompose D in the same way as Q:

$$D_+ = \Lambda_+ D \quad \text{and} \quad D_- = \Lambda_- D, \tag{4.5}$$

which satisfy

$$\begin{aligned} &\{D_\pm{}^\alpha, D_\pm{}^\beta\} = 0, \\ &\{D_+{}^\alpha, D_-{}^\beta\} = M(\Lambda_- C^{-1})^{\alpha\beta} = -M(\Lambda_+ C^{-1})^{\alpha\beta}. \end{aligned} \tag{4.6}$$

The basis states for an irreducible representation of P_n are generated by first applying to a rest state the rotations of the little group to obtain a manifold of rest states, and by then applying the boosts that generate states with arbitrary momentum. To obtain the basis states for an irreducible representation of SP_n we may first apply the operators Q_+ and their products to the manifold of rest states just described and we may then apply boosts to the resulting states. If the initial rest state is annihilated by Q_-, one will thereby obtain an irreducible representation of SP_n. Therefore an irreducible superfield can be generated from an irreducible $\mathrm{SO}(n-1)$ representation $|\Omega\rangle$, serving as a Clifford lowest state $(Q_-|\Omega\rangle = 0)$, by application of the $\Lambda/2$ lifting operators $Q_+{}^\alpha$. One obtains

$$|\Omega\rangle, Q_+{}^\alpha|\Omega\rangle, Q_+{}^{[\alpha_1} Q_+{}^{\alpha_2]}|\Omega\rangle, \ldots, Q_+{}^{[\alpha_1} \cdots Q_+{}^{\alpha_{\Lambda/2}]}|\Omega\rangle. \tag{4.7}$$

The dimension of this representation is then

$$\dim \Omega \times \sum_{j=0}^{\Lambda/2} \binom{\Lambda/2}{j} = 2^{\Lambda/2} \times \dim \Omega. \tag{4.8}$$

Equivalently, we can generate it from a $\mathrm{SO}(n-1)$ state $|\Omega'\rangle$, serving as a Clifford highest state: $Q_+|\Omega'\rangle = 0$ by application of the $\Lambda/2$ lowering operators Q^α_-,

$$\begin{aligned} &|\Omega'\rangle, Q_-{}^\alpha|\Omega'\rangle, Q_-{}^{[\alpha_1} Q_-{}^{\alpha_2]}|\Omega'\rangle, \\ &\quad \ldots, Q_-{}^{[\alpha_1} \cdots Q_-{}^{\alpha_{\Lambda/2}]}|\Omega'\rangle. \end{aligned} \tag{4.9}$$

The dimension of this representation is

$$\dim \Omega' \times 2^{\Lambda/2}. \tag{4.10}$$

In a similar way, a general superfield may be obtained from an irreducible superfield (a super-Poincaré state) $|\bar{\Omega}\rangle$ acting now as a Clifford lowest state $D_-|\bar{\Omega}\rangle = 0$, by application of the $\Lambda/2$ raising operators $D_+{}^\alpha$,

$$\begin{aligned} &|\bar{\Omega}\rangle, D_+{}^\alpha|\bar{\Omega}\rangle, D_+{}^{[\alpha_1} D_+{}^{\alpha_2]}|\bar{\Omega}\rangle \\ &\quad \cdots D_+{}^{[\alpha_1} \cdots D_+{}^{\alpha_{\Lambda/2}]}|\bar{\Omega}\rangle, \end{aligned} \tag{4.11}$$

or from a Clifford highest state $|\bar{\Omega}'\rangle$, $D_+|\bar{\Omega}'\rangle = 0$, by application of the $\Lambda/2$ lowering operators $D_-{}^\alpha$:

$$\begin{aligned} &|\bar{\Omega}'\rangle, D_-{}^\alpha|\bar{\Omega}\rangle, D_-{}^{[\alpha_1} D_-{}^{\alpha_2]}|\bar{\Omega}\rangle, \\ &\quad \cdots D_-{}^{[\alpha_1} \cdots D_-{}^{\alpha_{\Lambda/2}]}|\bar{\Omega}'\rangle, \end{aligned} \tag{4.12}$$

with dimensions $2^{\Lambda/2} \dim \bar{\Omega}$ and $2^{\Lambda/2} \dim \bar{\Omega}'$, respectively. For instance, the dimension of the general scalar superfield is $2^{\Lambda/2} \times 2^{\Lambda/2}$.

Now we particularize to the 11-dimensional case: $n = 11$ and $\Lambda = 32$. The spinor D splits into two parts D_+ and D_-, which transform according to the representations $[\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\ -\frac{1}{2}]$ and $[\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}]$ of $\mathrm{SO}(10)$ in the rest frame, re-

TABLE I. Decomposition of totally antisymmetric Kronecker product.

j	$D_+^{[\alpha_1}\cdots D_+^{\alpha_j]}$
0	$[0\,0\,0\,0\,0]$
1	$[\tfrac12\,\tfrac12\,\tfrac12\,\tfrac12\,-\tfrac12]$
2	$[1\,1\,1\,0\,0]$
3	$[\tfrac32\,\tfrac32\,\tfrac12\,\tfrac12\,\tfrac12]$
4	$[2\,2\,0\,0\,0]\oplus[2\,1\,1\,1\,1]$
5	$[\tfrac52\,\tfrac32\,\tfrac12\,\tfrac12\,\tfrac12]\oplus[\tfrac32\,\tfrac32\,\tfrac32\,\tfrac32\,\tfrac32]$
6	$[3\,1\,1\,0\,0]\oplus[2\,2\,1\,1\,1]$
7	$[\tfrac72\,\tfrac12\,\tfrac12\,\tfrac12\,-\tfrac12]\oplus[\tfrac52\,\tfrac32\,\tfrac32\,\tfrac12\,\tfrac12]$
8	$[4\,0\,0\,0\,0]\oplus[3\,1\,1\,1\,0]\oplus[2\,2\,2\,0\,0]$
9	$[\tfrac72\,\tfrac12\,\tfrac12\,\tfrac12\,\tfrac12]\oplus[\tfrac52\,\tfrac32\,\tfrac32\,\tfrac12\,-\tfrac12]$
10	$[3\,1\,1\,0\,0]\oplus[2\,2\,1\,1\,-1]$
11	$[\tfrac52\,\tfrac32\,\tfrac12\,\tfrac12\,-\tfrac12]\oplus[\tfrac32\,\tfrac32\,\tfrac32\,\tfrac32\,-\tfrac32]$
12	$[2\,2\,0\,0\,0]\oplus[2\,1\,1\,1\,-1]$
13	$[\tfrac32\,\tfrac32\,\tfrac12\,\tfrac12\,-\tfrac12]$
14	$[1\,1\,1\,0\,0]$
15	$[\tfrac12\,\tfrac12\,\tfrac12\,\tfrac12\,\tfrac12]$
16	$[0\,0\,0\,0\,0]$

spectively. Then

$$D_+{}^{[\alpha_1}\cdots D_+{}^{\alpha_j]}$$

corresponds to the totally antisymmetrized jth Kronecker power of $[\tfrac12\tfrac12\tfrac12\tfrac12\;-\tfrac12]$, which has the dimension $\binom{16}{j}$. We have reduced these product representations into their irreducible components. The results are given in Table I. If D_- is used then Table I is inverted.

Table I tells us immediately the "superspins" (highest superweights) included in the scalar superfield in 11 dimensions. It contains 27 irreducible superfields, some of them degenerate. The number of ordinary fields is huge. Later we shall see how to separate the different irreducible superfields.

V. THE SCALAR SUPERFIELD

We consider the scalar superfield. For this case one has

$$J_{AB} = -(x_A P_B - x_B P_A) + \frac{1}{2} i\bar\theta\Gamma_{AB}\frac{\partial}{\partial\bar\theta}. \qquad (5.1)$$

Using (5.1), (4.3), (4.4), and (3.7) one finds for U_{AB} the following simple expression:

$$U_{AB} = -\tfrac14 i(P^E/P^2)\bar D\Gamma_{EAB}D. \qquad (5.2)$$

One easily checks that this U_{AB} satisfies (3.8) if D^α satisfies (4.2). However, (5.2) describes generators U_{AB} for only a special choice of J_{AB}. Therefore only a small number of representations of (3.8) and (3.9) can be constructed if U_{AB} is restricted to (5.2). In order to identify these particular representations of SO$(n-1)$ we compute the eigenvalues of the Casimirs that follow from (5.2). These same eigenvalues will also give us the representations of SP$_n$ included in the scalar superfield.

VI. CASIMIRS AS FUNCTIONS OF COVARIANT DERIVATIVES

The Casimirs are given by (3.4) in terms of (4.3). For example,

$$C_1 = \tfrac{1}{16}(P^E P^F/P^4)(\bar D\Gamma_{EAB}D)(\bar D\Gamma_F{}^{AB}D). \qquad (6.1)$$

As described in Appendix B, C_1 may be simplified by repeated Fierz transformations where the precise anticommutation relations of the D^α must be taken into account. One gets the result

$$C_1 = \tfrac38(1/P^2)[(\bar DD)^2 - 16^2P^2]. \qquad (6.2)$$

All the even Casimirs may be expressed in terms of

$$Y_{AB} = U_{AC}U^C{}_B. \qquad (6.3)$$

Then

$$C_m = \operatorname{tr} Y^m, \quad m = 1,2,3,4. \qquad (6.4)$$

We may write

$$Y_{AB} = X\Pi_{AB} + Z_{AB} + 4iU_{AB}, \qquad (6.5)$$

where

$$\Pi_{AB} = \delta_{AB} - P_AP_B/P^2, \quad Z_{AB} = Z_{BA}. \qquad (6.6)$$

The antisymmetric part of Y_{AB} is simply $4iU_{AB}$. In addition

$$P^AY_{AB} = 0, \quad P^AZ_{AB} = 0. \qquad (6.7)$$

Then

$$\operatorname{tr} Y = 10X + \operatorname{tr} Z$$

or

$$C_1 = 10X + \operatorname{tr} Z. \qquad (6.8)$$

According to (6.8), X is fixed by C_1 alone if Z vanishes. If Z happens to vanish for a particular representation then it is possible to express all Casimirs as functions of $(\bar DD)^2$ and P^2 for that representation as one sees from (6.4), (6.5), (6.7), and (6.2). In fact one finds if $Z=0$,

$$C_2 = -(C_1/10)(160 - C_1),$$
$$C_3 = (C_1/100)(C_1{}^2 - 480C_1 + 25\,600), \qquad (6.9)$$
$$C_4 = -(C_1/1000)[40(320-4C_1)^2 - C_1(160-C_1)^2].$$

Here C_1 is exact for all representations but the other expressions are correct only if Z vanishes for that representation.

VII. EIGENVALUES OF $(\bar DD)^2$ AND CASIMIRS

To evaluate the Casimirs according to the above formulas one needs only the corresponding eigenvalues of $(\bar DD)^2$. The latter may be determined as follows. Since D, like Q, is a spinor in the 11-dimensional space it has $2^5 = 32$ components. Therefore if a completely antisymmetric multilinear form has more than 32 factors, it will vanish. Since these forms are polynomials in $(\bar DD)^2$, a vanishing form provides a characteristic equation for the eigenvalues of this operator. One may construct completely antisymmetric functions of the D^α as follows:

$$D^{\alpha_1\alpha_2} = \tfrac12[D^{\alpha_1},D^{\alpha_2}] = -D^{\alpha_2\alpha_1}, \qquad (7.1a)$$
$$D^{\alpha_1\alpha_2\alpha_3} = \tfrac12\{D^{\alpha_3},D^{\alpha_1\alpha_2}\} = -D^{\alpha_3\alpha_2\alpha_1}. \qquad (7.1b)$$

By forming commutators and anticommutators alternately we obtain completely antisymmetric functions of the D^α. Let

$$G = \bar DD = C_{\alpha\beta}D^\alpha D^\beta = C_{\alpha\beta}D^{\alpha\beta}, \qquad (7.2)$$
$$G_m = C_{\alpha_1\beta_1}\cdots C_{\alpha_m\beta_m}D^{\alpha_1\cdots\beta_m}, \qquad (7.3)$$
$$G_1 = G, \quad G_0 = 1, \quad G_{-1} = 0. \qquad (7.4)$$

The product GG_m can be decomposed using the formulas in Appendix A. Then we get the recursion relation

$$G_{m+1} = GG_m - (m/2)(\Lambda - 2m + 2)P^2 G_{m-1}, \tag{7.5}$$

where

$$\Lambda = 2^{[n/2]}.$$

Here $\Lambda = 32$.

Since there are only 32 different D^{α}, any antisymmetric functions containing more than 32 D^{α} must vanish. Then

$$G_{17} = C_{\alpha_1\beta_1} \cdots C_{\alpha_{17}\beta_{17}} D^{\alpha_1\beta_1 \cdots \alpha_{17}\beta_{17}} = 0. \tag{7.6}$$

If (7.6) is expanded by the recursion formula (7.5) one obtains a polynomial in G^2. Let

$$G^2 = x^2 P^2. \tag{7.7}$$

The algebraic equation (7.6) becomes

$$x \sum_0^8 A_s x^{2s} = 0, \tag{7.8}$$

where

$$A_8 = 1,$$
$$A_7 = -816,$$
$$A_6 = 262\,752,$$
$$A_5 = -42\,828\,032,$$
$$A_4 = 3, 773, 223, 168,$$
$$A_3 = -177, 891, 237, 888,$$
$$A_2 = 4, 165, 906, 530, 304,$$
$$A_1 = -40, 683, 662, 475, 264,$$
$$A_0 = 106, 542, 032, 486, 400.$$

The 17 roots of (7.8) are

$$x_k = \pm 2k, \quad k = 0,...,8, \tag{7.9}$$

or

$$(\bar{D}D)^2 = (2k)^2 P^2, \quad \bar{D}D = \pm 2k\sqrt{P^2}.$$

The corresponding eigenvalues of C_1 are

$$c_{1k} = \tfrac{3}{8}\,[x_k{}^2 - 16^2]. \tag{7.10}$$

The eigenvalues of the Casimirs C_2, C_3, and C_4 may also be computed from Eq. (6.9) for particular representations if Z vanishes. If Z does not vanish, it is still possible in principle to find the eigenvalues of C_2, C_3, and C_4 by obtaining polynomial equations for these operators just as for $\bar{D}D$. If the procedure were fully carried out, one would thereby determine the irreducible representations of SO(10) that can be realized in terms of the generators (5.2) and the commutation rules (4.2). This procedure is not practical, however, and in Sec. VIII we calculate the eigenvalues of C_2, C_3, and C_4 by a different method. After this is done we discover a few representations for which Eqs. (6.9) are indeed valid and for which we conclude that Z vanishes.

VIII. EIGENVALUES OF CASIMIRS OF SO(10)

We return to the general definition of C_n in (3.3). One is then no longer limited to the particular construction of C_n in terms of the covariant derivatives D^{α} such as (5.1). The eigenvalues of C_n, the Pfaffian and similar operators have been computed quite generally for the classical groups by Perelomov and Popov. Using their formulas we get the eigenvalues for tr U^k of SO(2ν) (see Ref. 3):

$$\mathrm{Tr}\, U^k = i^k \sum_{j=1}^{2\nu} \sum_{l=1}^{2\nu} (a^k)_{jl}, \tag{8.1}$$

where

$$a_{jk} = (l_j + \alpha)\delta_{jk} + (\beta/2)(1 + \epsilon_j)\delta_{j,2\nu-k+1} - \theta_{jk}, \tag{8.2}$$

$$\theta_{jk} = \begin{cases} 1, & \text{if } k > j, \\ 0, & \text{otherwise}, \end{cases} \tag{8.3}$$

$$\epsilon_j = \begin{cases} 1, & \text{if } j \leqslant \nu, \\ -1, & \text{if } \nu < j \leqslant 2\nu, \end{cases} \tag{8.4}$$

$$l_j = \begin{cases} m_j + r_j, & j \leqslant \nu, \\ -m_{2\nu-j+1} - r_{2\nu-j+1}, & \nu < j \leqslant 2\nu, \end{cases} \tag{8.5}$$

$$r_j = \nu\epsilon_j - j, \quad j \leqslant \nu, \tag{8.6}$$

and $m_j, j \leqslant \nu$, are the components of the highest weight of the representation. So the eigenvalues of C_n are

$$c_n = (-)^n \sum_{j,k} (a^{2n})_{jk}. \tag{8.7}$$

The eigenvalues for the Pfaffian C'_ν are given by

$$c'_\nu = 2^\nu \nu! l_1 l_2 \cdots l_\nu. \tag{8.8}$$

With this information we can compute Table II. The formula (8.1) gives eigenvalues of the Casimirs for all representations of the rotation group. It gives in particular the values corresponding to the D-representations, i.e., those listed in Table I. On the other hand the list in Table II contains all representations in which C_1 corresponds to one of the 17

TABLE II. Casimirs of irreducible representations of SO(10).

Highest weight	C_1	C_2	C_3	C_4	C'_5	
0 0 0 0 0	0	0	0	0	0	*
$\frac12\frac12\frac12\frac12\frac12$	$-\frac{45}{2}$	$\frac{3285}{8}$	$-\frac{265725}{32}$	$\frac{21523365}{128}$	113 400	*
$\frac12\frac12\frac12\frac12\ -\frac12$	$-\frac{45}{2}$	$\frac{3285}{8}$	$-\frac{265725}{32}$	$\frac{21523365}{128}$	− 113 400	*
1 1 1 0 0	− 42	1050	− 45 402	2173 290	0	
$\frac32\frac32\frac12\frac12\frac12$	$-\frac{117}{2}$	$\frac{14373}{8}$	$-\frac{3372309}{32}$	$\frac{918435141}{128}$	178 200	
$\frac32\frac32\frac12\frac12\ -\frac12$	$-\frac{117}{2}$	$\frac{14373}{8}$	$-\frac{3372309}{32}$	$\frac{918435141}{128}$	− 178 200	
2 1 1 1 1	− 72	2304	− 146 016	12 644 064	552 960	
2 1 1 1 − 1	− 72	2304	− 146 016	12 644 064	− 552 960	
2 2 0 0 0	− 72	2880	− 221 472	17 892 000	0	
$\frac52\frac32\frac12\frac12\frac12$	$-\frac{165}{2}$	$\frac{27525}{8}$	$-\frac{9147045}{32}$	$\frac{3567155205}{128}$	210 600	
$\frac52\frac32\frac12\frac12\ -\frac12$	$-\frac{165}{2}$	$\frac{27525}{8}$	$-\frac{9147045}{32}$	$\frac{3567155205}{128}$	− 210 600	
$\frac32\frac32\frac32\frac32\frac32$	$-\frac{165}{2}$	$\frac{16005}{8}$	$-\frac{1900965}{32}$	$\frac{229696005}{128}$	1247 400	*
$\frac32\frac32\frac32\frac32\ -\frac32$	$-\frac{165}{2}$	$\frac{16005}{8}$	$-\frac{1900965}{32}$	$\frac{229696005}{128}$	− 1247 400	*
2 2 1 1 1	− 90	2970	− 176 490	12 858 570	691 200	
2 2 1 1 − 1	− 90	2970	− 176 490	12 858 570	− 691 200	
3 1 1 0 0	− 90	4410	− 429 930	48 865 770	0	
$\frac72\frac12\frac12\frac12\frac12$	$-\frac{189}{2}$	$\frac{45045}{8}$	$-\frac{20628909}{32}$	$\frac{10708941285}{128}$	189 000	
$\frac72\frac12\frac12\frac12\ -\frac12$	$-\frac{189}{2}$	$\frac{45045}{8}$	$-\frac{20628909}{32}$	$\frac{10708941285}{128}$	− 189 000	
$\frac52\frac32\frac32\frac12\frac12$	$-\frac{189}{2}$	$\frac{28917}{8}$	$-\frac{8549037}{32}$	$\frac{3139651557}{128}$	294 840	
$\frac52\frac32\frac32\frac12\ -\frac12$	$-\frac{189}{2}$	$\frac{28917}{8}$	$-\frac{8549037}{32}$	$\frac{3139651557}{128}$	− 294 840	
4 0 0 0 0	− 96	7296	− 961 152	136 989 312	0	
3 1 1 1 0	− 96	4416	− 411 072	46 813 632	0	
2 2 2 0 0	− 96	3264	− 191 040	12 098 112	0	

roots of (7.8). The lists of representations in the two tables are the same.

In eight cases ($j\leqslant 3, j\geqslant 13$) the value of C_1 alone is sufficient to label the representation. It is also interesting to note that the Casimirs of some of the other representations in Table II (the starred representations) are correctly given by the simple formulas (6.9) of Sec. VI.

A straightforward method for extracting the irreducible parts involves the construction of projection operators belonging to the different Casimirs.[4] The operator

$$\Lambda_i(C) = \frac{\Pi_{j\neq i}(C-c_j)}{\Pi_{j\neq i}(c_i-c_j)} \tag{8.9}$$

projects out the eigenvalue c_i of C. These operators clearly satisfy

$$\Lambda_i(C)\Lambda_j(C) = \delta_{ij}\Lambda_i(C).$$

Construction of these operators becomes possible with the aid of the information contained in Table II. One may then, in principle, project out the irreducible superfields as follows:

$$\Psi(c_{1m_1},c_{2m_2},c_{3m_3},c_{4m_4},c_{5m_5}) = \prod_{i=1}^{5}\Lambda_{m_i}(C_i)\Psi, \tag{8.10}$$

where Ψ is the general superfield.

Since some representations appear twice, the above procedure will not completely separate all the irreducible representations. To remove the remaining degeneracy we need another operator, which is conveniently furnished by $\bar{D}D$. In this case the procedure is not so complicated as it may appear since there is very little degeneracy left after the eigenvalue of $\bar{D}D$ is chosen. Consequently the additional projection operators that are needed have very few factors. The full set of 27 projection operators is listed in Table III.

TABLE III. Projection operators for irreducible parts of the scalar superfield.

Superweights	Projection operators
$0\,0\,0\,0\,0$	$\Lambda_{\pm 16}(\bar{D}D)$
$\frac12\frac12\frac12\frac12 \pm\frac12$	$\Lambda_{\pm 14}(\bar{D}D)$
$1\,1\,1\,0\,0$	$\Lambda_{\pm 12}(\bar{D}D)$
$\frac32\frac32\frac12\frac12 \mp\frac12$	$\Lambda_{\pm 10}(\bar{D}D)$
$2\,1\,1\,1 \mp 1$	$\Lambda_{\pm 8}(\bar{D}D)\,\dfrac{C_2-2880}{2304-2880}$
$2\,2\,0\,0\,0$	$\Lambda_{\pm 8}(\bar{D}D)\,\dfrac{C_2-2304}{2880-2304}$
$\frac32\frac32\frac32\frac32 \mp\frac32$	$\Lambda_{\pm 6}(\bar{D}D)\,\dfrac{C_2-\frac{27525}{8}}{\frac{16005}{8}-\frac{27525}{8}}$
$\frac52\frac32\frac12\frac12 \mp\frac12$	$\Lambda_{\pm 6}(\bar{D}D)\,\dfrac{C_2-\frac{16005}{8}}{\frac{27525}{8}-\frac{16005}{8}}$
$3\,1\,1\,0\,0$	$\Lambda_{\pm 4}(\bar{D}D)\,\dfrac{C_2-2970}{4410-2970}$
$2\,2\,1\,1 \mp 1$	$\Lambda_{\pm 4}(\bar{D}D)\,\dfrac{C_2-4410}{2970-4410}$
$\frac52\frac32\frac32\frac12 \mp\frac12$	$\Lambda_{\pm 2}(\bar{D}D)\,\dfrac{C_2-\frac{45045}{8}}{\frac{28917}{8}-\frac{45045}{8}}$
$\frac72\frac12\frac12\frac12 \pm\frac12$	$\Lambda_{\pm 2}(\bar{D}D)\,\dfrac{C_2-\frac{28917}{8}}{\frac{45045}{8}-\frac{28917}{8}}$
$2\,2\,2\,0\,0$	$\Lambda_{\pm 0}(\bar{D}D)\,\dfrac{(C_2-7296)(C_2-4416)}{(3264-7296)(3264-4416)}$
$3\,1\,1\,1\,0$	$\Lambda_{\pm 0}(\bar{D}D)\,\dfrac{(C_2-7296)(C_2-3264)}{(4416-7296)(4416-3264)}$
$4\,0\,0\,0\,0$	$\Lambda_{\pm 0}(\bar{D}D)\,\dfrac{(C_2-4416)(C_2-3264)}{(7296-4416)(7296-3264)}$

Here $\Lambda_\lambda(\bar{D}D)$ satisfies $\bar{D}D\Lambda_\lambda\ (\bar{D}D) = \lambda M\Lambda_\lambda\ (\bar{D}D)$.

In spite of this simplification the method of projection operators does not seem the most useful approach here. Instead we adopt a different procedure based on the observation that C_1 depends only on the operator $\bar{D}D$. Therefore one obtains the eigenstates of C_1 by solving the following differential equation:

$$\bar{D}D\psi_n = \epsilon_n\psi_n. \tag{8.11}$$

The eigenvalues of $\bar{D}D$ have already been found and are equally spaced as for a harmonic oscillator. Since the forms of D_+, D_-, and $\bar{D}D$ also resemble the lifting and lowering operators and the Hamiltonian of the harmonic oscillator, it is natural to consider the "Grassmann Gaussian"

$$\psi = e^{\lambda\bar\theta\theta} = e^{\lambda\theta^\alpha C_{\alpha\beta}\theta^\beta}, \tag{8.12}$$

the analog of the lowest state of the oscillator. In the next section we pursue this approach by finding all solutions of (8.11) with the aid of the ansatz

$$\psi(x,\theta) = e^{\lambda\bar\theta\theta}F(x). \tag{8.13}$$

IX. IRREDUCIBLE SUPERFIELDS

The reducible scalar superfield in 11-dimensional Kaluza–Klein space may be written as

$$\Phi(\theta,x) = \sum_{m=0}^{32}\theta^{\alpha_1}\cdots\theta^{\alpha_m}F_{\alpha_1\cdots\alpha_m}(x), \tag{9.1}$$

where θ^α is the 32-component anticommuting spinor appropriate to a 10- or 11-dimensional space and $F_{\alpha_1\cdots\alpha_m}$ is a completely antisymmetric multispinor. The space of completely antisymmetric multispinors has dimensionality

$$\sum_{m=0}^{32}\binom{32}{m} = 2^{32}$$

and is highly reducible.

To obtain its irreducible subspaces let us consider, as suggested in (8.13),

$$\psi(x,\theta) = e^{\lambda\bar\theta\theta}F(x). \tag{9.2}$$

Then

$$\bar{D}D\psi(x,\theta) = [64\lambda - \bar\theta\theta(4\lambda^2 - \tfrac14 P^2)]e^{\lambda\bar\theta\theta}F(x).$$

If we impose

$$P^2F = 16\lambda^2F \equiv M^2F, \tag{9.3}$$

$$\lambda = \pm M/4, \tag{9.4}$$

then

$$\bar{D}D\psi_\pm = \bar{D}D\,e^{\pm(M/4)\bar\theta\theta}F = \pm 16M\psi_\pm \tag{9.5}$$

and

$$C_1\psi_\pm(x,\theta) = 0. \tag{9.6}$$

According to Table II, (9.6) implies that all other Casimirs also vanish and that ψ_+ and ψ_- each belong to one of the irreducible representations with highest weight $[0\,0\,0\,0\,0] = [0]$.

Since Q and D anticommute, $Q^{\alpha_1}\cdots Q^{\alpha_j}e^{\pm(M/4)\bar\theta\theta}\times F_{\alpha_1\cdots\alpha_j}(x)$ correspond to the same $\bar{D}D$ eigenvalue as $\dot\psi_\pm$ if $(P^2-M^2)F_{\alpha_1\cdots\alpha_j}(x)=0$.

But

$$Q_\mp e^{\pm(M/4)\bar\theta\theta}=0. \tag{9.7}$$

So we get the two irreducible parts

$$\Phi_{+[0]}=\sum_{j=0}^{16}Q_+^{\alpha_1}\cdots Q_+^{\alpha_j}e^{(M/4)\bar\theta\theta}F_{\alpha_1\cdots\alpha_j}(x), \quad (9.8)_+$$

$$\Phi_{-[0]}=\sum_{j=0}^{16}Q_-^{\alpha_1}\cdots Q_-^{\alpha_j}e^{-(M/4)\bar\theta\theta}F_{\alpha_1\cdots\alpha_j}(x), \quad (9.8)_-$$

which are the most general solutions, respectively, to the equations

$$\bar{D}D\Phi_{+[0]}=16M\Phi_{+[0]}, \quad (9.9)_+$$

$$\bar{D}D\Phi_{-[0]}=-16M\Phi_{-[0]}. \quad (9.9)_-$$

Here the $F_{\alpha_1\cdots\alpha_j}(x)$ are completely antisymmetric multispinors satisfying

$$P^2F_{\alpha_1\cdots\alpha_j}(x)=M^2F_{\alpha_1\cdots\alpha_j}(x), \tag{9.10}$$

and the indices α_k now have only 16 possible values.

The above solutions can also be generated by applying supersymmetry transformations to $\dot\psi_\pm$. The number of components in $\Phi_{\pm[0]}$ is 2^{16}. This is of course, the dimensionality of the irreducible representation [0]. The original reducible representation, on the other hand, is of much greater dimensionality, namely 2^{32}.

To obtain the remaining eigenstates of $\bar{D}D$ we need the lifting and lowering operators. These turn out to be the positive and negative "energy" projections of D

$$D_\pm=\Lambda_\pm i\left(\frac{\partial}{\partial\bar\theta}-\frac{1}{2}\not{P}\theta\right)=\Lambda_\pm i\left(\frac{\partial}{\partial\bar\theta}\mp\frac{1}{2}M\theta\right). \tag{9.11}$$

To establish this point we compute

$$\bar{D}DD^\alpha=D^\alpha\bar{D}D+[\bar{D}D,D^\alpha]=D^\alpha\bar{D}D+2\not{P}^\alpha{}_\gamma D^\gamma.$$

Multiplying by $\Lambda_\pm$ we find

$$\bar{D}DD_\pm=D_\pm\bar{D}D\pm 2MD_\pm. \tag{9.12}$$

Then $D_\pm$ raises the eigenvalues by $\pm 2M$, and of course we have

$$D_\pm e^{\pm(M/4)\bar\theta\theta}=0. \tag{9.13}$$

The expression

$$\sum_{j=0}^{16}Q_+^{\alpha_1}\cdots Q_+^{\alpha_j}D_-^{\beta_1}\cdots D_-^{\beta_m}e^{(M/4)\bar\theta\theta}F_{\alpha_1\cdots\alpha_j;\beta_1\cdots\beta_m}(x) \tag{9.14}$$

is the most general solution to the equation

$$\bar{D}D\Phi=(16-2m)M\Phi,\quad m=0,1,2,\ldots,16, \tag{9.15}$$

and the expression

$$\sum_{j=0}^{16}Q_-^{\alpha_1}\cdots Q_-^{\alpha_j}D_+^{\beta_1}\cdots D_+^{\beta_m}e^{-(M/4)\bar\theta\theta}F_{\alpha_1\cdots\alpha_j;\beta_1\cdots\beta_m}(x) \tag{9.16}$$

solves the equation

$$\bar{D}D\Phi=(-16+2m)M\Phi\quad m=0,1,2,\ldots,16. \tag{9.17}$$

Only one chain must be considered since the other one is redundant. The dimensions of one chain add up to

$$\sum_{j=0}^{16}2^{16}\binom{16}{m}=2^{32},$$

as it should.

As has already been pointed out, the values $m=0,1,2,3,13,14,15,16$ give irreducible superfields. The others need further projection.

X. SIMPLIFIED REPRESENTATIONS

Equation (8.11) shows that there is a factorization that separates the x- and θ-dependent operators in $D_\pm$. We can rewrite (9.11) as

$$D_\pm=i\Lambda_\pm e^{\pm(M/4)\bar\theta\theta}\frac{\partial}{\partial\bar\theta}e^{\mp(M/4)\bar\theta\theta}. \tag{10.1}$$

Similarly for Q,

$$Q_\pm=i\Lambda_\pm e^{\mp(M/4)\bar\theta\theta}\frac{\partial}{\partial\bar\theta}e^{\pm(M/4)\bar\theta\theta}. \tag{10.2}$$

By using these factored representations of D_- and Q_+ one may simplify the expressions for the superfields given in the preceding paragraph:

$$\begin{aligned}\Phi_m^+&=i^p\Lambda^{\alpha_1}_{+\gamma_1}\cdots\Lambda^{\alpha_p}_{+\gamma_p}e^{-(M/4)\bar\theta\theta}\frac{\partial}{\partial\bar\theta_{\gamma_1}}\cdots\frac{\partial}{\partial\bar\theta_{\gamma_p}}e^{(M/4)\bar\theta\theta}\\&\quad\times i^m\Lambda^{\alpha_1}_{-\delta_1}\cdots\Lambda^{\beta_m}_{-\delta_m}e^{-(M/4)\bar\theta\theta}\frac{\partial}{\partial\bar\theta_{\delta_1}}\cdots\frac{\partial}{\partial\bar\theta_{\delta_m}}e^{(M/4)\bar\theta\theta}e^{(M/4)\bar\theta\theta}F_{\alpha_1\cdots\alpha_p;\beta_1\cdots\beta_m}(x)\\&=\sum_{p=0}^{16}i^{p+m}\Lambda^{\alpha_1}_{+\gamma_1}\cdots\Lambda^{\alpha_p}_{+\gamma_p}\Lambda^{\beta_1}_{-\delta_1}\cdots\Lambda^{\beta_m}_{-\delta_m}e^{(M/4)\bar\theta\theta}H^{\gamma_1\cdots\gamma_p\delta_1\cdots\delta_m}(\theta)F_{\alpha_1\cdots\alpha_p;\beta_1\cdots\beta_m}(x),\end{aligned} \tag{10.3}$$

where we shall define the Grassmann–Hermite polynomials by

$$H^{\delta_1\cdots\delta_m}(\theta)=e^{-(M/2)\bar\theta\theta}\frac{\partial}{\partial\bar\theta_{\delta_1}}\cdots\frac{\partial}{\partial\bar\theta_{\delta_m}}e^{(M/2)\bar\theta\theta}. \tag{10.4}$$

These are multinomials in the θ variables and completely antisymmetric in all indices.

Finally

$$\Phi_m^+=e^{(M/4)\bar\theta\theta}\sum_{p=0}^{16}H^{\gamma_1\cdots\gamma_p\delta_1\cdots\delta_m}(\theta)\bar\psi_{\gamma_1\cdots\gamma_p;\delta_1\cdots\delta_m}(x), \tag{10.5}$$

where the $\bar{\psi}_{\gamma_1 \cdots \gamma_p;\delta_1 \cdots \delta_m}$ are Grassmann–Bargmann–Wigner multispinors with the properties

$$\Lambda^{\gamma_s}{}_{-\beta}\, \bar{\psi}_{\gamma_1 \cdots \gamma_s \cdots \gamma_p;\delta_1 \cdots \delta_m} = 0, \tag{10.6}$$

$$\Lambda^{\delta_s}{}_{+\beta}\, \bar{\psi}_{\gamma_1 \cdots \gamma_p;\delta_1 \cdots \delta_s \cdots \delta_m} = 0, \tag{10.7}$$

or alternatively

$$\Lambda^{\beta}{}_{+\gamma_s} \psi^{\gamma_1 \cdots \gamma_s \cdots \gamma_p;\delta_1 \cdots \delta_m} = 0, \tag{10.8}$$

$$\Lambda^{\beta}{}_{-\delta_s} \psi^{\gamma_1 \cdots \gamma_p;\delta_1 \cdots \delta_s \cdots \delta_m} = 0, \tag{10.9}$$

where

$$\psi^{\gamma_1 \cdots \gamma_p;\delta_1 \cdots \delta_m} = \bar{\psi}_{\alpha_1 \cdots \alpha_p;\beta_1 \cdots \beta_m} (C^{-1})^{\alpha_1\gamma_1} \ldots (C^{-1})^{\beta_m\delta_m}. \tag{10.10}$$

The ψ are antisymmetric in the sets $[\gamma_1 \cdots \gamma_p]$ and $[\delta_1 \cdots \delta_m]$ separately. However, only their completely antisymmetric projections contribute to Φ_m^+.

XI. INTEGRALS

The Grassmann–Hermite functions have properties entirely analogous to those of the familiar Hermite functions. To discuss orthogonality properties one may define a generating function

$$g(\theta,t) = \sum_{m=0}^{\infty} \frac{\epsilon(m)}{m!} \bar{t}_{\alpha_1} \cdots \bar{t}_{\alpha_m} H^{\alpha_1 \cdots \alpha_m}(\theta)$$

$$= e^{-(M/2)\bar{\theta}\theta} \sum_{m=0}^{\infty} \frac{1}{m!} \left(\bar{t} \frac{\partial}{\partial \bar{\theta}}\right)^m e^{(M/2)\bar{\theta}\theta}, \tag{11.1}$$

where the t_α anticommute among themselves and with the θ_α, and where

$$\bar{t} \frac{\partial}{\partial \bar{\theta}} = \bar{t}_\alpha \frac{\partial}{\partial \bar{\theta}_\alpha}, \tag{11.2}$$

$$\epsilon(m) = (-)^{(m/2)(m-1)}. \tag{11.3}$$

Then

$$g(\theta,t) = e^{-(M/2)\bar{\theta}\theta} e^{\bar{t}(\partial/\partial\bar{\theta})} e^{(M/2)\bar{\theta}\theta}$$

$$= e^{-(M/2)\bar{\theta}\theta} e^{(M/2)(\bar{\theta}+\bar{t})(\theta+t)} \tag{11.4}$$

or

$$g(\theta,t) = e^{(M/2)(\bar{t}t + 2\bar{t}\theta)}. \tag{11.5}$$

Then the orthogonality properties may be established by consideration of the integral

$$J = \int [d\theta] e^{(M/2)\bar{\theta}\theta} g(\theta,t) g(\theta,s). \tag{11.6}$$

This integral may be expanded according to (11.1) as

$$J = \sum_{p,m} \frac{\epsilon(m)\epsilon(p)}{m!p!} \bar{t}_{\alpha_1} \cdots \bar{t}_{\alpha_p} \bar{s}_{\beta_1} \cdots \bar{s}_{\beta_m}$$

$$\times \int [d\theta] e^{(M/2)\bar{\theta}\theta} H^{\alpha_1 \cdots \alpha_p}(\theta) H^{\beta_1 \cdots \beta_m}(\theta), \tag{11.7}$$

and it may also be evaluated according to (11.5) as

$$J = \int [d\theta] e^{(M/2)\bar{\theta}\theta} e^{(M/2)(\bar{s}s + \bar{t}t + 2\bar{s}\theta + 2\bar{t}\theta)}$$

$$= e^{-M\bar{s}t} \int [d\phi] e^{(M/2)\bar{\phi}\phi} \tag{11.8}$$

where $\phi = \theta + s + t$. But

$$\int [d\phi] e^{(M/2)\bar{\phi}\phi}$$

$$= \int [d\phi] \sum_{j=0}^{16} \frac{1}{j!} \left(\frac{M}{2}\right)^j (\bar{\phi}\phi)^j$$

$$= \sum_{j=0}^{16} \left(\frac{M}{2}\right)^j \frac{1}{j!} \int [d\phi] C_{\alpha_1\beta_1} \cdots C_{\alpha_j\beta_j} \phi^{\alpha_1}\phi^{\beta_1} \cdots \phi^{\alpha_j}\phi^{\beta_j}$$

$$= \left(\frac{M}{2}\right)^{\Lambda/2} \frac{\epsilon(\Lambda)}{(\Lambda/2)!} \mathrm{Pf}(C), \tag{11.9}$$

where the usual rules for integration of Grassmann variables have been used, and Pf(C) is the Pfaffian or square root of the determinant of C:

$$\mathrm{Pf}(C) = (\det C)^{1/2}. \tag{11.10}$$

Then

$$J = e^{-M\bar{s}t} \left(\frac{M}{2}\right)^{\Lambda/2} \frac{\mathrm{Pf}(C)}{(\Lambda/2)!}. \tag{11.11}$$

To compare with (11.7) we expand (11.11). Then

$$J(s,t) = \sum_{m,p} \frac{\epsilon(m)\epsilon(p)}{m!p!} \bar{t}_{\alpha_1} \cdots \bar{t}_{\alpha_p} \bar{s}_{\beta_1} \cdots \bar{s}_{\beta_m} I^{\alpha_1 \cdots \alpha_p \beta_1 \cdots \beta_m}$$

$$= \left(\frac{M}{2}\right)^{\Lambda/2} \frac{\mathrm{Pf}(C)}{(\Lambda/2)!} \sum_j \frac{(-M)^j}{j!} (\bar{t}_{\alpha_1}\bar{s}_{\beta_1}) \cdots (\bar{t}_{\alpha_j}\bar{s}_{\beta_j})$$

$$\times (C^{-1})^{\alpha_1\beta_1} \ldots (C^{-1})^{\alpha_j\beta_j}, \tag{11.12}$$

where

$$I^{\alpha_1 \cdots \alpha_p \beta_1 \cdots \beta_m} = \int [d\theta] e^{(M/2)\bar{\theta}\theta} H^{\alpha_1 \cdots \alpha_p}(\theta) H^{\beta_1 \cdots \beta_m}(\theta). \tag{11.13}$$

There are no terms on the right of (11.12) for which $m \neq p$. Therefore these functions are orthogonal:

$$I^{\alpha_1 \cdots \alpha_p \beta_1 \cdots \beta_m} = 0, \quad m \neq p. \tag{11.14}$$

If $m = p$,

$$I^{\alpha_1 \cdots \alpha_p \beta_1 \cdots \beta_m}$$

$$= A_m M^{\Lambda/2 + m} \mathrm{Pf}(C) (C^{-1})^{[\alpha_1[\beta_1} \ldots (C^{-1})^{\alpha_m]\beta_m]}, \tag{11.15}$$

where

$$A_m = \frac{(-)^m \epsilon(m) m!}{2^{\Lambda/2}(\Lambda/2)!}$$

and the indices α and β are antisymmetrized separately.

XII. LAGRANGIANS

To see how one obtains Lagrangians corresponding to (10.5), consider for instance the invariant θ integral:

$$I_m(x) = \int [d\theta]\, \Phi_m^+ \Phi_m^+ = \sum_q \bar{\psi}_{\alpha_1 \cdots \alpha_q; \beta_1 \cdots \beta_m}(x) \bar{\psi}_{\gamma_1 \cdots \gamma_q; \delta_1 \cdots \delta_m}(x)$$

$$\times \int [d\theta]\, e^{(M/2)\bar{\theta}\theta} H^{\alpha_1 \cdots \alpha_q \beta_1 \cdots \beta_m}(\theta) H^{\gamma_1 \cdots \gamma_q \delta_1 \cdots \delta_m}(\theta), \tag{12.1}$$

where

$$\bar{\psi}_{\alpha_1 \cdots \alpha_q; \beta_1 \cdots \beta_m}(x) = i^{q+m} \Lambda^{\gamma_1}_{+\ \alpha_1} \cdots \Lambda^{\gamma_q}_{+\ \alpha_q} \Lambda^{\delta_1}_{-\ \beta_1} \cdots \Lambda^{\delta_m}_{-\ \beta_m} \overline{F}_{\gamma_1 \cdots \gamma_q; \delta_1 \cdots \delta_m}(x).$$

Then

$$I_m = \sum_q B_{q+m}(-)^{q+m} \bar{\psi}_{\alpha_1 \cdots \alpha_q; \beta_1 \cdots \beta_m}(x) \psi^{[\alpha_1 \cdots \alpha_q; \beta_1 \cdots \beta_m]}(x), \tag{12.2}$$

where

$$B_m = A_m M^{16+m}\,\mathrm{Pf}(C).$$

As previously remarked, Φ_m^+ is irreducible only if $m<3$ or $m>13$. Let us consider $m = 1$:

$$I_1(x) = \sum_q (-i)^{q+1} B_{q+1} (\Lambda^{\gamma_1}_{+\ \alpha_1} \cdots \Lambda^{\gamma_q}_{+\ \alpha_q} \Lambda^{\delta}_{-\ \beta} \overline{F}_{\gamma_1 \cdots \gamma_q; \delta}(x)) \psi^{[\alpha_1 \cdots \alpha_q; \beta]}(x). \tag{12.3}$$

Up to a complete divergence we can write

$$\begin{aligned} I_1(x) &= \sum_q (-i)^{q+1} B_{q+1} \overline{F}_{\gamma_1 \cdots \gamma_q; \delta}(x) \Lambda^{\gamma_1}_{-\ \alpha_1} \cdots \Lambda^{\gamma_q}_{-\ \alpha_q} \Lambda^{\delta}_{+\ \beta} \psi^{[\alpha_1 \cdots \alpha_q; \beta]}(x) \\ &= -\sum_q \sum_{j=1}^q (-i)^{q+1} B_{q+1} \left(\frac{q}{q+1}\right) \overline{F}_{\gamma_1 \cdots \gamma_q; \delta}(x) \Lambda^{\gamma_1}_{-\ [\alpha_1} \cdots \Lambda^{\gamma_{j-1}}_{-\ \alpha_{j-1}} \Lambda^{\delta}_{+\ \alpha_j} \Lambda^{\gamma_{j+1}}_{-\ \alpha_{j+1}} \cdots \Lambda^{\gamma_q}_{-\ \alpha_q]} \Lambda^{\gamma_j}_{-\ \beta} \psi^{\alpha_1 \cdots \alpha_q; \beta}(x) \\ &\quad + \sum_q (-i)^{q+1} B_{q+1} \left(\frac{1}{q+1}\right) \overline{F}_{\gamma_1 \cdots \gamma_q; \delta}(x) \Lambda^{\gamma_1}_{-\ \alpha_1} \cdots \Lambda^{\gamma_q}_{-\ \alpha_q} \Lambda^{\delta}_{+\ \beta} \psi^{\alpha_1 \cdots \alpha_q; \beta}(x) \\ &= -\sum_q \sum_{j=1}^q B_{q+1} \left(\frac{q}{q+1}\right) \overline{F}_{\gamma_1 \cdots \gamma_q; \delta}(x) (\Lambda^2_-)^{\gamma_1}{}_{\beta_1} \cdots (\Lambda^2_-)^{\gamma_{j-1}}{}_{\beta_{j-1}} \\ &\quad \times (\Lambda^2_-)^{\gamma_{j+1}}{}_{\beta_{j+1}} \cdots (\Lambda^2_-)^{\gamma_q}{}_{\beta_q} (P^2 - M^2) \delta^{\delta}{}_{\beta_j} (P^2 - M^2) \delta^{\gamma_j}{}_{\alpha} F^{\beta_1 \cdots \beta_j; \alpha}(x) \\ &\quad + \sum_q B_{q+1} \left(\frac{1}{q+1}\right) \overline{F}_{\gamma_1 \cdots \gamma_q; \delta}(x) (\Lambda^2_-)^{\gamma_1}{}_{\beta_1} \cdots (\Lambda^2_-)^{\gamma_q}{}_{\beta_q} (\Lambda^2_+)^{\delta}{}_{\alpha} F^{\beta_1 \cdots \beta_j; \alpha}(x). \end{aligned} \tag{12.4}$$

If we use the irreducibility condition at this point, we just get

$$I_1(x) = \sum_q B_{q+1} \left(\frac{1}{q+1}\right) \overline{F}_{\gamma_1 \cdots \gamma_q; \delta}(x) \Lambda^{\gamma_1}_{-\ \alpha_1} \cdots \Lambda^{\gamma_q}_{-\ \alpha_q} \Lambda^{\delta}_{+\ \beta} F^{\alpha_1 \cdots \alpha_q; \beta}(x). \tag{12.5}$$

One can of course redefine fields in many ways to give a different form to the Lagrangian. For instance, with $\bar{\psi}_{\alpha_1 \cdots \alpha_q; \beta} = \Lambda^{\delta}_{-\ \beta} \bar{\chi}_{\alpha_1 \cdots \alpha_q; \delta}$ we get simply

$$I_1(x) = \sum_q \frac{B_{q+1}}{q+1} \bar{\chi}_{\alpha_1 \cdots \alpha_q; \beta} \Lambda^{\beta}_{+\ \delta} \chi^{\alpha_1 \cdots \alpha_q; \delta}, \tag{12.6}$$

after using the irreducibility condition.

XIII. REMARKS

In a more satisfactory formulation,[5] one writes the Lagrangian $\mathscr{L} = \int [d\theta] \Phi \Pi \Phi$, where Π is a projection operator such that the irreducibility conditions are derived from the superfield equation of motion, rather than being imposed by hand. In that way, the dimensions of Π are dictated by the leading physical field, which we have not selected here, and the fact that the dimensions of $\mathscr{L}$ are $(\text{length})^{-4}$.

There are two ways of introducing interactions into this noninteracting Lagrangian. In the first method the group is regarded as global or rigid. One may then form invariant interactions from products of the irreducible superfields that we have found. The simplest of these models correspond to nonlinear theories in which the interaction is of fourth degree. These cases are analogs of the nonlinear scalar and spinor fields.

If there are interaction terms, then of course the individual fields no longer satisfy the Klein–Gordon equation and by Sec. IX these superfields are no longer irreducible. However the field algebra still will close under supersymmetry transformations so that the formulation of the interacting theory is still off shell.

In the second and more fundamental approach to interactions, the group is regarded as local and the interactions

TABLE IV. Coefficients for 11 dimensions in Eq. (A8).

$C_p^{(k)}$	$p \backslash k$	0	1	2	3	4	5	G_p	E_p
	0	1	11	− 110	− 990	7920	55440	1024	− 32
	1	1	− 9	− 70	450	2160	− 5040	0	32
	2	1	7	− 38	− 126	− 144	− 5040	0	32
	3	1	− 5	− 14	− 30	− 528	1680	0	− 32
	4	1	3	2	66	− 144	1680	0	− 32
	5	1	− 1	10	− 30	240	− 1200	0	32

Here $G_p = \sum_{k=0}^{5} \frac{\epsilon(k)}{k!} C_p^{(k)}$ and $E_p = \sum_{k=0}^{5} \frac{(-)^k}{k!} C_p^{(k)}$.

appear via the displacement field or connection. To implement this approach one gauges the graded Lie algebra.

Although the analysis of massive representations given in this paper is of interest in itself, it was intended to be preliminary to an attack on the massless case so that the transition to local symmetries could be effected in the conventional manner.

In descending to space-time there are various possibilities as illustrated by the different versions of Cremmer and Julia[6] and of de Wit and Nicolai[7] and the discussion of Ref. 1.

ACKNOWLEDGMENT

We acknowledge helpful discussions with V. S. Varadarajan.

APPENDIX A: THE DIRAC ALGEBRA

The Dirac algebra in n dimensions is formed by $2^{[n/2]} \times 2^{[n/2]}$ matrices satisfying

$$\{\Gamma_A, \Gamma_B\} = 2\eta_{AB}, \quad A,B = 0,1,\ldots,n-1, \tag{A1}$$

where

$$\eta_{AB} = \mathrm{diag}(+\ -\cdots-).$$

We can define totally antisymmetric Γ-tensors by

$$\Gamma_{A_1\cdots A_p} = \tfrac{1}{2}(\Gamma_{A_1\cdots A_{p-1}}\Gamma_{A_p} - (-)^p\Gamma_{A_p}\Gamma_{A_1\cdots A_{p-1}}). \tag{A2}$$

In odd dimensions the Γ-tensors of rank $\leqslant [n/2]$ are sufficient to span the $2^{[n/2]} \times 2^{[n/2]}$ matrix space since

$$\Gamma_0\Gamma_1\Gamma_2\cdots\Gamma_{n-1} = \alpha I \quad [\alpha^2 = (-)^{[n/2]}] \tag{A3}$$

in odd dimensions.

From (A1) and (A2) we can derive formulas for all commutators and anticommutators of Γ-tensors:

$$\begin{aligned}
&\tfrac{1}{2}\{\Gamma^{M_1\cdots M_{2k}}, \Gamma_{N_1\cdots N_m}\} \\
&\quad = \sum_{j=0}^{a} (-)^j \frac{m!}{(m-2j)!}\binom{2k}{2j} \\
&\qquad \times \delta^{[M_1}_{[N_1}\cdots\delta^{M_{2j}}_{N_{2j}}\Gamma^{M_{2j+1}\cdots M_{2k}]}{}_{N_{2j+1}\cdots N_m]}, \\
&\qquad \text{with } a = \mathrm{Min}\{k,[m/2]\},
\end{aligned} \tag{A4}$$

$$\begin{aligned}
&\tfrac{1}{2}\{\Gamma^{M_1\cdots M_{2k+1}}, \Gamma_{N_1\cdots N_{2p+1}}\} \\
&\quad = \sum_{j=0}^{a} (-)^j \frac{(2p+1)!}{[2(p-j)]!}\binom{2k+1}{2j+1} \\
&\qquad \times \delta^{[M_1}_{[N_1}\cdots\delta^{M_{2j+1}}_{N_{2j+1}}\Gamma^{M_{2j+2}\cdots M_{2k+1}]}{}_{N_{2j+2}\cdots N_{2p+1}]}, \\
&\qquad a = \mathrm{Min}\{k,p\},
\end{aligned} \tag{A5}$$

$$\begin{aligned}
&\tfrac{1}{2}[\Gamma^{M_1\cdots M_{2k}}, \Gamma_{N_1\cdots N_m}] \\
&\quad = \sum_{j=0}^{a} (-)^{j+1} \frac{m!}{(m-2j-1)!}\binom{2k}{2j+1} \\
&\qquad \times \delta^{[M_1}_{[N_1}\cdots\delta^{M_{2j+1}}_{N_{2j+1}}\Gamma^{M_{2j+2}\cdots M_{2k}]}{}_{N_{2j+2}\cdots N_m]}, \\
&\qquad a = \mathrm{Min}\{k-1,[(m-1)/2]\},
\end{aligned} \tag{A6}$$

$$\begin{aligned}
&\tfrac{1}{2}[\Gamma^{M_1\cdots M_{2k+1}}, \Gamma_{N_1\cdots N_{2p+1}}] \\
&\quad = \sum_{j=0}^{a} (-)^j \frac{(2p+1)!}{(2p+1-2j)!}\binom{2k+1}{2j} \\
&\qquad \times \delta^{[M_1}_{[N_1}\cdots\delta^{M_{2j}}_{N_{2j}}\Gamma^{M_{2j+1}\cdots M_{2k+1}]}{}_{N_{2j+2}\cdots N_{2p+1}]}, \\
&\qquad a = \mathrm{Min}(k,p).
\end{aligned} \tag{A7}$$

These formulas can be used to decompose products of Γ-tensors (or D-tensors as in Sec. VII).

Constants very useful in calculations are $C_p^{(k)}$ defined by

$$\Gamma_{A_1\cdots A_k}\Gamma_{B_1\cdots B_p}\Gamma^{A_1\cdots A_k} = C_p^{(k)}\Gamma_{B_1\cdots B_p}. \tag{A8}$$

They satisfy the recursion relations

$$\begin{aligned}
C_m^{(k)} &= (-)^{k-1}C_m^{(k-1)}C_m^{(1)} \\
&\qquad + (k-1)(n-k+2)C_m^{(k-2)}, \quad k \geqslant 1, \\
C_m^{(0)} &= 1, \quad C_m^{(1)} = (-)^m(n-2m),
\end{aligned} \tag{A9}$$

$$C_0^{(k)} = (-)^{(k/2)(k-1)}k!\binom{n}{k} = \epsilon(k)k!\binom{n}{k}. \tag{A10}$$

In 11 dimensions we have the results in Table IV.

APPENDIX B: FIERZ REARRANGEMENTS

In general, if Q_1, Q_2, Q_3 and Q_4 are Majorana spinors, and

$$\{Q_i^\alpha, Q_j^\beta\} = 2\xi_{ij}^{\alpha\beta}, \tag{B1}$$

then one has the identity

$$\begin{aligned}
&\overline{Q}_1 M Q_2 \overline{Q}_3 N Q_4 \\
&\quad = -\frac{1}{\Lambda}\sum_j \lambda(j)\overline{Q}_1 M O_j N Q_4 \overline{Q}_3 O_j Q_2 \\
&\qquad + \frac{1}{\Lambda}\sum_j \lambda(j)\overline{Q}_1 M O_j N Q_4 \,\mathrm{Tr}[O_j(2\xi_{23})C] \\
&\qquad - \frac{1}{\Lambda}\sum_j \lambda(j)\lambda'(j)\overline{Q}_1 M O_j N(2\xi_{24})C O_j Q_3 \\
&\qquad + \frac{1}{\Lambda}\sum_j \lambda(j)\overline{Q}_1 M O_j N(2\xi_{34})C O_j Q_2,
\end{aligned} \tag{B2}$$

where $\{O_j\}$ is an orthogonal basis in the space of $\Lambda \times \Lambda$ matrices, satisfying

$$\mathrm{Tr}[O_i O_j] = \Lambda\lambda(j)\delta_{ij}, \qquad \text{(B3)}$$

$$(CO_j)^T = \lambda'(j)CO_j. \qquad \text{(B4)}$$

For our purposes all the spinors are the same (namely D) and the basis is given by Γ-tensors. We have then in 11 dimensions:

$$\bar{D}MD\bar{D}ND = -\frac{1}{32}\sum_{j=0}^{5}\frac{\epsilon(j)}{j!}\bar{D}M\Gamma^{B_1\cdots B_j}ND\bar{D}\Gamma_{B_1\cdots B_j}D - \bar{D}M\not{P}ND$$
$$-\frac{1}{32}\sum_{j=0}^{5}\frac{1}{j!}[(-)^j+\epsilon(j)]\bar{D}M\Gamma^{B_1\cdots B_j}N\not{P}\Gamma_{B_1\cdots B_j}D, \qquad \text{(B5)}$$

where the $1/j!$ factors are to compensate overcounting. Since D is a Majorana spinor satisfying

$$\{D^\alpha, D^\beta\} = -(\not{P}C^{-1})^{\alpha\beta},$$

we have

$$\bar{D}\Gamma_{A_1\cdots A_j}D = 0, \quad \text{for } j = 2,5, \qquad \text{(B6)}$$

$$\bar{D}\Gamma_A D = -16P_A. \qquad \text{(B7)}$$

Now we can "Fierz":

$$\bar{D}\not{P}\Gamma_{AB}D\bar{D}\Gamma^{AB}\not{P}D = -\frac{1}{32}\sum_{j=0}^{5}\frac{\epsilon(j)}{j!}\bar{D}\not{P}\Gamma_{AB}\Gamma^{C_1\cdots C_j}\Gamma^{AB}\not{P}D\bar{D}\Gamma_{C_1\cdots C_j}D$$
$$-\bar{D}\not{P}\Gamma_{AB}\not{P}\Gamma^{AB}\not{P}D - \frac{1}{32}\sum_{j=0}^{5}\frac{\epsilon(j)}{j!}\bar{D}\not{P}\Gamma_{AB}\Gamma^{C_1\cdots C_j}\Gamma^{AB}\not{P}\not{P}\Gamma_{C_1\cdots C_j}D. \qquad \text{(B8)}$$

We have made use of

$$-\frac{1}{32}\sum_{j=0}^{5}\frac{\epsilon(j)}{j!}\Gamma^{C_1\cdots C_j}\Gamma^{A_1\cdots A_k}\Gamma_{C_1\cdots C_j} = 0, \quad \text{if } k \neq 0$$

(see G_k in Appendix A).

After a little algebra one arrives at the equation

$$\left(1+\frac{C_3^{(2)}}{32}\right)\bar{D}\not{P}\Gamma_{AB}D\bar{D}\Gamma^{AB}\not{P}D$$
$$= -\frac{C_0^{(2)}}{32}P^2(\bar{D}D)^2 + 8P^4\left(C_1^{(2)} + \frac{1}{16}\sum_{j=0}^{5}\frac{(-)^j}{j!}C_j^{(2)}C_0^{(j)}\right)$$
$$-\frac{P^2}{32}\left(\frac{C_3^{(2)}}{3!}\bar{D}\Gamma^{B_1B_2B_3}D\bar{D}\Gamma_{B_1B_2B_3}D + \frac{C_4^{(2)}}{4!}\bar{D}\Gamma^{B_1\cdots B_4}D\bar{D}\Gamma_{B_1\cdots B_4}D\right) - \frac{C_4^{(2)}}{16}\frac{1}{3!}\bar{D}\not{P}\Gamma^{B_1B_2B_3}D\bar{D}\Gamma_{B_1B_2B_3}\not{P}D. \qquad \text{(B9)}$$

In the same way one can obtain the equation

$$\left(1+\frac{C_4^{(3)}}{3!16}\right)\bar{D}\not{P}\Gamma^{B_1B_2B_3}D\bar{D}\Gamma_{B_1B_2B_3}\not{P}D$$
$$= -\frac{C_0^{(3)}}{32}P^2(\bar{D}D)^2 + 8P^4\left(C_1^{(3)} + \frac{1}{16}\sum_{j=0}^{5}\frac{(-)^j}{j!}C_j^{(3)}C_0^{(j)}\right)$$
$$-\frac{P^2}{32}\left(\frac{C_3^{(3)}}{3!}\bar{D}\Gamma^{B_1B_2B_3}D\bar{D}\Gamma_{B_1B_2B_3}D + \frac{C_4^{(3)}}{4!}\bar{D}\Gamma^{B_1\cdots B_4}D\bar{D}\Gamma_{B_1\cdots B_4}D\right) + \frac{C_3^{(3)}}{32}\bar{D}\not{P}\Gamma_{AB}D\bar{D}\Gamma^{AB}\not{P}D. \qquad \text{(B10)}$$

With these two equations we can solve $\bar{D}\not{P}\Gamma_{AB}\ D\bar{D}\Gamma^{AB}\not{P}D$ in terms of $(\bar{D}D)^2$, P^2, $\bar{D}\Gamma^{B_1B_2B_3}D\ \bar{D}\Gamma_{B_1B_2B_3}D$, and $\bar{D}\Gamma^{B_1\cdots B_4}D\ \bar{D}\Gamma_{B_1\cdots B_4}D$.

We can continue this process with

$$\bar{D}\Gamma^{B_1\cdots B_k}D\bar{D}\Gamma_{\beta_1\cdots B_k}D$$
$$= -\frac{1}{32}\sum_{j=0}^{5}\frac{\epsilon(j)}{j!}C_j^{(k)}\bar{D}\Gamma^{A_1\cdots A_j}D\bar{D}\Gamma_{A_1\cdots A_j}D - C_1^{(k)}\bar{D}\not{P}D - \frac{1}{32}\sum_{j=0}^{5}\frac{(-)^j}{j!}C_j^{(k)}C_1^{(j)}\bar{D}\not{P}D. \qquad \text{(B11)}$$

For $k = 3,4$ we get two equations that allow us to solve for $\bar{D}\Gamma^{B_1B_2B_3}D\ \bar{D}\Gamma_{B_1B_2B_3}D$ and $\bar{D}\Gamma^{B_1\cdots B_4}\mathrm{D}\ \bar{D}\Gamma_{B_1\cdots B_4}D$ in terms of $(\bar{D}D)^2$ and P^2. After the necessary algebra one gets the final result

$$C_1 = \tfrac{3}{8}(1/P^2)[(\bar{D}D)^2 - 16^2 P^2]. \tag{B12}$$

[1] J. Ellis, M. K. Gaillard, and B. Zumino, Acta Phys. Pol. B **13**, 253 (1982). Here is a discussion of some of the physical difficulties that must be faced.

[2] A. Salam and J. Strathdee, Nucl. Phys. B **76**, 477 (1974).

[3] A. M. Perelomov and V. S. Popov, JETP Lett. **1**, 160 (1965); **2**, 20 (1965); Sov. Mat. Dokl. **8**, 631 (1967).

[4] E. Sokatchev, Nucl. Phys. B **99**, 96 (1975); J. Kim, J. Math. Phys. **25**, 2037 (1984); J. G. Taylor, Nucl. Phys. B **169**, 484 (1980).

[5] V. I. Ogievetsky and E. Sokatchev, J. Phys. A **10**, 2021 (1977).

[6] E. Cremmer and B. Julia, Nucl. Phys. B **159**, 141 (1979).

[7] B. de Wit and H. Nicolai, Nucl. Phys. B **208**, 323 (1982).

Superfield and irreducible superfield structure

M. Villasante
Department of Physics, University of California, Los Angeles, California 90024

(Received 14 November 1985; accepted for publication 5 February 1986)

Field contents for scalar superfields in different numbers of dimensions are tabulated. Tables of field contents for some irreducible superfields included in the scalar superfield in 11 dimensions are also given.

I. INTRODUCTION

The formulation of supersymmetric Kaluza–Klein theories would in principle demand knowledge of the contents of general superfields both in terms of irreducible superfields and of ordinary fields. Extensive tables of field contents for the scalar superfield in various numbers of dimensions have been given in Ref. 1. Unfortunately some of those tables contain errors and need revision while others must be completed. In Sec. II we give the correct results for the tables that need to be changed.

TABLE I. $D = 8$ scalar superfield.

θ sector	SO(7,1) representations
θ^0	[0]
$\theta^{(+)}$	$[\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$
$\theta^{(+)2}$	[1 1]
$\theta^{(+)}\theta^{(-)}$	[1 1 1],[1]
$\theta^{(+)3}$	$[\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}-\tfrac{1}{2}]$
$\theta^{(+)2}\theta^{(-)}$	$[\tfrac{3}{2}\tfrac{3}{2}\tfrac{1}{2}-\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}-\tfrac{1}{2}]$
$\theta^{(+)4}$	[2],[1 1 1 − 1]
$\theta^{(+)3}\theta^{(-)}$	[2 1 1 − 1],[2 1],[1 1 1],[1]
$\theta^{(+)2}\theta^{(-)2}$	[2 2],[2 1 1],[2],[1 1 1 1], [1 1 1 − 1],[1 1],[0]
$\theta^{(+)5}$	$[\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}-\tfrac{1}{2}]$
$\theta^{(+)4}\theta^{(-)}$	$[\tfrac{5}{2}\tfrac{1}{2}\tfrac{1}{2}-\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{3}{2}\tfrac{3}{2}-\tfrac{3}{2}]$, $[\tfrac{3}{2}\tfrac{3}{2}\tfrac{1}{2}-\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}-\tfrac{1}{2}]$
$\theta^{(+)3}\theta^{(-)2}$	$[\tfrac{5}{2}\tfrac{3}{2}\tfrac{1}{2}-\tfrac{1}{2}]$, $[\tfrac{5}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{3}{2}\tfrac{3}{2}-\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $2[\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}-\tfrac{1}{2}]$, $[\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$
$\theta^{(+)6}$	[1 1]
$\theta^{(+)5}\theta^{(-)}$	[2 1 1 − 1],[2 1],[1 1 1],[1]
$\theta^{(+)4}\theta^{(-)2}$	[3 1],[2 2 1 − 1],2[2 1 1],[2] [1 1 1 − 1],2[1 1]
$\theta^{(+)3}\theta^{(-)3}$	[3 1 1],[3],[2 2 1],[2 1 1 1], [2 1 1 − 1],2[2 1],2[1 1 1],[1]
$\theta^{(+)7}$	$[\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$
$\theta^{(+)6}\theta^{(-)}$	$[\tfrac{3}{2}\tfrac{3}{2}\tfrac{1}{2}-\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}-\tfrac{1}{2}]$
$\theta^{(+)5}\theta^{(-)2}$	$[\tfrac{5}{2}\tfrac{3}{2}\tfrac{1}{2}-\tfrac{1}{2}]$, $[\tfrac{5}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{3}{2}\tfrac{3}{2}-\tfrac{1}{2}]$ $[\tfrac{3}{2}\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $2[\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}-\tfrac{1}{2}]$, $[\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$
$\theta^{(+)4}\theta^{(-)3}$	$[\tfrac{7}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{5}{2}\tfrac{3}{2}\tfrac{3}{2}-\tfrac{1}{2}]$, $[\tfrac{5}{2}\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $2[\tfrac{5}{2}\tfrac{1}{2}\tfrac{1}{2}-\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{3}{2}\tfrac{3}{2}\tfrac{1}{2}]$, $2[\tfrac{3}{2}\tfrac{3}{2}\tfrac{1}{2}-\tfrac{1}{2}]$, $2[\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}-\tfrac{1}{2}]$
$\theta^{(+)8}$	[0]
$\theta^{(+)7}\theta^{(-)}$	[1 1 1],[1]
$\theta^{(+)6}\theta^{(-)2}$	[2 2],[2 1 1],[2],[1 1 1 1], [1 1 1 − 1],[1 1],[0]
$\theta^{(+)5}\theta^{(-)3}$	[3 1 1],[3],[2 2 1],[2 1 1 1], [2 1 1 − 1],2[2 1],2[1 1 1],[1]
$\theta^{(+)4}\theta^{(-)4}$	[4],[3 1 1 1],[3 1 1 − 1],[3 1], [2 2 2],[2 2],3[2 1 1],2[2], [1 1 1 1],[1 1 1 − 1],[1 1],[0]

The box $\theta^{(+)m}\theta^{(-)n}$ gives the SO(7,1) representations of the field $W_{\alpha_1\cdots\alpha_m\beta_1\cdots\beta_n}(x)$ in the expansion

$$\Phi(x,\theta^{(+)},\theta^{(-)}) = \sum_{m,n=0}^{8} \theta^{(+)\alpha_1}\cdots\theta^{(+)\alpha_m}\theta^{(-)\beta_1}\cdots\theta^{(-)\beta_n} W_{\alpha_1\cdots\alpha_m\beta_1\cdots\beta_n}(x) .$$

In Ref. 2 the analysis of the massive irreducible representations of the super-Poincaré algebra in higher dimensions was undertaken and a complete decomposition of the scalar superfield in 11 dimensions in its irreducible components was achieved. In Sec. III we will reproduce the table of these irreducible components and give the field content for the three smallest ones. What we search for are the representations corresponding to the massive counterpart of the supergravity multiplet in 11 dimensions,[3] which indeed appears later in Table VII.

II. FIELD CONTENTS OF SCALAR SUPERFIELDS

In the first three tables we follow the structure of Ref. 1, listing the field content as representations of SO$(d-1,1)$ for each θ sector. We list the irreducible representations by their highest weight vectors, omitting the zero components.

Table I shows the field content of the scalar superfield in eight dimensions. We list only the sectors $\theta^{(+)m}\theta^{(-)n}$ with $m \geqslant n$, $\theta^{(+)}$ and $\theta^{(-)}$ being the two Weyl projections of θ. The terms $\theta^{(+)n}\theta^{(-)m}$ can be obtained from $\theta^{(+)m}\theta^{(-)n}$ by mirror conjugation[1] (the representation $\bar{\lambda} = [\lambda_1\lambda_2\cdots\lambda_{p-1}-\lambda_p]$ is the mirror-conjugated representation of the irreducible representation $\lambda = [\lambda_1\lambda_2\cdots\lambda_{p-1}\lambda_p]$ of SO$(2p)$). We do not list the cases

TABLE II. $D = 9$ scalar superfield.

θ sector	SO(8,1) representations
θ^0	[0]
θ	$[\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$
θ^2	[1 1 1],[1 1]
θ^3	$[\tfrac{3}{2}\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$
θ^4	[2 2],[2 1 1 1],[2 1],[2],[1 1 1 1]
θ^5	$[\tfrac{5}{2}\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{5}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{3}{2}\tfrac{3}{2}\tfrac{3}{2}]$, $[\tfrac{3}{2}\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$
θ^6	[3 1 1],[3 1],[2 2 1 1],[2 1 1 1],[2 1 1],[2 1], [1 1 1],[1 1]
θ^7	$[\tfrac{7}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{5}{2}\tfrac{3}{2}\tfrac{3}{2}\tfrac{1}{2}]$, $[\tfrac{5}{2}\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{5}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{3}{2}\tfrac{3}{2}\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{3}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$, $[\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}]$
θ^8	[4],[3 1 1 1],[3 1 1],[3],[2 2 2],[2 2 1],[2 2], [2 1 1 1],[2 1 1],[2],[1 1 1 1],[1 1 1],[1],[0]

The box θ^n shows the SO(8,1) representations of the field $W_{\alpha_1\cdots\alpha_n}(x)$ in the expansion

$$\Phi(x,\theta) = \sum_{n=0}^{16} \theta^{\alpha_1}\cdots\theta^{\alpha_n} W_{\alpha_1\cdots\alpha_n}(x) .$$

0022-2488/86/061607-05$02.50

TABLE III. $D = 11$ scalar superfield.

θ sector	SO(10,1) representations
θ^0	[0]
θ	$[\frac12\frac12\frac12\frac12\frac12]$
θ^2	[1 1 1 1],[1 1 1] + θ^0
θ^3	$[\frac32\frac32\frac32\frac12\frac12], [\frac32\frac32\frac12\frac12\frac12] + \theta$
θ^4	[2 2 2],[2 2 1 1 1],[2 2 1],[2 2],[2 1 1 1 1] + θ^2
θ^5	$[\frac52\frac52\frac32\frac12\frac12], [\frac52\frac52\frac12\frac12\frac12], [\frac52\frac32\frac32\frac32\frac32], [\frac52\frac32\frac32\frac12\frac12], [\frac52\frac32\frac12\frac12\frac12], [\frac32\frac32\frac32\frac32\frac32] + \theta^3$
θ^6	[3 3 1 1],[3 3 1],[3 2 2 1 1],[3 2 1 1 1], [3 2 1 1],[3 2 1],[3 1 1 1],[3 1 1],[2 2 2 2 2], [2 2 2 1 1],[2 2 1 1 1] + θ^4
θ^7	$[\frac72\frac72\frac12\frac12\frac12], [\frac72\frac52\frac32\frac32\frac12], [\frac72\frac52\frac32\frac12\frac12], [\frac72\frac52\frac12\frac12\frac12], [\frac72\frac32\frac32\frac32\frac12], [\frac72\frac32\frac32\frac12\frac12], [\frac72\frac32\frac12\frac12\frac12], [\frac72\frac12\frac12\frac12\frac12], [\frac52\frac52\frac52\frac32\frac32], [\frac52\frac52\frac32\frac32\frac32], [\frac52\frac52\frac32\frac32\frac12], [\frac52\frac52\frac32\frac12\frac12], [\frac52\frac32\frac32\frac32\frac12], [\frac52\frac32\frac32\frac12\frac12] + \theta^5$
θ^8	[4 4],[4 3 1 1 1],[4 3 1 1],[4 3],[4 2 2 2], [4 2 2 1],[4 2 2],[4 2 1 1 1],[4 2 1 1],[4 2], [4 1 1 1 1],[4 1 1 1],[4 1],[4],[3 3 2 2 1], [3 3 2 1 1],[3 3 1 1 1],[3 3 1 1],[3 2 2 2 1], [3 2 2 2],[3 2 2 1 1],[3 2 2 1],[3 2 2],[3 2 1 1 1], [3 2 1 1],[3 1 1 1 1],[3 1 1 1],[2 2 2 2],[2 2 2 1], [2 2 2] + θ^6
θ^9	$[\frac92\frac72\frac12\frac12\frac12], [\frac92\frac52\frac32\frac32\frac12], [\frac92\frac52\frac32\frac12\frac12], [\frac92\frac52\frac12\frac12\frac12], [\frac92\frac32\frac32\frac32\frac12], [\frac92\frac32\frac32\frac12\frac12], [\frac92\frac32\frac12\frac12\frac12], [\frac92\frac12\frac12\frac12\frac12], [\frac72\frac72\frac32\frac32\frac32], [\frac72\frac72\frac32\frac32\frac12], [\frac72\frac72\frac12\frac12\frac12], [\frac72\frac52\frac52\frac52\frac12], [\frac72\frac52\frac52\frac32\frac12], [\frac72\frac52\frac52\frac12\frac12], [\frac72\frac52\frac32\frac32\frac32], 2[\frac72\frac52\frac32\frac32\frac12], [\frac72\frac52\frac32\frac12\frac12], [\frac72\frac52\frac12\frac12\frac12], [\frac72\frac32\frac32\frac32\frac32], 2[\frac72\frac32\frac32\frac32\frac12], [\frac72\frac32\frac32\frac12\frac12], [\frac72\frac32\frac12\frac12\frac12], [\frac72\frac12\frac12\frac12\frac12], [\frac52\frac52\frac52\frac52\frac12], [\frac52\frac52\frac52\frac32\frac12], [\frac52\frac52\frac52\frac12\frac12], [\frac52\frac52\frac32\frac32\frac12], [\frac52\frac52\frac32\frac12\frac12], [\frac52\frac32\frac32\frac32\frac12], [\frac52\frac32\frac32\frac12\frac12] + \theta^7$
θ^{10}	[5 3 1 1],[5 3 1],[5 2 2 1 1],[5 2 1 1 1],[5 2 1 1], [5 2 1],[5 1 1 1],[5 1 1],[4 4 1 1 1],[4 3 2 2 1], [4 3 2 2],[4 3 2 1 1],[4 3 2 1],[4 3 1 1 1], [4 3 1 1],[4 3 1],[4 2 2 2 1],[4 2 2 2],2[4 2 2 1 1], [4 2 2 1],2[4 2 1 1 1],[4 2 1 1],[4 2 1],[4 1 1 1 1], [4 1 1 1],[4 1 1],[3 3 3 3],[3 3 3 2],[3 3 3 1], [3 3 3],[3 3 2 2 1],[3 3 2 2],[3 3 2 1 1],[3 3 2 1], [3 3 1 1],[3 3 1],[3 2 2 2 1],[3 2 2 2],2[3 2 2 1 1], [3 2 2 1],[3 2 1 1 1],[3 2 1 1],[3 2 1],[3 1 1 1], [3 1 1],[2 2 2 1 1],[2 2 1 1 1] + θ^8
θ^{11}	$[\frac{11}{2}\frac52\frac32\frac12\frac12], [\frac{11}{2}\frac52\frac12\frac12\frac12], [\frac{11}{2}\frac32\frac32\frac32\frac32], [\frac{11}{2}\frac32\frac32\frac12\frac12], [\frac{11}{2}\frac32\frac12\frac12\frac12], [\frac92\frac72\frac32\frac32\frac12], [\frac92\frac72\frac32\frac12\frac12], [\frac92\frac52\frac52\frac32\frac32], [\frac92\frac52\frac52\frac32\frac12], [\frac92\frac52\frac32\frac32\frac32], 2[\frac92\frac52\frac32\frac32\frac12], 2[\frac92\frac52\frac32\frac12\frac12], [\frac92\frac52\frac12\frac12\frac12], [\frac92\frac32\frac32\frac32\frac32], [\frac92\frac32\frac32\frac32\frac12], 2[\frac92\frac32\frac32\frac12\frac12], [\frac92\frac32\frac12\frac12\frac12], [\frac72\frac72\frac52\frac52\frac12], [\frac72\frac72\frac52\frac32\frac12], [\frac72\frac72\frac52\frac12\frac12], [\frac72\frac72\frac32\frac32\frac12], [\frac72\frac72\frac32\frac12\frac12], [\frac72\frac52\frac52\frac52\frac12], [\frac72\frac52\frac52\frac32\frac32], 2[\frac72\frac52\frac52\frac32\frac12], [\frac72\frac52\frac52\frac12\frac12], [\frac72\frac52\frac32\frac32\frac32], 2[\frac72\frac52\frac32\frac32\frac12], 2[\frac72\frac52\frac32\frac12\frac12], [\frac72\frac52\frac12\frac12\frac12], [\frac72\frac32\frac32\frac32\frac32], [\frac72\frac32\frac32\frac32\frac12], 2[\frac72\frac32\frac32\frac12\frac12], [\frac72\frac32\frac12\frac12\frac12], [\frac52\frac52\frac52\frac32\frac32], [\frac52\frac52\frac52\frac32\frac12], [\frac52\frac52\frac32\frac32\frac32], [\frac52\frac52\frac32\frac32\frac12], [\frac52\frac52\frac32\frac12\frac12], [\frac52\frac52\frac12\frac12\frac12], [\frac52\frac32\frac32\frac32\frac32], [\frac52\frac32\frac32\frac12\frac12], [\frac52\frac32\frac12\frac12\frac12], [\frac32\frac32\frac32\frac32\frac32] + \theta^9$
θ^{12}	[6 2 2],[6 2 1 1 1],[6 2 1],[6 2],[6 1 1 1 1], [5 3 2 1 1],[5 3 2 1],[5 3 1 1 1],[5 3 1 1], [5 2 2 2 2],[5 2 2 2 1],[5 2 2 1 1],[5 2 2 1], [5 2 2],2[5 2 1 1 1],[5 2 1 1],[5 2 1],[5 2], [5 1 1 1 1],[4 4 2 2],[4 4 2 1],[4 4 2],[4 3 3 2 1], [4 3 3 1 1],[4 3 2 2 1],[4 3 2 2],2[4 3 2 1 1], 2[4 3 2 1],[4 3 2],[4 3 1 1 1],[4 3 1 1],[4 2 2 2 2], [4 2 2 2 1],[4 2 2 2],[4 2 2 1 1],2[4 2 2 1], 2[4 2 2],2[4 2 1 1 1],[4 2 1 1],[4 2 1],[4 2], [4 1 1 1 1],[3 3 3 2 1],[3 3 3 1 1],[3 3 2 2 1], 2[3 3 2 1 1],[3 3 2 1],[3 3 1 1 1],[3 3 1 1], [3 2 2 2 2],[3 2 2 2 1],[3 2 2 1 1],[3 2 2 1], [3 2 2],[3 2 1 1 1],[3 2 1],[3 2],[3 1 1 1 1], [2 2 2 2 2],[2 2 2 2 1],[2 2 2],[2 2 1 1 1],[2 2 1], [2 2],[2 1 1 1 1] + θ^{10}
θ^{13}	$[\frac{13}{2}\frac32\frac32\frac12\frac12], [\frac{13}{2}\frac32\frac12\frac12\frac12], [\frac{11}{2}\frac52\frac52\frac12\frac12], [\frac{11}{2}\frac52\frac32\frac32\frac32], [\frac{11}{2}\frac52\frac32\frac32\frac12], [\frac{11}{2}\frac52\frac32\frac12\frac12], [\frac{11}{2}\frac52\frac12\frac12\frac12], [\frac{11}{2}\frac32\frac32\frac32\frac32], [\frac{11}{2}\frac32\frac32\frac32\frac12], [\frac{11}{2}\frac32\frac32\frac12\frac12], [\frac{11}{2}\frac32\frac12\frac12\frac12], [\frac92\frac72\frac52\frac32\frac12], [\frac92\frac72\frac52\frac12\frac12], [\frac92\frac72\frac32\frac32\frac12], [\frac92\frac72\frac32\frac12\frac12], [\frac92\frac52\frac52\frac32\frac32], [\frac92\frac52\frac52\frac32\frac12], 2[\frac92\frac52\frac52\frac12\frac12], [\frac92\frac52\frac32\frac32\frac32], 2[\frac92\frac52\frac32\frac32\frac12], 2[\frac92\frac52\frac32\frac12\frac12], [\frac92\frac52\frac12\frac12\frac12], [\frac92\frac32\frac32\frac32\frac32], [\frac92\frac32\frac32\frac32\frac12], [\frac92\frac32\frac32\frac12\frac12], [\frac92\frac32\frac12\frac12\frac12], [\frac72\frac72\frac72\frac32\frac32], [\frac72\frac72\frac52\frac32\frac32], [\frac72\frac72\frac52\frac32\frac12], [\frac72\frac72\frac52\frac12\frac12], [\frac72\frac72\frac32\frac32\frac32], [\frac72\frac72\frac32\frac32\frac12], [\frac72\frac72\frac32\frac12\frac12], [\frac72\frac52\frac52\frac52\frac32], [\frac72\frac52\frac52\frac32\frac32], [\frac72\frac52\frac52\frac32\frac12], 2[\frac72\frac52\frac52\frac12\frac12], [\frac72\frac52\frac32\frac32\frac32], 2[\frac72\frac52\frac32\frac32\frac12], 2[\frac72\frac52\frac32\frac12\frac12], [\frac72\frac52\frac12\frac12\frac12], [\frac72\frac32\frac32\frac32\frac32], [\frac72\frac32\frac32\frac32\frac12], [\frac72\frac32\frac32\frac12\frac12], [\frac72\frac32\frac12\frac12\frac12], [\frac52\frac52\frac52\frac52\frac32], [\frac52\frac52\frac52\frac32\frac32],[\frac52\frac52\frac52\frac12\frac12], [\frac52\frac52\frac32\frac32\frac32], [\frac52\frac52\frac32\frac32\frac12], [\frac52\frac52\frac32\frac12\frac12], [\frac52\frac52\frac12\frac12\frac12], [\frac52\frac32\frac32\frac32\frac32], [\frac52\frac32\frac32\frac32\frac12], [\frac52\frac32\frac32\frac12\frac12], [\frac52\frac32\frac12\frac12\frac12], [\frac32\frac32\frac32\frac12\frac12], [\frac32\frac32\frac12\frac12\frac12] + \theta^{11}$
θ^{14}	[7 1 1 1],[7 1 1],[6 2 2 1 1],[6 2 2 1],[6 2 1 1 1], [6 2 1 1],[6 1 1 1],[6 1 1],[5 3 3 1],[5 3 3], [5 3 2 2 1],[5 3 2 1 1],[5 3 2 1],[5 3 2],[5 3 1 1], [5 3 1],[5 2 2 2 1],2[5 2 2 1 1],[5 2 2 1], [5 2 1 1 1],[5 2 1 1],[5 1 1 1],[5 1 1], [4 4 3 1 1],[4 4 2 1 1],[4 4 1 1 1],[4 3 3 2 2], [4 3 3 1 1],[4 3 3 1],[4 3 3],[4 3 2 2 2], [4 3 2 2 1],2[4 3 2 1 1],[4 3 2 1],[4 3 2], [4 3 1 1 1],[4 3 1 1],[4 3 1],[4 2 2 2 1], 2[4 2 2 1 1],[4 2 2 1],[4 2 2],[4 1 1 1],[4 1 1], [3 3 3 2 2],[3 3 3 1],[3 3 3],[3 3 2 2 2], [3 3 2 2 1],[3 3 2 1 1],[3 3 2 1],[3 3 2],[3 3 1 1], [3 3 1],[3 2 2 2 1],[3 2 2 2],[3 2 2 1 1],[3 2 2 1], [3 2 2],[3 1 1 1],[3 1 1],[2 2 2 2],[2 2 2 1], [2 2 2],[2 1 1 1],[2 1 1],[1 1 1 1],[1 1 1] + θ^{12}
θ^{15}	$[\frac{15}{2}\frac12\frac12\frac12\frac12], [\frac{13}{2}\frac32\frac32\frac32\frac12], [\frac{13}{2}\frac32\frac32\frac12\frac12], [\frac{13}{2}\frac12\frac12\frac12\frac12], [\frac{11}{2}\frac52\frac52\frac32\frac12], [\frac{11}{2}\frac52\frac52\frac12\frac12], [\frac{11}{2}\frac52\frac32\frac32\frac12], [\frac{11}{2}\frac52\frac32\frac12\frac12], [\frac{11}{2}\frac32\frac32\frac32\frac12], [\frac{11}{2}\frac32\frac32\frac12\frac12], [\frac{11}{2}\frac12\frac12\frac12\frac12],[\frac92\frac72\frac72\frac12\frac12], [\frac92\frac72\frac52\frac32\frac32], [\frac92\frac72\frac52\frac12\frac12], [\frac92\frac72\frac32\frac32\frac32], [\frac92\frac72\frac32\frac12\frac12], [\frac92\frac72\frac12\frac12\frac12], [\frac92\frac52\frac52\frac32\frac32], [\frac92\frac52\frac52\frac32\frac12], [\frac92\frac52\frac52\frac12\frac12], [\frac92\frac52\frac32\frac32\frac32], [\frac92\frac52\frac32\frac32\frac12], [\frac92\frac52\frac32\frac12\frac12], [\frac92\frac32\frac32\frac32\frac12], [\frac92\frac32\frac32\frac12\frac12], [\frac92\frac12\frac12\frac12\frac12], [\frac72\frac72\frac72\frac12\frac12], [\frac72\frac72\frac52\frac52\frac52], [\frac72\frac72\frac52\frac32\frac32], [\frac72\frac72\frac52\frac12\frac12], [\frac72\frac72\frac32\frac32\frac32], [\frac72\frac72\frac32\frac12\frac12], [\frac72\frac72\frac12\frac12\frac12], [\frac72\frac52\frac52\frac52\frac52], [\frac72\frac52\frac52\frac32\frac32], [\frac72\frac52\frac52\frac32\frac12], [\frac72\frac52\frac52\frac12\frac12], [\frac72\frac52\frac32\frac32\frac32], [\frac72\frac52\frac32\frac32\frac12], [\frac72\frac52\frac32\frac12\frac12], [\frac72\frac32\frac32\frac32\frac12], [\frac72\frac32\frac32\frac12\frac12], [\frac72\frac12\frac12\frac12\frac12], [\frac52\frac52\frac52\frac32\frac12], [\frac52\frac52\frac52\frac12\frac12], [\frac52\frac52\frac32\frac32\frac12], [\frac52\frac52\frac32\frac12\frac12], [\frac52\frac32\frac32\frac32\frac12], [\frac52\frac32\frac32\frac12\frac12], [\frac52\frac12\frac12\frac12\frac12], [\frac32\frac32\frac32\frac32\frac12], [\frac32\frac32\frac32\frac12\frac12], [\frac32\frac12\frac12\frac12\frac12], [\frac12\frac12\frac12\frac12\frac12] + \theta^{13}$
θ^{16}	[8],[7 1 1 1 1],[7],[6 2 2 2],[6 2 2 1],[6 2 2], [6 1 1 1 1],[6],[5 3 3 1 1],[5 3 2 1 1],[5 3 1 1 1], [5 2 2 2],[5 2 2 1],[5 2 2],[5 1 1 1 1],[5],[4 4 4], [4 4 3],[4 4 2 2 2],[4 4 2],[4 4 1],[4 4], [4 3 3 1 1],[4 3 2 2 2],[4 3 2 1 1],[4 3 1 1 1], [4 2 2 2 2],[4 2 2 2],[4 2 2 1],[4 2 2],[4 1 1 1 1], [4],[3 3 3 3 3],[3 3 3 1 1],[3 3 2 1 1],[3 3 1 1 1], [3 2 2 2],[3 2 2 1],[3 2 2],[3 1 1 1 1],[3], [2 2 2 2],[2 2 2 1],[2 2 2],[2 1 1 1 1],[2], [1 1 1 1 1],[1],[0] + θ^{14}

The box θ^n lists the SO(10,1) representations of the field $W_{\alpha_1\cdots\alpha_n}(x)$ in the expansion

$$\Phi(x,\theta) = \sum_{n=0}^{32} \theta^{\alpha_1}\cdots\theta^{\alpha_n} W_{\alpha_1\cdots\alpha_n}(x) .$$

TABLE IV. Irreducible superfields included in the scalar superfield in 11 dimensions.

j	Δ^j_-
0	$[0 0 0 0 0]$
1	$[\frac12\frac12\frac12\frac12 -\frac12]$
2	$[1 1 1 0 0]$
3	$[\frac32\frac32\frac12\frac12\frac12]$
4	$[2 2 0 0 0] \oplus [2 1 1 1 1]$
5	$[\frac52\frac32\frac12\frac12\frac12] \oplus [\frac32\frac32\frac32\frac32\frac32]$
6	$[3 1 1 0 0] \oplus [2 2 1 1 1]$
7	$[\frac72\frac12\frac12\frac12 -\frac12] \oplus [\frac52\frac32\frac32\frac12\frac12]$
8	$[4 0 0 0 0] \oplus [3 1 1 1 0] \oplus [2 2 2 0 0]$
9	$[\frac72\frac12\frac12\frac12\frac12] \oplus [\frac52\frac32\frac32\frac12 -\frac12]$
10	$[3 1 1 0 0] \oplus [2 2 1 1 -1]$
11	$[\frac52\frac32\frac12\frac12 -\frac12] \oplus [\frac32\frac32\frac32\frac32 -\frac32]$
12	$[2 2 0 0 0] \oplus [2 1 1 1 -1]$
13	$[\frac32\frac32\frac12\frac12 -\frac12]$
14	$[1 1 1 0 0]$
15	$[\frac12\frac12\frac12\frac12\frac12]$
16	$[0 0 0 0 0]$

$m + n > 8$ since $\theta^{(\pm)4+j} = \theta^{(\pm)4-j}$ representation wise.

Table II, in similar fashion, lists the SO(8,1) representations contained in the nine-dimensional scalar superfield, where we stop at θ^8, since for θ^{8+n} we have the same representations as for θ^{8-n}. Table III gives the results for the scalar superfield in $d = 11$, where we need to go only up to θ^{16}. In this table, the representations appearing in the θ^n sector appear again in the θ^{n+2} sector, for $n \leqslant 14$. In order to save space, we limit ourselves to list the new representations in θ^{n+2} and simply write $+\theta^n$ at the end to indicate the set of representations in the θ^n sector.

In ten dimensions one can have "chiral" $\Phi(x,\theta^{(+)})$ and "antichiral" $\Phi(x,\theta^{(-)})$ superfields, whose contents have been correctly reported in Ref. 1. The content of the full scalar superfield $\Phi(x,\theta)$ can be obtained straightforwardly[4] by reducing the SO(10,1) representations given in Table III to SO(9,1) and will not be done here.

III. STRUCTURE OF THE SCALAR SUPERFIELD IN 11 DIMENSIONS

Now we briefly mention some facts about the super-Poincaré algebra in d dimensions SP_d. This algebra contains

TABLE V. Irreducible superfield $[\frac12\frac12\frac12\frac12\frac12]$.

Factor	SO(10) representations
2[0]	$2[\frac12\frac12\frac12\frac12\frac12]$
$[\frac12\frac12\frac12\frac12 -\frac12]$	[1 1 1 1],[1 1],[0]
$[\frac12\frac12\frac12\frac12\frac12]$	[1 1 1 1 1],[1 1 1],[1]
2[1 1 1]	$2[\frac32\frac32\frac32\frac12\frac12]$, $2[\frac32\frac32\frac12\frac12 -\frac12]$, $2[\frac32\frac12\frac12\frac12\frac12]$, $2[\frac12\frac12\frac12\frac12 -\frac12]$
$[\frac32\frac32\frac12\frac12 -\frac12]$	[2 2 1 1],[2 2],[2 1 1 1 − 1],[2 1 1],[1 1 1 1],[1 1]
$[\frac32\frac32\frac12\frac12\frac12]$	[2 2 1 1 1],[2 2 1],[2 1 1 1],[2 1],[1 1 1 1 1], [1 1 1]
2[2 2]	$2[\frac52\frac52\frac12\frac12\frac12]$, $2[\frac52\frac32\frac12\frac12 -\frac12]$, $2[\frac32\frac32\frac12\frac12\frac12]$
[2 1 1 1 1]	$[\frac52\frac32\frac32\frac32\frac32]$, $[\frac52\frac32\frac32\frac12\frac12]$, $[\frac52\frac12\frac12\frac12\frac12]$, $[\frac32\frac32\frac32\frac32\frac12]$, $[\frac32\frac32\frac12\frac12\frac12]$
[2 1 1 1 − 1]	$[\frac52\frac32\frac32\frac32 -\frac12]$, $[\frac52\frac32\frac12\frac12 -\frac12]$, $[\frac32\frac32\frac32\frac32 -\frac32]$, $[\frac32\frac32\frac32\frac12 -\frac12]$, $[\frac32\frac12\frac12\frac12 -\frac12]$
$[\frac52\frac32\frac12\frac12\frac12]$	[3 2 1 1 1],[3 2 1],[3 1 1 1],[3 1],[2 2 1 1],[2 2], [2 1 1 1 1],[2 1 1]
$[\frac52\frac32\frac12\frac12 -\frac12]$	[3 2 1 1],[3 2],[3 1 1 1 − 1],[3 1 1],[2 2 1 1 − 1], [2 2 1],[2 1 1 1],[2 1]
$[\frac32\frac32\frac32\frac32\frac32]$	[2 2 2 2 2],[2 2 2 1 1],[2 1 1 1 1]
$[\frac32\frac32\frac32\frac32 -\frac32]$	[2 2 2 2 − 1],[2 2 1 1 − 1],[1 1 1 1 − 1]
2[3 1 1]	$2[\frac72\frac32\frac32\frac12\frac12]$, $2[\frac72\frac32\frac12\frac12 -\frac12]$, $2[\frac72\frac12\frac12\frac12\frac12]$, $2[\frac52\frac32\frac32\frac12 -\frac12]$, $2[\frac52\frac32\frac12\frac12\frac12]$, $2[\frac52\frac12\frac12\frac12 -\frac12]$
[2 2 1 1 1]	$[\frac52\frac52\frac32\frac32\frac32]$, $[\frac52\frac52\frac32\frac12\frac12]$, $[\frac52\frac32\frac32\frac32\frac12]$, $[\frac52\frac32\frac12\frac12\frac12]$, $[\frac32\frac32\frac32\frac32\frac32]$, $[\frac32\frac32\frac32\frac12\frac12]$
[2 2 1 1 − 1]	$[\frac52\frac52\frac32\frac32 -\frac12]$, $[\frac52\frac52\frac12\frac12 -\frac12]$ $[\frac52\frac32\frac32\frac32 -\frac32]$, $[\frac52\frac32\frac32\frac12 -\frac12]$, $[\frac32\frac32\frac32\frac32 -\frac12]$, $[\frac32\frac32\frac12\frac12 -\frac12]$
$[\frac72\frac12\frac12\frac12\frac12]$	[4 1 1 1 1],[4 1 1],[4],[3 1 1 1],[3 1]
$[\frac72\frac12\frac12\frac12 -\frac12]$	[4 1 1 1],[4 1],[3 1 1 1 − 1],[3 1 1],[3]
$[\frac52\frac32\frac32\frac12\frac12]$	[3 2 2 1 1],[3 2 2],[3 2 1 1],[3 1 1 1 1],[3 1 1], [2 2 2 1],[2 2 1 1 1],[2 2 1],[2 1 1 1]
$[\frac52\frac32\frac32\frac12 -\frac12]$	[3 2 2 1],[3 2 1 1 − 1],[3 2 1],[3 1 1 1],[2 2 2 1 − 1], [2 2 2],[2 2 1 1],[2 1 1 1 − 1],[2 1 1]
[4]	$[\frac92\frac12\frac12\frac12\frac12]$, $[\frac72\frac12\frac12\frac12 -\frac12]$
[3 1 1 1]	$[\frac72\frac32\frac32\frac32\frac12]$, $[\frac72\frac32\frac32\frac12 -\frac12]$, $[\frac72\frac32\frac12\frac12\frac12]$, $[\frac72\frac12\frac12\frac12 -\frac12]$, $[\frac52\frac32\frac32\frac32 -\frac12]$, $[\frac52\frac32\frac32\frac12\frac12]$, $[\frac52\frac32\frac12\frac12 -\frac12]$, $[\frac52\frac12\frac12\frac12\frac12]$
[2 2 2]	$[\frac52\frac52\frac52\frac12\frac12]$, $[\frac52\frac52\frac32\frac12 -\frac12]$, $[\frac52\frac32\frac32\frac12\frac12]$, $[\frac32\frac32\frac32\frac12 -\frac12]$

TABLE VI. Irreducible superfield $[\frac12\frac12\frac12\frac12 -\frac12]$.

Factor	SO(10) representations
2[0]	$2[\frac12\frac12\frac12\frac12 -\frac12]$
$[\frac12\frac12\frac12\frac12\frac12]$	[1 1 1 1],[1 1],[0]
$[\frac12\frac12\frac12\frac12 -\frac12]$	[1 1 1 1 − 1],[1 1 1],[1]
2[1 1 1]	$2[\frac32\frac32\frac32\frac12 -\frac12]$, $2[\frac32\frac32\frac12\frac12\frac12]$, $2[\frac32\frac12\frac12\frac12 -\frac12]$, $2[\frac12\frac12\frac12\frac12\frac12]$
$[\frac32\frac32\frac12\frac12\frac12]$	[2 2 1 1],[2 2],[2 1 1 1 1],[2 1 1],[1 1 1 1],[1 1]
$[\frac32\frac32\frac12\frac12 -\frac12]$	[2 2 1 1 − 1],[2 2 1],[2 1 1 1],[2 1],[1 1 1 1 − 1], [1 1 1]
2[2 2]	$2[\frac52\frac52\frac12\frac12 -\frac12]$, $2[\frac52\frac32\frac12\frac12\frac12]$, $2[\frac32\frac32\frac12\frac12 -\frac12]$
[2 1 1 1 − 1]	$[\frac52\frac32\frac32\frac32 -\frac32]$, $[\frac32\frac32\frac32\frac12 -\frac12]$, $[\frac52\frac12\frac12\frac12 -\frac12]$, $[\frac32\frac32\frac32\frac32 -\frac12]$, $[\frac32\frac32\frac12\frac12 -\frac12]$
[2 1 1 1 1]	$[\frac52\frac32\frac32\frac32\frac12]$, $[\frac52\frac32\frac12\frac12\frac12]$, $[\frac32\frac32\frac32\frac32\frac32]$, $[\frac32\frac32\frac32\frac12\frac12]$, $[\frac32\frac12\frac12\frac12\frac12]$
$[\frac52\frac32\frac12\frac12 -\frac12]$	[3 2 1 1 − 1],[3 2 1],[3 1 1 1],[3 1],[2 2 1 1],[2 2], [2 1 1 1 − 1],[2 1 1]
$[\frac52\frac32\frac12\frac12\frac12]$	[3 2 1 1],[3 2],[3 1 1 1 1],[3 1 1],[2 2 1 1 1], [2 2 1],[2 1 1 1],[2 1]
$[\frac32\frac32\frac32\frac32 -\frac32]$	[2 2 2 2 − 2],[2 2 2 1 − 1],[2 1 1 1 − 1]
$[\frac32\frac32\frac32\frac32\frac32]$	[2 2 2 2 1],[2 2 1 1 1],[1 1 1 1 1]
2[3 1 1]	$2[\frac72\frac32\frac32\frac12 -\frac12]$, $2[\frac72\frac32\frac12\frac12\frac12]$, $2[\frac72\frac12\frac12\frac12 -\frac12]$, $2[\frac52\frac32\frac32\frac12\frac12]$, $2[\frac52\frac32\frac12\frac12 -\frac12]$, $2[\frac52\frac12\frac12\frac12\frac12]$
[2 2 1 1 − 1]	$[\frac52\frac52\frac32\frac32 -\frac32]$, $[\frac52\frac52\frac32\frac12 -\frac12]$, $[\frac52\frac32\frac32\frac32 -\frac12]$, $[\frac52\frac32\frac12\frac12 -\frac12]$, $[\frac32\frac32\frac32\frac32 -\frac32]$, $[\frac32\frac32\frac32\frac12 -\frac12]$
[2 2 1 1 1]	$[\frac52\frac52\frac32\frac32\frac12]$, $[\frac52\frac52\frac12\frac12\frac12]$ $[\frac52\frac32\frac32\frac32\frac32]$, $[\frac52\frac32\frac32\frac12\frac12]$, $[\frac32\frac32\frac32\frac32\frac12]$, $[\frac32\frac32\frac12\frac12\frac12]$
$[\frac72\frac12\frac12\frac12 -\frac12]$	[4 1 1 1 − 1],[4 1 1],[4],[3 1 1 1],[3 1]
$[\frac72\frac12\frac12\frac12\frac12]$	[4 1 1 1],[4 1],[3 1 1 1 1],[3 1 1],[3]
$[\frac52\frac32\frac32\frac12 -\frac12]$	[3 2 2 1 − 1],[3 2 2],[3 2 1 1],[3 1 1 1 − 1],[3 1 1], [2 2 2 1],[2 2 1 1 − 1],[2 2 1],[2 1 1 1]
$[\frac52\frac32\frac32\frac12\frac12]$	[3 2 2 1],[3 2 1 1 1],[3 2 1],[3 1 1 1],[2 2 2 1 1], [2 2 2],[2 2 1 1],[2 1 1 1 1],[2 1 1]
[4]	$[\frac92\frac12\frac12\frac12 -\frac12]$, $[\frac72\frac12\frac12\frac12\frac12]$
[3 1 1 1]	$[\frac72\frac32\frac32\frac32 -\frac12]$, $[\frac72\frac32\frac32\frac12\frac12]$, $[\frac72\frac32\frac12\frac12 -\frac12]$, $[\frac72\frac12\frac12\frac12\frac12]$, $[\frac52\frac32\frac32\frac32\frac12]$, $[\frac52\frac32\frac32\frac12 -\frac12]$, $[\frac52\frac32\frac12\frac12\frac12]$, $[\frac52\frac12\frac12\frac12 -\frac12]$
[2 2 2]	$[\frac52\frac52\frac52\frac12 -\frac12]$, $[\frac52\frac52\frac32\frac12\frac12]$, $[\frac52\frac32\frac32\frac12 -\frac12]$, $[\frac32\frac32\frac32\frac12\frac12]$

TABLE VII. Irreducible superfield [1 1 1].

Factor	SO(10) representations
2[0]	2[1 1 1]
$[\frac12\frac12\frac12\frac12\frac12]$	$[\frac32\frac32\frac32\frac12\frac12], [\frac32\frac32\frac12\frac12-\frac12], [\frac32\frac12\frac12\frac12\frac12], [\frac12\frac12\frac12\frac12-\frac12]$
$[\frac12\frac12\frac12\frac12-\frac12]$	$[\frac32\frac32\frac32\frac12-\frac12], [\frac32\frac32\frac12\frac12\frac12], [\frac32\frac12\frac12\frac12-\frac12], [\frac12\frac12\frac12\frac12\frac12]$
2[1 1 1]	2[2 2 2],2[2 2 1 1],2[2 2],2[2 1 1 1 1],2[2 1 1 1 − 1], 2[2 1 1],2[2],4[1 1 1 1],2[1 1],2[0]
$[\frac32\frac32\frac12\frac12\frac12]$	$[\frac52\frac52\frac32\frac12\frac12], [\frac52\frac52\frac12\frac12-\frac12], [\frac52\frac32\frac32\frac32\frac12], [\frac52\frac32\frac32\frac12-\frac12],$ $2[\frac52\frac32\frac12\frac12\frac12], [\frac52\frac12\frac12\frac12-\frac12], [\frac32\frac32\frac32\frac32\frac32], [\frac32\frac32\frac32\frac32-\frac12],$ $2[\frac32\frac32\frac32\frac12\frac12], 2[\frac32\frac32\frac12\frac12-\frac12], 2[\frac32\frac12\frac12\frac12\frac12], [\frac12\frac12\frac12\frac12-\frac12]$
$[\frac32\frac32\frac12\frac12-\frac12]$	$[\frac52\frac52\frac32\frac12-\frac12], [\frac52\frac52\frac12\frac12\frac12], [\frac52\frac32\frac32\frac32-\frac12], [\frac52\frac32\frac32\frac12\frac12],$ $2[\frac52\frac32\frac12\frac12-\frac12], [\frac52\frac12\frac12\frac12\frac12], [\frac32\frac32\frac32\frac32-\frac32], [\frac32\frac32\frac32\frac32\frac12],$ $2[\frac32\frac32\frac32\frac12-\frac12], 2[\frac32\frac32\frac12\frac12\frac12], 2[\frac32\frac12\frac12\frac12-\frac12], [\frac12\frac12\frac12\frac12\frac12]$
2[2 2]	2[3 3 1],2[3 2 1 1],2[3 2],2[3 1 1],2[2 2 1 1 1], 2[2 2 1 1 − 1],2[2 2 1],2[2 1 1 1],2[2 1],2[1 1 1]
[2 1 1 1 1]	[3 2 2 1 1],[3 2 1 1],[3 1 1 1 1],[3 1 1], [2 2 2 2 1],[2 2 2 1],2[2 2 1 1 1],[2 2 1], 2[2 1 1 1],[2 1],[1 1 1 1 1],[1 1 1]
[2 1 1 1 − 1]	[3 2 2 1 − 1],[3 2 1 1],[3 1 1 1 − 1],[3 1 1], [2 2 2 2 − 1],[2 2 2 1],2[2 2 1 1 − 1],[2 2 1], 2[2 1 1 1],[2 1],[1 1 1 1 − 1],[1 1 1]
$[\frac52\frac32\frac12\frac12\frac12]$	$[\frac72\frac52\frac32\frac12\frac12], [\frac72\frac52\frac12\frac12-\frac12], [\frac72\frac32\frac32\frac32\frac12], [\frac72\frac32\frac32\frac12-\frac12],$ $2[\frac72\frac32\frac12\frac12\frac12], [\frac72\frac12\frac12\frac12-\frac12], [\frac52\frac52\frac32\frac32\frac12], [\frac52\frac52\frac32\frac12-\frac12],$ $2[\frac52\frac52\frac12\frac12\frac12], [\frac52\frac32\frac32\frac32\frac32], [\frac52\frac32\frac32\frac32-\frac12], 3[\frac52\frac32\frac32\frac12\frac12],$ $3[\frac52\frac32\frac12\frac12-\frac12], 2[\frac52\frac12\frac12\frac12\frac12], [\frac32\frac32\frac32\frac32\frac12], [\frac32\frac32\frac32\frac12-\frac12],$ $2[\frac32\frac32\frac12\frac12\frac12], [\frac32\frac12\frac12\frac12-\frac12]$
$[\frac52\frac32\frac12\frac12-\frac12]$	$[\frac72\frac52\frac32\frac12-\frac12], [\frac72\frac52\frac12\frac12\frac12], [\frac72\frac32\frac32\frac32-\frac12], [\frac72\frac32\frac32\frac12\frac12],$ $2[\frac72\frac32\frac12\frac12-\frac12], [\frac72\frac12\frac12\frac12\frac12], [\frac52\frac52\frac32\frac32-\frac12], [\frac52\frac52\frac32\frac12\frac12],$ $2[\frac52\frac52\frac12\frac12-\frac12], [\frac52\frac32\frac32\frac32-\frac32], [\frac52\frac32\frac32\frac32\frac12], 3[\frac52\frac32\frac32\frac12-\frac12],$ $3[\frac52\frac32\frac12\frac12\frac12], 2[\frac52\frac12\frac12\frac12-\frac12], [\frac32\frac32\frac32\frac32-\frac12], [\frac32\frac32\frac32\frac12\frac12],$ $2[\frac32\frac32\frac12\frac12-\frac12], [\frac32\frac12\frac12\frac12\frac12]$
$[\frac32\frac32\frac32\frac32\frac32]$	$[\frac52\frac52\frac52\frac32\frac32], [\frac52\frac52\frac32\frac32\frac12], [\frac52\frac32\frac32\frac32\frac32], [\frac52\frac32\frac32\frac12\frac12],$ $[\frac32\frac32\frac32\frac32\frac12], [\frac32\frac32\frac12\frac12\frac12]$
$[\frac32\frac32\frac32\frac32-\frac32]$	$[\frac52\frac52\frac52\frac32-\frac32], [\frac52\frac52\frac32\frac32-\frac12], [\frac52\frac32\frac32\frac32-\frac32], [\frac52\frac32\frac32\frac12-\frac12],$ $[\frac32\frac32\frac32\frac32-\frac12], [\frac32\frac32\frac12\frac12-\frac12]$
2[3 1 1]	2[4 2 2],2[4 2 1 1],2[4 2],2[4 1 1 1 1],2[4 1 1 1 − 1], 2[4 1 1],2[4],2[3 2 2 1],2[3 2 1 1 1],2[3 2 1 1 − 1], 4[3 2 1],6[3 1 1 1],4[3 1],2[2 2 2],2[2 2 1 1], 2[2 2],2[2 1 1 1 1],2[2 1 1 1 − 1],2[2 1 1],2[2]
[2 2 1 1 1]	[3 3 2 1 1],[3 3 1 1],[3 2 2 2 1],[3 2 2 1], 2[3 2 1 1 1],[3 2 1],[3 1 1 1],[2 2 2 2 2], [2 2 2 2],2[2 2 2 1 1],[2 2 2],2[2 2 1 1],[2 2], 2[2 1 1 1 1],[2 1 1],[1 1 1 1]
[2 2 1 1 − 1]	[3 3 2 1 − 1],[3 3 1 1],[3 2 2 2 − 1],[3 2 2 1], 2[3 2 1 1 − 1],[3 2 1],[3 1 1 1],[2 2 2 2 − 2], [2 2 2 2],2[2 2 2 1 − 1],[2 2 2],2[2 2 1 1],[2 2], 2[2 1 1 1 − 1],[2 1 1],[1 1 1 1]
$[\frac72\frac12\frac12\frac12\frac12]$	$[\frac92\frac32\frac32\frac12\frac12], [\frac92\frac32\frac12\frac12-\frac12], [\frac92\frac12\frac12\frac12\frac12], [\frac72\frac32\frac32\frac32\frac12],$ $[\frac72\frac32\frac32\frac12-\frac12], 2[\frac72\frac32\frac12\frac12\frac12], 2[\frac72\frac12\frac12\frac12-\frac12], [\frac52\frac32\frac32\frac12\frac12],$ $[\frac52\frac32\frac12\frac12-\frac12], [\frac52\frac12\frac12\frac12\frac12]$
$[\frac72\frac12\frac12\frac12-\frac12]$	$[\frac92\frac32\frac32\frac12-\frac12], [\frac92\frac32\frac12\frac12\frac12], [\frac92\frac12\frac12\frac12-\frac12], [\frac72\frac32\frac32\frac32-\frac12],$ $[\frac72\frac32\frac32\frac12\frac12], 2[\frac72\frac32\frac12\frac12-\frac12], 2[\frac72\frac12\frac12\frac12\frac12], [\frac52\frac32\frac32\frac12-\frac12],$ $[\frac52\frac32\frac12\frac12\frac12], [\frac52\frac12\frac12\frac12-\frac12]$
$[\frac52\frac32\frac32\frac12\frac12]$	$[\frac72\frac52\frac52\frac12\frac12], [\frac72\frac52\frac32\frac32\frac12], [\frac72\frac52\frac32\frac12-\frac12], [\frac72\frac52\frac12\frac12\frac12],$ $[\frac72\frac32\frac32\frac32\frac32], [\frac72\frac32\frac32\frac32-\frac12], 2[\frac72\frac32\frac32\frac12\frac12], [\frac72\frac32\frac12\frac12-\frac12],$ $[\frac72\frac12\frac12\frac12\frac12], [\frac52\frac52\frac52\frac32\frac12], [\frac52\frac52\frac52\frac12-\frac12], [\frac52\frac52\frac32\frac32\frac32],$ $[\frac52\frac52\frac32\frac32-\frac12], 3[\frac52\frac52\frac32\frac12\frac12], [\frac52\frac52\frac12\frac12-\frac12], 3[\frac52\frac32\frac32\frac32\frac12],$ $3[\frac52\frac32\frac32\frac12-\frac12], 3[\frac52\frac32\frac12\frac12\frac12], [\frac52\frac12\frac12\frac12-\frac12], [\frac32\frac32\frac32\frac32\frac32],$ $[\frac32\frac32\frac32\frac32-\frac12], 2[\frac32\frac32\frac32\frac12\frac12], [\frac32\frac32\frac12\frac12-\frac12], [\frac32\frac12\frac12\frac12\frac12]$
$[\frac52\frac32\frac32\frac12-\frac12]$	$[\frac72\frac52\frac52\frac12-\frac12], [\frac72\frac52\frac32\frac32-\frac12], [\frac72\frac52\frac32\frac12\frac12], [\frac72\frac52\frac12\frac12-\frac12],$ $[\frac72\frac32\frac32\frac32-\frac32], [\frac72\frac32\frac32\frac32\frac12], 2[\frac72\frac32\frac32\frac12-\frac12], [\frac72\frac32\frac12\frac12\frac12],$ $[\frac72\frac12\frac12\frac12-\frac12], [\frac52\frac52\frac52\frac32-\frac12], [\frac52\frac52\frac52\frac12\frac12], [\frac52\frac52\frac32\frac32-\frac32],$ $[\frac52\frac52\frac32\frac32\frac12], 3[\frac52\frac52\frac32\frac12-\frac12], [\frac52\frac52\frac12\frac12\frac12], 3[\frac52\frac32\frac32\frac32-\frac12],$ $3[\frac52\frac32\frac32\frac12\frac12], 3[\frac52\frac32\frac12\frac12-\frac12], [\frac52\frac12\frac12\frac12\frac12], [\frac32\frac32\frac32\frac32-\frac32],$ $[\frac32\frac32\frac32\frac32\frac12], 2[\frac32\frac32\frac32\frac12-\frac12], [\frac32\frac32\frac12\frac12\frac12], [\frac32\frac12\frac12\frac12-\frac12]$

TABLE VII. (*Continued.*)

[4]	[5 1 1],[4 1 1 1],[4 1],[3 1 1]
[3 1 1 1]	[4 2 2 1],[4 2 1 1 1],[4 2 1 1 − 1],[4 2 1],2[4 1 1 1], [4 1],[3 2 2 2],[3 2 2 1 1],[3 2 2 1 − 1],[3 2 2], 3[3 2 1 1],[3 2],2[3 1 1 1 1],2[3 1 1 1 − 1],3[3 1 1], [3],[2 2 2 1],[2 2 1 1 1],[2 2 1 1 − 1],[2 2 1], 2[2 1 1 1],[2 1]
[2 2 2]	[3 3 3],[3 3 2 1],[3 3 1],[3 2 2 1 1],[3 2 2 1 − 1], [3 2 2],[3 2 1 1],[3 1 1],2[2 2 2 1],[2 2 1 1 1], [2 2 1 1 − 1],[2 2 1],[2 1 1 1],[1 1 1]

the generators of the Poincaré group as well as a fermionic generator Q satisfying the Majorana condition. The square of the momentum P^2 is a Casimir operator for this algebra and its eigenvalue M^2 partially characterizes the representations. When $M^2 \neq 0$ a complete set of Casimirs can be found for the algebra, the remaining ones corresponding to Casimirs of SO$(d-1)$, which is the "little algebra" for SP$_d$ (see Ref. 2).

Therefore the massive $(M^2>0)$ irreducible representations of SP$_d$ are characterized by M and some SO$(d-1)$ irreducible representation. The highest weight describing this SO$(d-1)$ representation is what we call superweight.[2]

The eigenvalues of the Casimir operators, given in the form demanded by the scalar superfield, give us the irreducible superfields included in the former. The corresponding superweights can be obtained alternatively by computing totally antisymmetric Kronecker powers[5] of either of the two basic spinorial representations of SO(10), as thoroughly explained in Ref. 2.

Thus, in Table IV we display the irreducible superfields given by their superweights, which are included in the scalar superfield in 11 dimensions, as successive totally antisymmetrized Kronecker powers of the spinorial representation $\Delta_- = [\frac12\frac12\frac12\frac12 -\frac12]$ of SO(10).

Each of these irreducible superfields contains a multitude of ordinary fields, of course. These ordinary fields are representations of the Poincaré group whose mass is the same as the parent superfield and which are further labeled by the highest weight of some SO(10) irreducible representation. The set of highest weights included in one given superweight is obtained by performing the Kronecker product of that superweight with each of the superweights listed in Table IV. Each product is the Kronecker product of two SO(10) representations.

Unlike Ref. 1 we favor the technique involving Schur functions to compute Kronecker products,[6] which can be implemented in computer programs.

In Tables V–VII we report the field contents of the irreducible superfields characterized by $[\frac12\frac12\frac12\frac12\frac12]$, $[\frac12\frac12\frac12\frac12 -\frac12]$, and [1 1 1] in 11 dimensions. In these tables we list representations of the Poincaré group [the mass being omitted, this means SO(10) representations] rather than SO(10,1) by giving the result of the Kronecker product of the representation under "Factor" with the representation characterizing the corresponding irreducible superfield. We have to include under "Factor" all the irreducible representations of Table IV. From the discussion of the previous paragraph it is clear that the field content of the representation [0] is already listed in Table IV.

The reason to go up to $[1\ 1\ 1]$ and no further is that the smaller representations do not contain all three fields $[2]$, $[\frac{3}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}]$, and $[1\ 1\ 1]$ corresponding to the "massive supergravity multiplet"; $[1\ 1\ 1]$ is the smallest piece containing $[2]$. All the bigger irreducible superfields also contain the full multiplet.

Viewed as representations of SO(10,1), the massive counterpart of the supergravity multiplet is given again by $[2]$, $[\frac{3}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}]$, and $[1\ 1\ 1]$, which indeed appear in Table III: $[2]$ in the θ^{16} and $[\frac{3}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}]$ in the θ^{15} as well as the θ^{17} sector.

This tells us that it may not be necessary to go beyond the scalar superfield in order to formulate supergravity in 11 dimensions. Even the irreducible pieces that do not contain the massive counterpart of the supergravity multiplet could contain the supergravity multiplet itself in the limit $M^2 \to 0$. This remains unexplored so far.

[1] L. B. Litov, Bulg. J. Phys. **11**, 141 (1984).

[2] R. Finkelstein and M. Villasante, UCLA preprint UCLA/84/TEP/13, 1984.

[3] E. Cremmer, B. Julia, and J. Scherk, Phys. Lett. B **76**, 409 (1978).

[4] A. O. Barut and R. Raczka, *Theory of Group Representations and their Applications* (Polish Scientific, Warsaw, 1977), p. 228.

[5] B. G. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974), p. 115. See also, G. R. E. Black and B. G. Wybourne, J. Phys. A **16**, 2405 (1983).

[6] P. H. Butler and B. G. Wybourne, J. Phys. (Paris) **30**, 655 1969; B. G. Wybourne, *Symmetry Principles in Atomic Spectroscopy* (Wiley, New York, 1970); G. R. E. Black, R. C. King, and B. G. Wybourne, J. Phys. A **16**, 1555 (1983).

Local properties of quantum systems

Boris Leaf
Department of Physics, State University of New York, College at Cortland, Cortland, New York 13045

(Received 28 May 1985; accepted for publication 5 February 1986)

Local properties of a quantum system are defined as the expectation values of its observables in a microstate of some complete set of commuting observables. An equation for the time evolution of local properties is obtained for any system whose statistical operator (density matrix) evolves by unitary transformation in accordance with the von Neumann equation. The formalism is applied to the example of a system of one particle. In this case the local properties are fields, the time-evolution equation is an equation of continuity with source terms. For constants of motion the source terms vanish, giving equations of continuity for the fields. For each scalar field a flux vector for its transport current is defined. For momentum, a stress tensor is obtained. The effects on local properties of realization of a latent ensemble of the statistical operator (an entropy-increasing mechanism recently proposed to explain approach to equilibrium) are also considered; a non-negative local entropy production is identified, as well as a discontinuous redistribution of local properties among microstates.

I. INTRODUCTION

The properties of physical systems, including thermodynamic properties, are ensemble-average values of the corresponding observables of the system. An ensemble average is calculated as the trace of the product of the observable and the statistical operator that represents the state of the system.[1] Accordingly, these averages give global properties of the system. We use the term global to describe properties of the entire system. It is frequently the case, however, that we are interested in the value of a property, not for the system as a whole, but for some subsystem within the system, such as a particular spatial point or set points inside an extended system. For example, thermodynamic systems are characteristically subsystems of a global system, the "universe," with complementary subsystems, which are the "surroundings."

In order to discuss properties of subsystems we introduce the concept of a local property, which may be formulated in the following way. Each observable of the system is an operator belonging to some complete set of commuting observables.[2,3] The spectrum of simultaneous eigenvalues of such a set is nondegenerate; each eigenstate defines a microstate of the system for the observables in the complete set.[4] By a local property is meant the expectation of an observable, not for the system as a whole, but for an individual microstate of a complete set of commuting observables; in particular, the set of coordinate operators of the system whose microstates constitute the physical space of the system.[5] The definition of local property is given in (2.6). The property in a subspace of the system, which comprises a subset of all the microstates of a complete set of commuting observables, is additive: it is the sum of the local properties at these microstates in the subspace. The global property is the sum of local properties at all the microstates of the set of observables, a sum over all eigenvalues of the spectrum of these observables. From the defining equation (2.6) of a local property, we obtain the equation for its rate of change in time (2.11).

The main result of this paper is Eq. (2.11), which describes the rate of change of a local property when the statistical operator of the system evolves unitarily in accordance with the von Neumann equation. If the property is globally conserved, this equation becomes the local conservation law (2.14), which is satisfied, for example, by the local entropy, as shown in (2.20).

In Sec. III the formalism is applied to the example of a system that consists of one particle. Although globally the system is a particle, locally at a microstate of the coordinate operator, it is characterized by fields. Each local property is a field, whose time evolution, derived from (2.11), is the equation of continuity with source terms (3.7). For each scalar observable of the one-particle system there is a flux vector whose divergence appears in (3.7); it specifies the transport current for that observable. Several examples for various observables are considered. If the observable is a globally conserved property, the source terms in (3.7) vanish, leaving the equation of continuity (3.9) for the conserved field. Thus an equation of continuity for probability (number of particles) is obtained in (3.10), which takes the familiar form (3.12) when the statistical operator of the system is a pure state.[6] An equation of continuity for the local entropy is obtained in (3.13). The equation of change for the local momentum (3.15) permits identification of the stress tensor (3.16). Equations for local energy, kinetic, potential, and total, are obtained in (3.17)–(3.19), in which the corresponding flux vectors are readily identifiable.

The time evolution equation (2.11), on which the results of Secs. II and III are based, applies to systems whose statistical operators evolve by unitary transformation in accordance with the von Neumann equation. In unitary transformations the global entropy remains constant. In order to explain the approach of systems to equilibrium, it was proposed in a recent paper[4] that the statistical operator changes sporadically by realizations of latent ensembles, as well as by isentropic, unitary transformations. Non-negative increases in global entropy occur when latent ensembles are realized, increases that vanish when equilibrium is attained. As a final application of the concept of local properties introduced in Sec. II, the effects on local properties resulting from realiza-

0022-2488/86/061612-06$02.50

tion of a latent ensemble are considered in Sec. IV. Following an explanation of terminology and a recapitulation of the mechanism of realization of latent ensembles (described in detail in Ref. 4), the attendant changes in local properties are considered. In general, a discontinuous redistribution of local values among the microstates occurs, leaving the global values of observables and entropy unchanged. But, in addition, non-negative local entropy changes are produced at each microstate; the sum of these local entropy productions, taken over the spectrum of microstates, agrees with the global entropy increase for the system when a latent ensemble is realized.

II. LOCAL PROPERTIES

Consider, in the Schrödinger picture, an observable $\hat{Q}_t$ (with explicit dependence on time t) of a system in a state specified by statistical operator $\hat{w}_t$. The expectation value of $\hat{Q}_t$ for the system at time t is

$$\langle \hat{Q}_t \rangle_t = \operatorname{tr}(\hat{w}_t \hat{Q}_t), \quad \operatorname{tr} \hat{w}_t = 1. \tag{2.1}$$

According to the von Neumann equation for a system with Hamiltonian $\hat{H}_t$,

$$\frac{\partial \hat{w}_t}{\partial t} + \frac{i}{\hbar}[\hat{H}_t,\hat{w}_t]_- = 0, \tag{2.2}$$

so that

$$\frac{d\langle \hat{Q}_t \rangle_t}{dt} = \left\langle \frac{\partial \hat{Q}_t}{\partial t} + \frac{i}{\hbar}[\hat{H}_t,\hat{Q}_t]_- \right\rangle_t. \tag{2.3}$$

The expectation $\langle \hat{Q}_t \rangle_t$ in (2.1) is an average value for the entire system, a global property; it is the mean value predicted by theory for the experimental measurement of $\hat{Q}_t$ on the system in state $\hat{w}_t$. We are interested in predicting the values of $\hat{Q}_t$ measured on subsystems within the system, such as the value at a particular spatial point or set of points inside an extended system in a state $\hat{w}_t$. For this purpose we now introduce the notion of a local property.

Let $\hat{\mathbf{x}}$ represent a complete commuting set of observables of the system. Its spectrum of eigenvalues is nondegenerate and, for finite systems, discrete.[4,7] The projector

$$\hat{P}(\mathbf{x}) = |\mathbf{x}\rangle\langle\mathbf{x}|, \quad \operatorname{tr} \hat{P}(\mathbf{x}) = 1, \tag{2.4}$$

specifies the microstate of $\hat{\mathbf{x}}$ with eigenvector $|\mathbf{x}\rangle$. The completeness condition is

$$\int_{\mathbf{x}} \hat{P}(\mathbf{x}) = \hat{1}, \tag{2.5}$$

where $\int_{\mathbf{x}}$ indicates a spectral summation on the eigenvalues $\mathbf{x}$. A local property is the expectation value of $\hat{Q}_t$ in a microstate $\hat{P}(\mathbf{x})$, defined to be

$$\begin{aligned}\langle \hat{Q}_t \rangle(t,\mathbf{x}) &= \tfrac{1}{2} \operatorname{tr}(\hat{P}(\mathbf{x})[\hat{w}_t,\hat{Q}_t]_+) \\ &= \tfrac{1}{2}\langle \mathbf{x}|[\hat{w}_t,\hat{Q}_t]_+|\mathbf{x}\rangle.\end{aligned} \tag{2.6}$$

From the completeness condition (2.5) it follows that

$$\langle \hat{Q}_t \rangle_t = \int_{\mathbf{x}} \langle \hat{Q}_t \rangle(t,\mathbf{x}), \tag{2.7}$$

so that the global property is a sum of the local properties on the entire set of eigenvalues $\mathbf{x}$ or $\hat{\mathbf{x}}$. If the spectral sum in (2.7) is restricted to a subset of these eigenvalues, then $\langle \hat{Q}_t \rangle_t$ will be the value of the property on the subsystem that comprises just this subset of eigenvalues of $\hat{\mathbf{x}}$. For a subsystem the value of a property is the sum of local values on the microstates that it comprises.

The time evolution of the global property $\langle \hat{Q}_t \rangle_t$ is given in (2.3). We now derive the equation for the time evolution of the local property $\langle \hat{Q}_t \rangle_t(t,\mathbf{x})$. Using (2.2) and (2.6) we obtain

$$\begin{aligned}\frac{\partial \langle \hat{Q}_t \rangle(t,\mathbf{x})}{\partial t} &= (2i\hbar)^{-1} \operatorname{tr}([\hat{H}_t,\hat{w}_t]_- [\hat{P}(\mathbf{x}),\hat{Q}_t]_+) \\ &\quad + \tfrac{1}{2} \operatorname{tr}\left(\hat{w}_t \left[\hat{P}(\mathbf{x}), \frac{\partial \hat{Q}_t}{\partial t}\right]_+\right).\end{aligned} \tag{2.8}$$

Note that since $[\hat{H}_t,\hat{w}_t]_- = 0$ if the system is in equilibrium, then for an observable $\hat{Q}$ without explicit time dependence,

$$\frac{\partial \langle \hat{Q} \rangle(t,\mathbf{x})}{\partial t} = 0, \tag{2.9}$$

so that the local property $\langle \hat{Q} \rangle(t,\mathbf{x})$, as well as the global property $\langle \hat{Q} \rangle_t$ in (2.3), is independent of time. For operators $\hat{A}$, $\hat{B}$, $\hat{C}$, and $\hat{D}$, it is readily verified that

$$\begin{aligned}&\operatorname{tr}([\hat{A},\hat{B}]_-[\hat{C},\hat{D}]_+) \\ &\quad = \operatorname{tr}([\hat{C},\hat{A}]_-[\hat{D},\hat{B}]_+) + \operatorname{tr}([\hat{D},\hat{A}]_-[\hat{C},\hat{B}]_+).\end{aligned} \tag{2.10}$$

Accordingly, from (2.8),

$$\begin{aligned}\frac{\partial \langle \hat{Q}_t \rangle(t,\mathbf{x})}{\partial t} &= (2i\hbar)^{-1} \operatorname{tr}([\hat{P}(\mathbf{x}),\hat{H}_t]_-[\hat{w}_t,\hat{Q}_t]_+) \\ &\quad + (2i\hbar)^{-1} \operatorname{tr}([\hat{Q}_t,\hat{H}_t]_-[\hat{P}(\mathbf{x}),\hat{w}_t]_+) \\ &\quad + \tfrac{1}{2} \operatorname{tr}\left(\hat{P}(\mathbf{x})\left[\hat{w}_t,\frac{\partial \hat{Q}_t}{\partial t}\right]_+\right)\end{aligned}$$

or, from (2.6),

$$\begin{aligned}\frac{\partial \langle \hat{Q}_t \rangle(t,\mathbf{x})}{\partial t} &= \frac{i}{2\hbar} \operatorname{tr}([\hat{H}_t,\hat{P}(\mathbf{x})]_-[\hat{w}_t,\hat{Q}_t]_+) \\ &\quad + \left\langle \frac{\partial \hat{Q}_t}{\partial t} + \frac{i}{\hbar}[\hat{H}_t,\hat{Q}_t]_- \right\rangle(t,\mathbf{x}).\end{aligned} \tag{2.11}$$

This is the equation for the time evolution of a local property $\langle \hat{Q}_t \rangle(t,\mathbf{x})$. Note that (2.3) is recovered upon spectral summation of (2.11) on the set of eigenvalues $\mathbf{x}$, because of the completeness condition (2.5). But again, as noted after (2.7), spectral summation over a subset of eigenvalues gives $d\langle \hat{Q}_t \rangle/dt$ for the subsystem comprising this subset of eigenvalues of $\hat{\mathbf{x}}$.

Comparison of (2.11) and (2.3) shows that if $\hat{Q}_t$ satisfies the operator equation

$$\frac{\partial \hat{Q}_t}{\partial t} + \frac{i}{\hbar}[\hat{H}_t,\hat{Q}_t]_- = 0, \tag{2.12}$$

then the global property $\langle \hat{Q}_t \rangle_t$ is conserved,

$$\frac{d\langle \hat{Q}_t \rangle_t}{dt} = 0, \tag{2.13}$$

and the local property $\langle \hat{Q}_t \rangle(t,\mathbf{x})$ is subject to a local conservation law

$$\frac{\partial\langle\hat{Q}_t\rangle(t,\mathbf{x})}{\partial t}=\frac{i}{2\hbar}\operatorname{tr}([\hat{H}_t,\hat{P}(\mathbf{x})]_-[\hat{w}_t,\hat{Q}_t]_+). \quad (2.14)$$

[Note that (2.12), in the Schrödinger picture, is equivalent to $d\hat{Q}_t^H/dt=0$, where the Heisenberg operator $\hat{Q}_t^H=\hat{U}_{t,t_0}^{-1}\hat{Q}_t\hat{U}_{t,t_0}$, and $\hat{U}_{t,t_0}$ is defined in (2.16) below.] We consider several examples. The unit operator and, according to (2.2), $\hat{w}_t$ both satisfy (2.12). More generally, the solution of the von Neumann equation (2.2) can be written as

$$\hat{w}_t=\hat{U}_{t,t_0}\hat{w}_{t_0}\hat{U}_{t,t_0}^{-1}, \quad (2.15)$$

where $\hat{U}_{t,t_0}$ is a unitary operator satisfying

$$\frac{\partial\hat{U}_{t,t_0}}{\partial t}+\frac{i}{\hbar}\hat{H}_t\hat{U}_{t,t_0}=0,\quad \hat{U}_{t_0,t_0}=\hat{1}. \quad (2.16)$$

Therefore, for any integer $n\geqslant 0$,

$$\hat{w}_t^n=\hat{U}_{t,t_0}\hat{w}_{t_0}^n\hat{U}_{t,t_0}^{-1}. \quad (2.17)$$

Also

$$\ln\hat{w}_t=\hat{U}_{t,t_0}(\ln\hat{w}_{t_0})\hat{U}_{t,t_0}^{-1}. \quad (2.18)$$

According, $\hat{w}_t^n$ and $\ln\hat{w}_t$ satisfy (2.12). From (2.6) and (2.14), therefore, the local conservation laws for these properties are

$$\frac{\partial\operatorname{tr}(\hat{P}(\mathbf{x})\hat{w}_t^n)}{\partial t}=\frac{i}{\hbar}\operatorname{tr}([\hat{H}_t,\hat{P}(\mathbf{x})]_-\hat{w}_t^n),\quad n\geqslant 1, \quad (2.19)$$

$$\frac{\partial\operatorname{tr}(\hat{P}(\mathbf{x})\hat{S}_t)}{\partial t}=\frac{i}{\hbar}\operatorname{tr}([\hat{H}_t,\hat{P}(\mathbf{x})]_-\hat{S}_t). \quad (2.20)$$

Here $\hat{S}_t$ is the entropy operator,

$$\hat{S}_t=-\hat{w}_t\ln\hat{w}_t \quad (2.21)$$

and $\operatorname{tr}(\hat{P}(\mathbf{x})\hat{S}_t)$ is the local entropy in the microstate $\hat{P}(\mathbf{x})$. The entropy of a subsystem comprising several microstates is the sum of the local entropies in these microstates. The global entropy $\operatorname{tr}\hat{S}_t$ is the von Neumann entropy,[8] which is conserved when $\hat{w}_t$ evolves unitarily in accordance with (2.15).

If $\hat{Q}$ is any time-independent variable that commutes with $\hat{H}$ (also time independent), then $\hat{Q}$ satisfies (2.12). Therefore $\langle\hat{Q}\rangle_t$ is conserved according to (2.13)—a "constant of motion"—and $\langle\hat{Q}\rangle(t,\mathbf{x})$ obeys the local conservation law (2.14).

III. ONE-PARTICLE SYSTEM

The general discussion in Sec. II now will be applied to a specific system. This example will elucidate the physical meaning of the first term on the right-hand side of (2.11). We consider a system consisting of one particle, identify $\hat{\mathbf{x}}$ as the coordinate operator, and assume that

$$\hat{H}_t=\hat{\mathbf{p}}^2/2m+V_t(\hat{\mathbf{x}}). \quad (3.1)$$

Here $\hat{\mathbf{p}}$ is the momentum; m, the mass; and $V_t(\hat{\mathbf{x}})$, the potential energy at time t.

When $\hat{\mathbf{x}}$ is the coordinate operator, the local properties defined in (2.6) are fields. The spatial coordinates $\mathbf{x}$ for these fields are the eigenvalues of $\hat{\mathbf{x}}$ for the system. In quantum mechanics space comprises the global set of eigenvalues $\mathbf{x}$ of $\hat{\mathbf{x}}$ (see Ref. 5), in contrast to classical geometry, where coordinate variables are defined on a space that is the three-dimensional continuum of the real numbers.

We now specify (2.11) for the Hamiltonian (3.1) of the one-particle system. The first term on the right-hand side becomes

$$\begin{aligned}(i/2\hbar)&\operatorname{tr}([\hat{H}_t,\hat{P}(\mathbf{x})]_-[\hat{w}_t,\hat{Q}_t]_+)\\&=(i/4m\hbar)\operatorname{tr}([\hat{\mathbf{p}}^2,\hat{P}(\mathbf{x})]_-[\hat{w}_t,\hat{Q}_t]_+)\\&=-(i/\hbar)\operatorname{tr}([\hat{P}(\mathbf{x}),\hat{\mathbf{p}}]_-\cdot\hat{\mathcal{Q}}_t)=-\operatorname{tr}(\hat{P}(\mathbf{x})\hat{\nabla}\cdot\mathcal{Q}_t),\end{aligned} \quad (3.2)$$

where $\mathcal{Q}_t$ is the flux-vector operator for the transport current associated with the scalar field $\hat{Q}_t$,

$$\hat{\mathcal{Q}}_t=(4m)^{-1}[\hat{\mathbf{p}},[\hat{w}_t,\hat{Q}_t]_+]_+, \quad (3.3)$$

and $\hat{\nabla}$ is the commutation operator[5]

$$\hat{\nabla}(\cdots)=(i/\hbar)[\hat{\mathbf{p}},\cdots]_-. \quad (3.4)$$

Furthermore, in coordinate respresentation,

$$\hat{\mathbf{p}}=\frac{\hbar}{i}\int_{\mathbf{x}}|\mathbf{x}\rangle\frac{\partial}{\partial\mathbf{x}}\langle\mathbf{x}|. \quad (3.5)$$

Note: Differentiation with respect to discrete-valued variables has been discussed in earlier works.[5,7] Therefore,

$$\begin{aligned}\operatorname{tr}(\hat{P}(\mathbf{x})\hat{\nabla}\cdot\hat{\mathcal{Q}}_t)&=\int_{\mathbf{x}'}\Big(\langle\mathbf{x}|\mathbf{x}'\rangle\frac{\partial}{\partial\mathbf{x}'}\cdot\langle\mathbf{x}'|\hat{\mathcal{Q}}_t|\mathbf{x}\rangle\\&\quad-\langle\mathbf{x}|\hat{\mathcal{Q}}_t|\mathbf{x}'\rangle\cdot\frac{\partial}{\partial\mathbf{x}'}\langle\mathbf{x}'|\mathbf{x}\rangle\Big)=\frac{\partial}{\partial\mathbf{x}}\cdot\mathcal{Q}(t,\mathbf{x}),\\&\mathcal{Q}(t,\mathbf{x})=\operatorname{tr}(\hat{P}(\mathbf{x})\hat{\mathcal{Q}}_t).\end{aligned} \quad (3.6)$$

Thus, for the one-particle system, (2.11) becomes

$$\frac{\partial\langle\hat{Q}_t\rangle(t,\mathbf{x})}{\partial t}=-\frac{\partial}{\partial\mathbf{x}}\cdot\mathcal{Q}(t,\mathbf{x})+\left\langle\frac{\partial\hat{Q}_t}{\partial t}+\frac{i}{\hbar}[\hat{H}_t,\hat{Q}_t]_-\right\rangle(t,\mathbf{x}), \quad (3.7)$$

which has the form of an equation of continuity with source terms. From (3.3), (3.5), and (3.6), the local flux vector $\mathcal{Q}(t,\mathbf{x})$ is given by

$$\begin{aligned}\mathcal{Q}(t,\mathbf{x})&=(4m)^{-1}\operatorname{tr}(\hat{P}(\mathbf{x})[\hat{\mathbf{p}},[\hat{w}_t,\hat{Q}_t]_+]_+)\\&=\frac{\hbar}{4mi}\left\{\left(\frac{\partial}{\partial\mathbf{x}}\langle\mathbf{x}|\right)[\hat{w}_t,\hat{Q}_t]_+|\mathbf{x}\rangle\right.\\&\quad\left.-\langle\mathbf{x}|[\hat{w}_t,\hat{Q}_t]_+\left(\frac{\partial}{\partial\mathbf{x}}|\mathbf{x}\rangle\right)\right\}.\end{aligned} \quad (3.8)$$

For observables $\hat{Q}_t$ that satisfy (2.12), the source terms in (3.7) vanish, and the local conservation law (2.14) becomes the equation of continuity

$$\frac{\partial\langle\hat{Q}_t\rangle(t,\mathbf{x})}{\partial t}=-\frac{\partial}{\partial\mathbf{x}}\cdot\mathcal{Q}(t,\mathbf{x}). \quad (3.9)$$

For example, from (2.19) for $n=1$,

$$\frac{\partial\operatorname{tr}(\hat{P}(\mathbf{x})\hat{w}_t)}{\partial t}=-\frac{\partial}{\partial\mathbf{x}}\cdot\mathcal{J}(t,\mathbf{x}), \quad (3.10)$$

which is the equation of continuity for probability, with flux vector $\mathcal{J}(t,\mathbf{x})$ given by (3.8) ($\hat{Q}_t=\hat{1}$ in this case),

$$\begin{aligned}\mathcal{J}(t,\mathbf{x})&=(2m)^{-1}\operatorname{tr}(\hat{P}(\mathbf{x})[\hat{\mathbf{p}},\hat{w}_t]_+)\\&=\frac{\hbar}{2mi}\left\{\left(\frac{\partial}{\partial\mathbf{x}}\langle\mathbf{x}|\right)\hat{w}_t|\mathbf{x}\rangle-\langle\mathbf{x}|\hat{w}_t\left(\frac{\partial}{\partial\mathbf{x}}|\mathbf{x}\rangle\right)\right\}.\end{aligned} \quad (3.11)$$

In the pure case, in which $\hat{w}_t = |\psi_t\rangle\langle\psi_t|$ and $\langle\mathbf{x}|\psi_t\rangle = \psi_t(\mathbf{x})$, (3.10) and (3.11) give the well-known equation of continuity[6]

$$\frac{\partial\psi_t^*(\mathbf{x})\psi_t(\mathbf{x})}{\partial t} = -\frac{\partial}{\partial\mathbf{x}}\cdot\frac{\hbar}{2mi}\left\{\psi_t^*(\mathbf{x})\frac{\partial\psi_t(\mathbf{x})}{\partial\mathbf{x}} - \psi_t(\mathbf{x})\frac{\partial\psi_t^*(\mathbf{x})}{\partial\mathbf{x}}\right\}. \qquad (3.12)$$

Similarly, the local conservation law for entropy (2.20) becomes the equation of continuity,

$$\frac{\partial\,\mathrm{tr}(\hat{P}(\mathbf{x})\hat{S}_t)}{\partial t} = -\frac{\partial}{\partial\mathbf{x}}\cdot\mathscr{S}(t,\mathbf{x}), \qquad (3.13)$$

with entropy flux vector

$$\mathscr{S}(t,\mathbf{x}) = (2m)^{-1}\,\mathrm{tr}(\hat{P}(\mathbf{x})[\hat{\mathbf{p}},\hat{\mathbf{S}}_t]_+). \qquad (3.14)$$

We may obtain the rate of change of local properties for other observables of the one-particle system from (3.7). For example, for the momentum, with $\hat{\mathbf{v}} = \mathbf{p}/m$,

$$\frac{\partial\langle\hat{\mathbf{p}}\rangle(t,\mathbf{x})}{\partial t} = -\frac{\partial}{\partial\mathbf{x}}\cdot\frac{1}{4}\,\mathrm{tr}(\hat{P}(\mathbf{x})[\hat{\mathbf{v}},[\hat{w}_t,\hat{\mathbf{p}}]_+]_+) - \langle\hat{\nabla}V_t(\hat{\mathbf{x}})\rangle(t,\mathbf{x}). \qquad (3.15)$$

Note: We use the convention in the divergence term that the scalar product is between $\partial/\partial\mathbf{x}$ and the adjacent vector $\hat{\mathbf{v}}$ *as written*, so that $\partial/\partial\mathbf{x}$ dots $\hat{\mathbf{v}}$ in all four terms arising in expansion of $[\hat{\mathbf{v}},[\hat{w}_t,\hat{\mathbf{p}}]_+]_+$. In (3.15), we identify the stress tensor as the local symmetric tensor

$$\boldsymbol{\Psi}(t,\mathbf{x}) = (m/4)\,\mathrm{tr}(\hat{P}(\mathbf{x})[\hat{\mathbf{v}},[\hat{w}_t,\hat{\mathbf{v}}]_+]_+). \qquad (3.16)$$

The rate of increase of the local kinetic energy is

$$\frac{\partial\langle\frac{1}{2}m\hat{\mathbf{v}}^2\rangle(t,\mathbf{x})}{\partial t} = -\frac{\partial}{\partial\mathbf{x}}\cdot\frac{1}{4}\,\mathrm{tr}\left(\hat{P}(\mathbf{x})\left[\hat{\mathbf{v}},\left[\hat{w}_t,\frac{1}{2}m\hat{\mathbf{v}}^2\right]_+\right]_+\right) - \tfrac{1}{2}\langle[\hat{\mathbf{v}}\cdot,\hat{\nabla}V_t(\hat{\mathbf{x}})]_+\rangle(t,\mathbf{x}), \qquad (3.17)$$

with the kinetic energy flux vector related to the stress tensor of (3.16). The rate of increase of the local potential energy is

$$\frac{\partial\langle V_t(\hat{\mathbf{x}})\rangle(t,\mathbf{x})}{\partial t} = -\frac{\partial}{\partial\mathbf{x}}\cdot\frac{1}{4}\,\mathrm{tr}(\hat{P}(\mathbf{x})[\hat{\mathbf{v}},[\hat{w}_t,V_t(\hat{\mathbf{x}})]_+]_+) + \left\langle\frac{\partial V_t(\hat{\mathbf{x}})}{\partial t} + \frac{1}{2}[\hat{\mathbf{v}}\cdot,\hat{\nabla}V_t(\hat{\mathbf{x}})]_+\right\rangle(t,\mathbf{x}). \qquad (3.18)$$

The rate of increase of the local total energy, the sum of (3.17) and (3.18), is

$$\frac{\partial\langle\hat{H}_t\rangle(t,\mathbf{x})}{\partial t} = -\frac{\partial}{\partial\mathbf{x}}\cdot\frac{1}{4}\,\mathrm{tr}(\hat{P}(\mathbf{x})[\hat{\mathbf{v}},[\hat{w}_t,\hat{H}_t]_+]_+) + \left\langle\frac{\partial V_t(\hat{\mathbf{x}})}{\partial t}\right\rangle(t,\mathbf{x}). \qquad (3.19)$$

The example of a one-particle system illustrates that the local level of description of quantum systems, according to (2.6) and (2.11), provides a suitable context for treating thermodynamic problems. A subsystem containing the microstates belonging to a subset $[\mathbf{x}]_1$ of the eigenvalues $\mathbf{x}$ of the global system, constitutes a thermodynamic system. The remainder of the microstates belonging to the complementary subset of $\mathbf{x}$ is the "surroundings" of the thermodynamic system. This system plus surroundings is the thermodynamic "universe," the global system in our terminology. For example, spectral summation on $[\mathbf{x}]_1$ in (3.19) gives

$$\frac{d\langle\hat{H}_t\rangle_1}{dt} = -\oint\mathscr{U}(t,\mathbf{x})\cdot d\boldsymbol{\sigma} + \left\langle\frac{\partial V_t(\hat{\mathbf{x}})}{\partial t}\right\rangle_1, \qquad (3.20)$$

where

$$\langle\hat{H}_t\rangle_1 = \int_{[\mathbf{x}]_1}\langle\hat{H}_t\rangle(t,\mathbf{x}), \qquad \left\langle\frac{\partial V_t(\hat{\mathbf{x}})}{\partial t}\right\rangle_1 = \int_{[\mathbf{x}]_1}\left\langle\frac{\partial V_t(\hat{\mathbf{x}})}{\partial t}\right\rangle(t,\mathbf{x}), \qquad (3.21)$$

and

$$\oint\mathscr{U}(t,\mathbf{x})\cdot d\boldsymbol{\sigma} = \int_{[\mathbf{x}]_1}\frac{\partial}{\partial\mathbf{x}}\cdot\mathscr{U}(t,\mathbf{x}) \qquad (3.22)$$

is the integral over the boundary surfaces enclosing the microstates belonging to the subset $[\mathbf{x}]_1$, assumed to be suitably connected so that Gauss's law applies. This surface integral represents the transport of energy into the thermodynamic system from its surroundings by the energy flux vector,

$$\mathscr{U}(t,\mathbf{x}) = \tfrac{1}{4}\,\mathrm{tr}(\hat{P}(\mathbf{x})[\hat{\mathbf{v}},[\hat{w}_t,\hat{H}_t]_+]_+). \qquad (3.23)$$

If the Hamiltonian $\hat{H}$ [i.e., $V(\hat{\mathbf{x}})$] is time independent in (3.1), the energy of the global system is conserved, and, according to (3.20), the energy increase of the thermodynamic system arises entirely from transport of energy from the surrounding through the boundary surfaces.

IV. REALIZATION OF LATENT ENSEMBLES

In the preceding sections the statistical operator $\hat{w}_t$ is given by (2.15), a solution of the von Neumann equation (2.2), evolving unitarily in time. This is a globally isentropic process as noted after (2.21). A second mechanism, in addition to unitary transformation, for change of the statistical operator has been proposed in a recent paper[4]: realization of latent ensembles. It also is a global process, occurring irreversibly with non-negative entropy increase. We wish to consider the effects of this global process on the local properties of the system at a microstate $\hat{P}(\mathbf{x})$.

We first give an outline of the mechanism of realization of latent ensembles. The initial value of the statistical operator at t_0 is assumed to be a $\hat{\mathbf{v}}$ ensemble, characterized by the spectral representation

$$w(\hat{\mathbf{v}}) = \int_{\mathbf{v}}|\mathbf{v}\rangle w(\mathbf{v})\langle\mathbf{v}|, \quad \int_{\mathbf{v}}w(\mathbf{v}) = 1. \qquad (4.1)$$

Here $\hat{\mathbf{v}}$ is a complete commuting set of observables of the system with eigenvectors $|\mathbf{v}\rangle$ belonging to the nondegenerate set of eigenvalues $\mathbf{v}$ of $\hat{\mathbf{v}}$. An eigenvalue $w(\mathbf{v})$ of $w(\hat{\mathbf{v}})$ gives the initial probability of the observable $\hat{\mathbf{v}}$ in the $\mathbf{v}$ microstate; it is specified by the projector $\hat{P}(\mathbf{v})$,

$$w(\mathbf{v}) = \mathrm{tr}[w(\hat{\mathbf{v}})\hat{P}(\mathbf{v})], \quad \hat{P}(\mathbf{v}) = |\mathbf{v}\rangle\langle\mathbf{v}|. \qquad (4.2)$$

These probabilities are determined in a $\mathbf{v}$ measurement situation at t_0. The statistical operator evolves unitarily from its initial value $w(\hat{\mathbf{v}})$ in accordance with the von Neumann equation (2.2), so that for $t \geqslant t_0$ it becomes

$$w_{t,t_0}(\hat{\mathbf{v}}) = \hat{\mathscr{U}}_{t,t_0} w(\hat{\mathbf{v}}) \hat{\mathscr{U}}_{t,t_0}^{-1}, \tag{4.3}$$

according to (2.15) and (2.16). If $\hat{\mathbf{q}}$ is another complete commuting set of observables, the latent $\hat{\mathbf{q}}$ ensemble of the system whose statistical operator is $w_{t,t_0}(\hat{\mathbf{v}})$ is, by definition,

$$\begin{aligned} w_{t,t_0}(\hat{\mathbf{v}};\hat{\mathbf{q}}) &= \int_{\mathbf{q}} \mathrm{tr}\left[w_{t,t_0}(\hat{\mathbf{v}}) \hat{P}(\mathbf{q}) \right] \hat{P}(\mathbf{q}), \\ \hat{P}(\mathbf{q}) &= |\mathbf{q}\rangle\langle\mathbf{q}|. \end{aligned} \tag{4.4}$$

Thus, $w_{t,t_0}(\hat{\mathbf{v}};\hat{\mathbf{q}})$ is the diagonal projection of $w_{t,t_0}(\hat{\mathbf{v}})$ in $\hat{\mathbf{q}}$ representation. The global expectation values of the microstates $\hat{P}(\mathbf{q})$, and hence of any function of $\hat{\mathbf{q}}$, are given by

$$\begin{aligned} \langle \hat{P}(\mathbf{q}) \rangle_t &= \mathrm{tr}\left[w_{t,t_0}(\hat{\mathbf{v}}) \hat{P}(\mathbf{q}) \right] \\ &= \mathrm{tr}\left[w_{t,t_0}(\hat{\mathbf{v}};\hat{\mathbf{q}}) \hat{P}(\mathbf{q}) \right]. \end{aligned} \tag{4.5}$$

In the event of a $\hat{\mathbf{q}}$-measurement situation at $t = t_1 > t_0$, the latent $\hat{\mathbf{q}}$ ensemble is realized as the statistical operator $w_{t,t_1}(\hat{\mathbf{q}})$ for $t \geqslant t$, replacing the previous statistical operator $w_{t,t_0}(\hat{\mathbf{v}})$,

$$\begin{aligned} w_{t,t_1}(\hat{\mathbf{q}}) &= \hat{\mathscr{U}}_{t,t_1} w(\hat{\mathbf{q}}) \hat{\mathscr{U}}_{t,t_1}^{-1}, \\ \hat{\mathscr{U}}_{t,t_1} &= \mathscr{U}_{t,t_0} \hat{\mathscr{U}}_{t_1,t_0}^{-1}, \end{aligned} \tag{4.6}$$

where the initial value at t_1 of the new statistical operator is

$$w(\hat{\mathbf{q}}) = \int_{\mathbf{q}} |\mathbf{q}\rangle w(\mathbf{q}) \langle\mathbf{q}| = w_{t_1,t_0}(\hat{\mathbf{v}};\hat{\mathbf{q}}), \tag{4.7}$$

the same as the latent $\hat{\mathbf{q}}$ ensemble at the moment of realization, $t = t_1$. From (4.5), (4.3), and (4.1), it follows that

$$\begin{aligned} w(\mathbf{q}) &= \mathrm{tr}\left[w_{t_1,t_0}(\hat{\mathbf{v}}) \hat{P}(\mathbf{q}) \right] \\ &= \int_{\mathbf{v}} |\langle\mathbf{q}| \hat{\mathscr{U}}_{t_1,t_0} |\mathbf{v}\rangle|^2 w(\mathbf{v}). \end{aligned} \tag{4.8}$$

According to this mechanism the statistical operator is a succession of ensembles each evolving unitarily from its initial state, which is a realization of a latent ensemble in the preceding one.[4]

We now consider the effects of the realization at $t = t_1$ of the latent $\hat{\mathbf{q}}$ ensemble in (4.4), a global process, on the local properties at a microstate specified by $\hat{P}(\mathbf{x})$ in (2.4). For $t < t_1$, the local expection of $\hat{P}(\mathbf{q})$ at $\mathbf{x}$ for the system with statistical operator $w_{t,t_0}(\hat{\mathbf{v}})$ is given by (2.6) as

$$\langle \hat{P}(\mathbf{q}) \rangle (t < t_1,\mathbf{x}) = \tfrac{1}{2} \mathrm{tr}(\hat{P}(x) \left[w_{t,t_0}(\hat{\mathbf{v}}), \hat{P}(\mathbf{q}) \right]_+). \tag{4.9}$$

For $t \geqslant t_1$, for the system with statistical operator $w_{t,t_1}(\hat{\mathbf{q}})$, it is

$$\langle \hat{P}(\mathbf{q}) \rangle (t \geqslant t_1,\mathbf{x}) = \tfrac{1}{2} \mathrm{tr}(\hat{P}(\mathbf{x}) \left[w_{t,t_1}(\hat{\mathbf{q}}), \hat{P}(\mathbf{q}) \right]_+). \tag{4.10}$$

As $t \rightarrow t_1$, $\langle \hat{P}(\mathbf{q}) \rangle (t < t_1,\mathbf{x})$ in (4.9) approaches the limit

$$\begin{aligned} \langle \hat{P}(\mathbf{q}) \rangle (t_1^-,\mathbf{x}) &= \tfrac{1}{2} \mathrm{tr}(\hat{P}(\mathbf{x}) \left[w_{t_1,t_0}(\hat{\mathbf{v}}), \hat{P}(\mathbf{q}) \right]_+) \\ &= \tfrac{1}{2} \left[\langle\mathbf{x}| w_{t_1,t_0}(\hat{\mathbf{v}}) |\mathbf{q}\rangle \langle\mathbf{q}|\mathbf{x}\rangle \right. \\ &\quad \left. + \langle\mathbf{x}|\mathbf{q}\rangle \langle\mathbf{q}| w_{t_1,t_0}(\hat{\mathbf{v}}) |\mathbf{x}\rangle \right] \end{aligned} \tag{4.11}$$

and $\langle \hat{P}(\mathbf{q}) \rangle (t \geqslant t_1,\mathbf{x})$ in (4.10) becomes

$$\begin{aligned} \langle \hat{P}(\mathbf{q}) \rangle (t_1^+,\mathbf{x}) &= \tfrac{1}{2} \mathrm{tr}(\hat{P}(\mathbf{x}) \left[w_{t_1,t_0}(\hat{\mathbf{v}};\hat{\mathbf{q}}), \hat{P}(\mathbf{q}) \right]_+) \\ &= |\langle\mathbf{x}|\mathbf{q}\rangle|^2 \langle\mathbf{q}| w_{t_1,t_0}(\hat{\mathbf{v}}) |\mathbf{q}\rangle, \end{aligned} \tag{4.12}$$

which is the local expectation of $\hat{P}(\mathbf{q})$ in the latent $\hat{\mathbf{q}}$ ensemble at t_1. In general, these two expectation values at t_1 are not equal, although when summed over $\mathbf{x}$ they give the same global expectation values according to (4.5). Accordingly, when the latent $\hat{\mathbf{q}}$ ensemble is realized at t_1, the local expectation value of $\hat{P}(\mathbf{q})$ at $\mathbf{x}$ changes discontinuously from its value in (4.11) to its value in (4.12). At $t = t_1$ a redistribution of local expectation values occurs among the spectral points $\mathbf{x}$ of $\hat{\mathbf{x}}$, while the global expectation value is unchanged according to (4.5). An exception is the case in which the set $\hat{\mathbf{q}}$ is the same as $\hat{\mathbf{x}}$. In this case with $\mathbf{q} = \mathbf{x}'$, another eigenvalue of $\hat{\mathbf{x}}$, the expectation value of $\hat{P}(\mathbf{x}')$ is the same in (4.11) and (4.12), namely,

$$\langle \hat{P}(\mathbf{x}') \rangle (t_1,\mathbf{x}) = \langle\mathbf{x}|\mathbf{x}'\rangle \langle\mathbf{x}| w_{t_1,t_0}(\hat{\mathbf{v}}) |\mathbf{x}\rangle \tag{4.13}$$

so that no redistribution occurs at t_1; the local expectation at $\mathbf{x}$ of any function of $\hat{\mathbf{x}}$ changes continuously from its value in $w_{t,t_0}(\hat{\mathbf{v}})$ for $t > t_1$ to its value in $w_{t,t_1}(\hat{\mathbf{x}})$ for $t \geqslant t_1$.

The von Neumann entropy of the system is given by $\mathrm{tr}\, \hat{S}_t$ in (2.21). From (4.3) this global entropy for $t < t_1$ is represented by

$$\begin{aligned} S\left[w_{t,t_0}(\hat{\mathbf{v}}) \right] &= -\mathrm{tr}\left[w_{t,t_0}(\hat{\mathbf{v}}) \ln w_{t,t_0}(\hat{\mathbf{v}}) \right] \\ &= -\int_{\mathbf{v}} w(\mathbf{v}) \ln w(\mathbf{v}), \end{aligned} \tag{4.14}$$

and for $t \geqslant t_1$, after realization of the latent $\hat{\mathbf{q}}$ ensemble, by

$$S\left[w_{t,t_1}(\hat{\mathbf{q}}) \right] = -\int_{\mathbf{q}} w(\mathbf{q}) \ln w(\mathbf{q}). \tag{4.15}$$

These expressions are independent of time, while the statistical operator evolves unitarily. As shown previously[4]

$$S\left[w_{t,t_1}(\hat{\mathbf{q}}) \right] - S\left[w_{t,t_0}(\hat{\mathbf{v}}) \right] \geqslant 0. \tag{4.16}$$

This non-negative global entropy increase attends the realization of the latent $\hat{\mathbf{q}}$ ensemble at t_1. It is of interest to consider the local entropy changes that this realization entails.

The local entropy at $(t,\mathbf{x})$ is given by $\mathrm{tr}(\hat{P}(\mathbf{x}) \hat{S}_t)$ in (2.21). We write it for $t \leqslant t_1$ as

$$\begin{aligned} S\left[w_{t,t_0}(\hat{\mathbf{v}}) \right](t,\mathbf{x}) &= -\mathrm{tr}(\hat{P}(\mathbf{x}) w_{t,t_0}(\hat{\mathbf{v}}) \ln w_{t,t_0}(\hat{\mathbf{v}})) \\ &= -\int_{\mathbf{v}} |\langle\mathbf{x}| \hat{U}_{t,t_0} |\mathbf{v}\rangle|^2 w(\mathbf{v}) \ln w(\mathbf{v}), \end{aligned} \tag{4.17}$$

and for $t \geqslant t_1$, as

$$S\left[w_{t,t_1}(\hat{\mathbf{q}}) \right](t,\mathbf{x}) = -\int_{\mathbf{q}} |\langle\mathbf{x}| \hat{U}_{t,t_1} |\mathbf{q}\rangle|^2 w(\mathbf{q}) \ln w(\mathbf{q}). \tag{4.18}$$

Spectral summation on $\mathbf{x}$ in (4.17) and (4.18) gives back the global entropies (4.14) and (4.15), respectively. The increase in local entropy at $\mathbf{x}$, on realization of the latent $\hat{\mathbf{q}}$ ensemble, is the difference in these expectations evaluated at t_1,

$$\begin{aligned} & S\left[w_{t,t_1}(\hat{\mathbf{q}}) \right](t_1,\mathbf{x}) - S\left[w_{t,t_0}(\hat{\mathbf{v}}) \right](t,\mathbf{x}) \\ &\quad = -\int_{\mathbf{q}} |\langle\mathbf{x}|\mathbf{q}\rangle|^2 w(\mathbf{q}) \ln w(\mathbf{q}) \\ &\qquad + \int_{\mathbf{v}} |\langle\mathbf{x}| \hat{U}_{t_1,t_0} |\mathbf{v}\rangle|^2 w(\mathbf{v}) \ln w(\mathbf{v}). \end{aligned} \tag{4.19}$$

This increase can be interpreted as follows. Define a local entropy quantity, denoted by $S\left[w_{t,t_0}(\hat{\mathbf{v}});\hat{\mathbf{q}}\right](t,\mathbf{x})$, for $t<t_1$:

$$\begin{aligned}
&S\left[w_{t,t_0}(\hat{\mathbf{v}});\hat{\mathbf{q}}\right](t,\mathbf{x})\\
&\quad=-\int_{\mathbf{q}}\operatorname{tr}(\widehat{P}(\mathbf{x})\widehat{P}(\mathbf{q}))\operatorname{tr}(\widehat{P}(\mathbf{q})w_{t,t_0}(\hat{\mathbf{v}}))\ln w_{t,t_0}(\hat{\mathbf{v}})\\
&\quad=-\int_{\mathbf{q}}\int_{\mathbf{v}}|\langle\mathbf{x}|\mathbf{q}\rangle|^2|\langle\mathbf{q}|\widehat{U}_{t,t_0}|\hat{\mathbf{v}}\rangle|^2 w(\mathbf{v})\ln w(\mathbf{v}). \qquad (4.20)
\end{aligned}$$

This is the local value at $\mathbf{x}$ of the diagonal projection of the entropy operator $\widehat{S}_t=-w_{t,t_0}(\hat{\mathbf{v}})\ln w_{t,t_0}(\hat{\mathbf{v}})$ in $\hat{\mathbf{q}}$ representation for $t<t_1$. The spectral summations on $\mathbf{x}$ of $S\left[w_{t,t_0}(\hat{\mathbf{v}})\right]\times(t,\mathbf{x})$ in (4.17) and $S\left[w_{t,t_0}(\hat{\mathbf{v}});\hat{\mathbf{q}}\right](t,\mathbf{x})$ in (4.20) both give the global value in (4.14),

$$\begin{aligned}
\int_{\mathbf{x}}S\left[w_{t,t_0}(\hat{\mathbf{v}})\right](t,\mathbf{x})&=\int_{\mathbf{x}}S\left[w_{t,t_0}(\hat{\mathbf{v}});\hat{\mathbf{q}}\right](t,\mathbf{x})\\
&=S\left[w_{t,t_0}(\hat{\mathbf{v}})\right]. \qquad (4.21)
\end{aligned}$$

Therefore, the local values $S\left[w_{t,t_0}(\hat{\mathbf{v}})\right](t,\mathbf{x})$ and $S\left[w_{t,t_0}(\hat{\mathbf{v}});\hat{\mathbf{q}}\right](t,\mathbf{x})$ represent two distributions among the spectral points $\mathbf{x}$ of the global entropy at $t<t_1$. Furthermore from (4.8), (4.18), and (4.20), at $t=t_1$,

$$\begin{aligned}
&S\left[w_{t,t_1}(\hat{\mathbf{q}})\right](t_1,\mathbf{x})\\
&\quad=-\int_{\mathbf{q}}|\langle\mathbf{x}|\mathbf{q}\rangle|^2 w(\mathbf{q})\ln w(\mathbf{q})\\
&\quad\geqslant-\int_{\mathbf{q}}\int_{\mathbf{v}}|\langle\mathbf{x}|\mathbf{q}\rangle|^2|\langle\mathbf{q}|\widehat{U}_{t_1,t_0}|\mathbf{v}\rangle|^2 w(\mathbf{v})\ln w(\mathbf{v})\\
&\quad=S\left[w_{t,t_0}(\hat{\mathbf{v}});\hat{\mathbf{q}}\right](t_1,\mathbf{x}), \qquad (4.22)
\end{aligned}$$

where the inequality follows since $-w(\mathbf{v})\ln w(\mathbf{v})$ is a concave function, and for any concave function $f(x)$,

$$\sum_i\lambda_i f(x_i)\leqslant f\left(\sum_i\lambda_i x_i\right), \qquad (4.23)$$

for $\Sigma_i\lambda_i=1$ and $\lambda_i\geqslant 0$. Accordingly, the increase in local entropy in (4.19) can be written as the sum of two terms. The first term,

$$S\left[w_{t,t_0}(\hat{\mathbf{v}});\hat{\mathbf{q}}\right](t_1,\mathbf{x})-S\left[w_{t,t_0}(\hat{\mathbf{v}})\right](t_1,\mathbf{x}), \qquad (4.24)$$

is the change at t_1 from the redistribution of local entropy expectation values among the spectral points $\mathbf{x}$, a redistribution that leaves the global entropy unchanged according to (4.21); this discontinuity in local entropy at $\mathbf{x}$ is analogous to the discontinuity in the local property $\langle\widehat{P}(\mathbf{q})\rangle(t,\mathbf{x})$ at t_1 discussed above following (4.12). The second term,

$$\mathrm{S}\left[w_{t,t_1}(\hat{\mathbf{q}})\right](t_1,\mathbf{x})-S\left[w_{t,t_0}(\hat{\mathbf{v}});\hat{\mathbf{q}}\right](t_1,\mathbf{x})\geqslant 0, \qquad (4.25)$$

is a non-negative local entropy production at $(t_1,\mathbf{x})$ according to (4.22). Summation on $\mathbf{x}$ in (4.25) gives the non-negative global entropy increase in (4.16). In the special case in which $\hat{\mathbf{q}}$ is the same as $\hat{\mathbf{x}}$, the term in (4.24) vanishes so that, as in the analogous case of the local property $\langle\widehat{P}(\mathbf{q})\rangle(t,\mathbf{x})$, no redistribution of local entropy occurs at t_1, but the non-negative local entropy production according to (4.25) remains.

The results of Sec. IV do not depend on the assumptions of the one-particle problem that was considered in Sec. III.

[1] J. von Neumann, *Mathematical Foundations of Quantum Mechanics*, translated by R. T. Beyer (Princeton U. P., Princeton, NJ, 1955).

[2] P. A. M. Dirac, *The Principles of Quantum Mechanics* (Clarendon, Oxford, 1947), 3rd ed., p. 57.

[3] A. Böhm, *Quantum Mechanics* (Springer, New York, 1979), Chap. IV.

[4] B. Leaf, J. Math. Phys. **26**, 1337 (1985).

[5] B. Leaf, J. Math. Phys. **25**, 535 (1984).

[6] See almost any quantum mechanics text, e.g., D. Park, *Introduction to Quantum Theory* (McGraw-Hill, New York, 1974), 2nd ed., p. 64.

[7] B. Leaf, Found. Phys. **12**, 583 (1982).

[8] A. Wehrl, Rev. Mod. Phys. **50**, 221 (1978).

Wave vector dependent susceptibility of a free electron gas in D dimensions and the singularity at $2k_F$

N. L. Sharma and M. Howard Lee
Department of Physics, University of Georgia, Athens, Georgia 30602

(Received 21 November 1985; accepted for publication 21 February 1986)

The static susceptibility of a free electron gas in D dimensions at $T=0$ is obtained by techniques of dimensional regularization. Our solutions for the susceptibility $\chi(k,D)$ are given in terms of the hypergeometric function. For any integer dimensions analytic expressions are possible. The high- and low-k series solutions are shown to be related by an analytic continuation if D is an odd integer, but not related if D is an even integer. The singularity at $2k_F$ is a branch point, whereupon the series solutions are absolutely convergent, yielding $\chi(k=2k_F,D)=(D-1)^{-1}$. The relationship of χkD has the appearance of a PVT diagram.

I. INTRODUCTION

The wave vector k dependent susceptibility $\chi(k)$ is a basic physical quantity in many-body physics.[1] It enters into a variety of physical relationships, e.g., dispersion relations, scaling laws. For a free electron gas this quantity is exactly known in spatial dimensions $D=1,2$, and 3.[2–6] Especially interesting is its singular behavior at $k=2k_F$ due to what is known as the Pauli blocking, where k_F is the Fermi wave vector. This singularity is responsible for the Kohn anomaly in the phonon spectrum of a metal.[7] The singularity in the susceptibility at $2k_F$ is D dependent. In $D=1$ the susceptibility has a logarithmic divergence. In $D=2$ the susceptibility is finite but its slope is discontinuous. In $D=3$ the slope has a logarithmic divergence. This trend suggests that the strength of the singularity becomes weaker with increasing dimensions. A precise knowledge of the D dependence would be of interest.

In addition to its own intrinsic interest, the susceptibility for a free electron gas is useful in other ways. If electrons now freely interact pairwise via the Coulomb force, the susceptibility for the interacting electron gas always can be written in the following form: $\chi^{\text{int}}(k)=\chi(k)/(1+\Lambda_k\chi(k))$, where Λ_k is some function of the interaction.[8] If $\Lambda_k=v_k$, where v_k is the Fourier transform of the Coulomb potential, one gets the simple random phase approximation (RPA) theory. If $\Lambda_k=v_k(1-G_k)$, where G_k is a local field term, one recovers the generalized RPA theory. Hence, the knowledge of the susceptibility for a free electron gas is essential to these RPA theories.

The above idea may be extended to the linear response theory of dynamic processes since $\chi(k)=\chi(k,\omega=0)$, where ω is the frequency. In dynamic theories, the knowledge of the static susceptibility is always presupposed.[1] For example, the relaxation function is normalized with respect to the static susceptibility. Moment sum rules are expressible in terms of the static susceptibility.[9]

More subtle is that the susceptibility may be defined by the Kubo scalar product (see Sec. II). The Kubo scalar product is an inner product that *realizes* an abstract Hilbert space. In this realized space time-dependent quantum statistical problems are all definable.[10] Hence, the existence of the susceptibility plays a central role in the study of time evolution of dynamic variables.[11] In such a study there is the possibility that the relaxation function may assume a mean-field form in all spatial dimensions greater than a certain critical value.[12] This kind of dynamic anomaly is signaled by a deformation of the realized Hilbert space.[13] Furthermore, the critical dimension may take on a noninteger value. The physics of noninteger dimensions is of current theoretical interest. See, e.g., fractals,[14] ϵ expansions,[15] kinetics of formation.[16]

The evaluation of the susceptibility for higher integer dimensions, e.g., $D=4$ may be carried out as was for $D=1$–3 (see Secs. II and III). We shall use techniques of dimensional regularization developed in particle physics[17] to obtain a solution for the susceptibility that is valid for any D, integers and nonintegers. This solution $\chi(k,D)$ might be viewed as, e.g., the PVT diagram of a homogeneous fluid. It traces a contour, which is a map of a continuous surface. This map is naturally divided into two regions (high k and low k) by the D line at $k=2k_F$. By moving alongside of this boundary line, one can examine the mathematical nature of the singularity at $2k_F$.

We find that $\chi(2k_F,D)=1/(D-1)$, $D>1$; $\chi'(2k_F,D)=-(D-2)/(D-1)(D-3)$, $D>3$, etc., where $\chi'=\partial\chi/\partial k$. We also find that the susceptibility is of two families, D odd and D even. For D odd, the singularity at $2k_F$ is all logarithmic in origin. For D even (high-k region) the singularity is of the square root. For D odd, the solution in one region is an analytic continuation into the other and $2k_F$ is a branch point. For D even, there are no such relationships and $k=2k_F$ is a branch point only at the high-k region.

II. STATIC SUSCEPTIBILITY

A free electron gas is described by the following Hamiltonian:

$$H=\sum_k\epsilon_k\, c_k^\dagger\, c_k\,, \tag{1}$$

where $\epsilon_k=k^2/2m$, m is the mass of the electron, and $c_k^\dagger$ and c_k are, respectively, the fermion creation and annihilation operators at wave vector k. Our units are such that $\hbar=1$. The longitudinal response to a weak static density-coupled perturbation is the static susceptibility given by the Kubo scalar product[9] (KSP),

0022-2488/86/061618-06$02.50

$$\chi(k) = (\rho_k, \rho_k) = \int_0^\beta d\lambda \, \langle e^{\lambda H} \rho_k e^{-\lambda H} \rho_k^\dagger \rangle \,, \tag{2}$$

where β is the inverse temperature, the brackets $\langle \cdots \rangle$ denote an ensemble average, † denotes Hermitian conjugation, and ρ_k is the density fluctuation operator defined as

$$\rho_k = \sum_p c_p^\dagger c_{p+k} \,. \tag{3}$$

For a free electron gas the KSP may be reduced to the well-known form

$$\chi(k) = 2 \sum_p \frac{f_p - f_{p+k}}{\epsilon_{p+k} - \epsilon_p} \,, \tag{4}$$

where f_k is the Fermi function. Converting the sum into an integral we can rewrite (4) in spatial dimensions D as

$$\chi_D(k) = 4(2\pi)^{-D} \int d^D p \, \frac{f_p}{\epsilon_{p+k} - \epsilon_p} \,. \tag{5}$$

In the second term of (4) we make use of the fact that f_p and ϵ_p are both functions of $|p|$. At $T = 0$ the Fermi function is a step function, i.e., $f_k = \theta(k_{\mathrm{F}} - k)$. Hence, for D small integers one can directly evaluate (5). For $D = 1$–3, the susceptibility is already known.[2–6] For comparison purposes, we shall list its normalized values $\widetilde{\chi}_D(k)/\chi_D(0)$, expressing k in units of k_{F}:

$$\widetilde{\chi}_1(k) = k^{-1} \ln |(2+k)/(2-k)| \,, \tag{6a}$$

$$\widetilde{\chi}_2(k) = 1 - \theta(k-2)(1 - 4/k^2)^{1/2} \,, \tag{6b}$$

$$\widetilde{\chi}_3(k) = \tfrac{1}{2} \left[1 + k^{-1}(1 - \tfrac{1}{4}k^2) \ln |(2+k)/(2-k)| \right] . \tag{6c}$$

III. EVALUATION OF THE SUSCEPTIBILITY FOR $D = 4$

The static susceptibility for $D = 4$ may be written down from (5):

$$\chi_4(k) = \frac{mk_{\mathrm{F}}^2}{\pi^3 k} \int_0^1 dp \, p^3 \int_0^{2\pi} d\Theta \, \frac{\sin^2 \Theta}{k + 2p \cos \Theta} \,. \tag{7}$$

We shall consider the angular integral first (denoted by Q). It may be converted into a contour integral on the unit circle by the substitution $u = e^{i\Theta}$,

$$Q = \frac{i}{4p} \oint du \, \frac{(u^2 - 1)^2}{u^2 (u - u_+)(u - u_-)} \,, \tag{8}$$

where

$$u_\pm = -(k/2p) \pm ((k/2p)^2 - 1)^{1/2} \,.$$

The zeros of the denominator are 0, u_+, and u_-. The zeros u_+ and u_- may be real or complex depending on whether $|k/2p|$ is greater or less than 1.

(i) $|k/2p| > 1$: The conjugate zeros u_+ and u_- are real, lying, respectively, inside and outside the unit circle. Hence, the zero u_- does not contribute to the integral. The residues at 0 and u_+ are, respectively, $-k/p$ and $2((k/2p)^2 - 1)^{1/2}$. Together,

$$Q = (\pi k/2p^2)[1 - (1 - (2p/k)^2)^{1/2}] \,.$$

(ii) $|k/2p| < 1$: The conjugate zeros are now complex and lie on the contour of integration. If we take the Cauchy principal value, their contributions cancel each other exactly, leaving only the zero at the origin to contribute to the integral, giving $Q = \pi k/2p^2$.

Both may be combined to read as

$$Q = (\pi k/2p^2)[1 - \theta(k^2 - 4p^2)(1 - (2p/k)^{1/2})] \,. \tag{9}$$

Using (9) in (7) we can now complete the radial part:

$$\begin{aligned} \widetilde{\chi}_4(k) &\equiv \chi_4(k)/\chi_4(0) \\ &= 1 - (k^2/6)[1 - \theta(k-2)(1 - 4/k^2)^{3/2}] \,, \end{aligned} \tag{10}$$

where $\chi_4(0) = mk_{\mathrm{F}}^2/4\pi^2$. We observe that the suceptibility and its derivative are both finite and continuous at $k = 2k_{\mathrm{F}}$.

For D even, generally the same idea may be used to evaluate the angular integral. There always is one pole of order $D - 2$ at the origin. The conjugate poles u_+ and u_- behave in the same manner as described for $D = 4$. For D odd, one cannot avail of this simplification and must resort to, e.g., integration by parts. In any event, the evaluation of the susceptibility by this standard approach becomes very tedious as $D \to \infty$. Also one is limited to integer dimensions only. We shall, therefore, consider another approach, i.e., dimensional regularization,[17] which may allow us to obtain the susceptibility possibly more simply and, more important, in any dimension.

IV. DIMENSIONAL REGULARIZATION

From (5) it is possible to express the zero temperature susceptibility for $D \geqslant 3$ as follows[18]:

$$\begin{aligned} \chi_D(k) &= A \int_0^1 dp \, p^{D-1} \int_0^\pi d\theta \, (\sin\theta)^{D-2} \left(\frac{k}{2} + p \cos\theta \right)^{-1} \\ &\equiv AI \,, \end{aligned} \tag{11}$$

where $A = A(k,D) = 2k_{\mathrm{F}}^D S_{D-1}/k\epsilon_{\mathrm{F}}(2\pi)^D$, where $S_D = 2(\Gamma(\tfrac{1}{2}))^D/\Gamma(\tfrac{1}{2}D)$, and k is in units of k_{F}. To evaluate this double integral, we exchange the order of integration. For $k > 2$ the integrand is well behaved in the given interval of p. One can, therefore, expand it in powers of $(2/k)$ and carry out the integration term by term. If, for $k < 2$, one attempts to expand it in powers of $(k/2)$, one encounters a pole in the interval of p. To avoid this apparent difficulty, we consider the following integral:

$$I_s = \int_0^\pi d\theta \, (\sin\theta)^{D-2} \int_0^1 dp \, p^{D-1} \left(\frac{k}{2} + p\cos\theta \right)^{2s-1} . \tag{12}$$

If we assume s is a positive integer, the new integrand is now well behaved in the interval of p for any k. Hence, one may expand it binomially and complete the integration term by term. Then one may possibly analytically continue I_s to obtain $I_0 \equiv I$. Clearly for $K > 2$ this process it unnecessary, hence it can be used as a direct test.

For any positive integer s, we obtain

$$I_s = \sum_{n=0}^{2s-1} \frac{(\tfrac{1}{2}k)^n \Gamma(2s)}{(D + 2s - n - 1)\Gamma(2s - n)\Gamma(n+1)} \times \int_0^\pi d\theta \, (\sin\theta)^{D-2} (\cos\theta)^{2s-n-1} \,, \tag{13}$$

where Γ is the gamma function. Next the angular integration, although cumbersome, is straightforward. Terms of even n vanish and terms of odd n are the beta functions. Letting $n \to 2n-1$, we get

$$I_s = \frac{1}{2} \sum_{n=1}^{s} \left(\frac{k}{2}\right)^{2n-1} \times \frac{\Gamma(2s)\Gamma(\frac{1}{2}D - \frac{1}{2})\Gamma(s-n+\frac{1}{2})}{\Gamma(2n)\Gamma(2s-2n+1)\Gamma(s-n+\frac{1}{2}D+1)} . \tag{14}$$

The above expression is appropriate for the $k<2$ expansion. To obtain an expression more suitable for the $k>2$ expansion, we rewrite (14) in ascending powers of $(2/k)$ by letting $s-n \to n$,

$$I_s = \frac{1}{2} \left(\frac{k}{2}\right)^{2s-1} \sum_{n=0}^{s-1} \left(\frac{2}{k}\right)^{2n} \times \frac{\Gamma(2s)\Gamma(\frac{1}{2}D - \frac{1}{2})\Gamma(n+\frac{1}{2})}{\Gamma(2s-2n)\Gamma(2n+1)\Gamma(n+\frac{1}{2}D+1)} . \tag{15}$$

Both expressions [(14) and (15)] are clearly well defined for any finite positive integer s. There are s terms in each expansion. We now take advantage of the gamma functions present in our expansions to perform analytic continuation. We first note that, for any $r>0$,

$$\lim_{t \to -1} \Gamma(t+1) = t\Gamma(t) = (-1)\cdots(-2)\cdots(-r+1)(-r)\Gamma(-r) .$$

Hence

$$\lim_{s \to 0} \frac{\Gamma(2s)}{\Gamma(2n)\Gamma(2s-2n+1)} = \frac{(-1)(-2)\cdots(-2n+1)\Gamma(-2n+1)}{\Gamma(2n)\Gamma(-2n+1)} = -1$$

and

$$\lim_{s \to 0} \frac{\Gamma(2s)}{\Gamma(2s-2n)\Gamma(2n+1)} = \frac{(-1)(-2)\cdots(-2n)\Gamma(-2n)}{\Gamma(-2n)\Gamma(2n+1)} = 1.$$

Using these results we get[19]

$$I_0 = -\frac{1}{2}\Gamma\left(\frac{1}{2}D - \frac{1}{2}\right) \sum_{n=1}^{\infty} \left(\frac{k}{2}\right)^{2n-1} \times \frac{\Gamma(-n+\frac{1}{2})}{\Gamma(-n+\frac{1}{2}D+1)}, \quad k<2, \tag{16a}$$

$$= \frac{1}{2}\Gamma\left(\frac{1}{2}D - \frac{1}{2}\right) \sum_{n=0}^{\infty} \left(\frac{2}{k}\right)^{2n+1} \times \frac{\Gamma(n+\frac{1}{2})}{\Gamma(n+\frac{1}{2}D+1)}, \quad k>2 . \tag{16b}$$

Finally, using the definition for A, $\chi(k=0,D) = (k_F/2\pi)^D S_D/\epsilon_F$ (see Ref. 20), where

$$S_D/S_{D-1} = \Gamma(\tfrac{1}{2})\,\Gamma(\tfrac{1}{2}D - \tfrac{1}{2})/\Gamma(\tfrac{1}{2}D) ,$$

we get

$$\widetilde{\chi}(k,D) \equiv \chi(k,D)/\chi(0,D)$$

$$= -\frac{1}{2} \frac{\Gamma(\frac{1}{2}D)}{\Gamma(\frac{1}{2})} \sum_{n=1}^{\infty} \left(\frac{k}{2}\right)^{2n-2} \times \frac{\Gamma(-n+\frac{1}{2})}{\Gamma(-n+\frac{1}{2}D+1)}, \quad k<2, \tag{17a}$$

$$= \frac{1}{2} \frac{\Gamma(\frac{1}{2}D)}{\Gamma(\frac{1}{2})} \sum_{n=0}^{\infty} \left(\frac{2}{k}\right)^{2n+2} \frac{\Gamma(n+\frac{1}{2})}{\Gamma(n+\frac{1}{2}D+1)}, \quad k>2 . \tag{17b}$$

Since the gamma functions are well defined for any arguments other than zero or negative integers, our solutions (17a) and (17b) are applicable to any value of D. Our series solutions agree term by term with the high- and low-k expansions of the susceptibility for $D = 1$–4 [Eqs. (6) and (10)] previously obtained by a conventional method.

V. APPLICATION OF THE HYPERGEOMETRIC FUNCTION

It is possible to express the susceptibility series [Eqs. (17a) and (17b)] in terms of the hypergeometric series $F(a,b;c;t)$ defined as follows[21]:

$$F(a,b;c;t) = \sum_{n=0}^{\infty} \frac{a_n b_n}{c_n n!}, \quad |t|<1, \tag{18}$$

where $a_n = \Gamma(n+a)/\Gamma(a)$, etc., $c \neq 0, -1, -2, \ldots$. The advantages of having the susceptibility given in the hypergeometric function (hgf) are evident. One can obtain analytic representations for integer dimensions. Properties of the hgf may be used to study the behavior of the susceptibility at the singular point $2k_F$. The high- and low-k expansions may be related through an analytic continuation.

For this purpose, we introduce $z = (\frac{1}{2}k)^2$ and let $\widetilde{\chi}_1(z,D) = \widetilde{\chi}(z<1,D)$ and $\widetilde{\chi}_2(z,D) = \widetilde{\chi}(z>1,D)$. We shall consider $\widetilde{\chi}_1(z,D)$ first. Using the identity

$$\Gamma(t-n) = (-1)^n \Gamma(t)\Gamma(-t+1)/\Gamma(-t+n+1)$$

in (17a) we obtain after some manipulations

$$\widetilde{\chi}(z,D) = -\frac{\Gamma(\frac{1}{2})}{D\Gamma(-\frac{1}{2}D)} \sum_{n=0}^{\infty} z^n \frac{\Gamma(n+1-\frac{1}{2}D)}{\Gamma(n+\frac{3}{2})} = F(1,1-\tfrac{1}{2}D;\tfrac{3}{2};z) . \tag{19a}$$

Similarly, we obtain from (17b) with the aid of (18)

$$\widetilde{\chi}_2(z,D) = D^{-1} z^{-1} F(1,\tfrac{1}{2};1+\tfrac{1}{2}D;z^{-1}) . \tag{19b}$$

Hence, together we have

$$\widetilde{\chi}(z,D) = F(1,1-\tfrac{1}{2}D;\tfrac{3}{2};z), \quad z<1 , \tag{20a}$$

$$= D^{-1} z^{-1} F(1,\tfrac{1}{2};1+\tfrac{1}{2}D;z^{-1}), \quad z>1 . \tag{20b}$$

We observe that for $D=1$ the high and low sides of the susceptibility have the same parameters of the hgf: $a=1$, $b=\frac{1}{2}$, $c=\frac{3}{2}$. For these values the hgf has an analytic representation

$$F(1,\tfrac{1}{2};\tfrac{3}{2};t^2) = \tfrac{1}{2} t^{-1} \ln(1+t)/(1-t), \quad |t|<1 . \tag{21}$$

The resulting susceptibility is in exact agreement with the $D=1$ result [Eq. (6a)].

To obtain analytic representations of the susceptibility for other integer dimensions, we study the hgf. Consider $F(1,1-\frac{1}{2}D;\frac{3}{2};t)$ first. For D even, $b=0,-1,-2,\ldots$, and for D odd, $b=\frac{1}{2},-\frac{1}{2},-\frac{3}{2},\ldots$. Hence, for D even, the hgf is a polynomial. For D odd, the hgf is *contiguous*,[21] i.e.,

$$F_{b-1}=[\tfrac{1}{2}+(1-b)(1-t)F_b]/(\tfrac{3}{2}-b)\,, \tag{22}$$

where $F_b=F(1,\mathrm{b};\frac{3}{2};t)$. Hence, using the known form for $F_{1/2}$, one can generate all others readily. Now since $F_{1/2}$ contains a logarithmic singularity [see (21)], all odd-dimensioned low-k susceptibility contains the same singularity.

We next consider $F(1,\frac{1}{2};1+\frac{1}{2}D;t)$. For D even (including 0), $c=1,2,3,\ldots$, and for D odd, $c=\frac{3}{2},\frac{5}{2},\frac{7}{2},\ldots$. In both cases the hgf is again contiguous,

$$F_{c+1}=c[1-(1-t)F_c]/(c-\tfrac{1}{2})t\,, \tag{23}$$

where $F_c=F(1,\frac{1}{2};c;t)$. Hence, now there are two "seeds," F_1 and $F_{3/2}$, where

$$F_1=F(1,\tfrac{1}{2};1;t)=(1-t)^{-1/2} \tag{24}$$

and $F_{3/2}$ is already given [see (21)]. Thus all the even-dimensioned high-k susceptibility has a square root singularity, while the odd-dimensioned high-k susceptibility has a logarithmic singularity.

Shown in Table I are analytic representations of the susceptibility for $D=0$–6 in the high- and low-k regions obtained by the relationships of the contiguous hgf. Even- and odd-dimensional cases are grouped separately to emphasize their distinctive singular behavior. These results for $D=1$–4 are in agreement with the previously established results [Eqs. (6) and (10)]. The agreement for $D=1$ and 2 is interesting in view of the original restriction imposed on Eq. (11), i.e., $D\geqslant 3$. Evidently, the dimensional regularization techniques used here has removed ultimately even this restriction. Illustrated in Fig. 1 is the susceptibility versus wave vector for a few low-integer values of D.

VI. BEHAVIOR NEAR $2k_F$

The hgf $F(a,b;c;t)$ is *absolutely* convergent on $|t|=1$ if $\mathrm{Re}(c-a-b)>0$ and has the value[21]

$$F(a,b;c;1)=\Gamma(c)\Gamma(c-a-b)/\Gamma(c-a)\Gamma(c-b)\,. \tag{25}$$

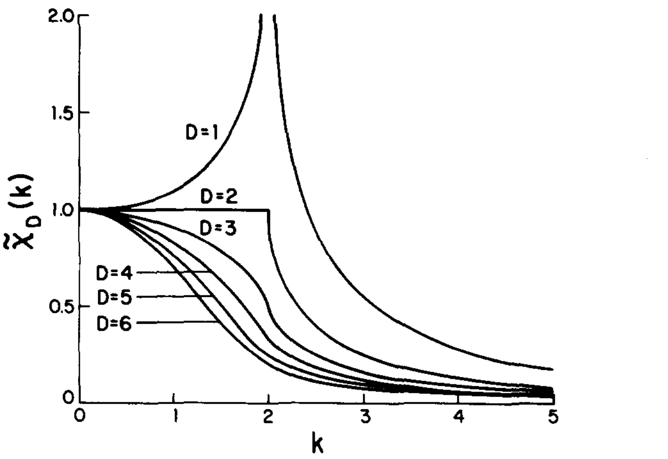

FIG. 1. The susceptibility versus wave vector at integer values of D. Here k is in units of k_F.

If applied to the high- and low-k sides, we find that

$$\tilde{\chi}_1(z=1,D)=\tilde{\chi}_2(z=1,D)=(D-1)^{-1}\,,\quad D>1\,. \tag{26}$$

Thus, the susceptibility is continuous at $z=1$ $(k=2k_F)$ except when $D=1$.

The slope at the boundary can be evaluated by using

$$\frac{\partial}{\partial t}F(a,b;c;t)=\frac{ab}{c}F(a+1,b+1;c+1;t) \tag{27}$$

and (25) provided now that $\mathrm{Re}(c-a-b-1)>0$. We obtain

$$\begin{aligned}\frac{\partial}{\partial z}\tilde{\chi}_1(z=1,D)&=\frac{\partial}{\partial z}\tilde{\chi}_2(z=1,D)\\&=-(D-2)/(D-1)(D-3)\,,\quad D>3\,.\end{aligned} \tag{28}$$

Similarly, we obtain

$$\begin{aligned}&\left(\frac{\partial}{\partial z}\right)^2\tilde{\chi}(z=1,D)\\&\quad=2(D-2)(D-4)/(D-1)(D-3)(D-5)\,,\\&\quad D>5\,.\end{aligned} \tag{29}$$

TABLE I. Analytic expressions of the susceptibility. These results are obtained by the relationships of the contiguous hgf. $L_1=\ln(1+z^{1/2})/(1-z^{1/2})$ and $L_2=\ln(1+z^{-1/2})/(1-z^{-1/2})$.

D	$\tilde{\chi}_1$	$\tilde{\chi}_2$
0	$z^{-1/2}(1-z)^{-1/2}\sin^{-1}z^{1/2}$	$D^{-1}z^{-1}(1-z^{-1})^{-1/2}$, $D\to 0$
2	1	$1-(1-z^{-1})^{1/2}$
4	$1-2z/3$	$1-2z/3(1-(1-z^{-1})^{3/2})$
6	$1-4z/3+8z^2/15$	$1-4z/3+8z^2/15(1-(1-z^{-1})^{5/2})$
1	$\frac{1}{2}z^{-1/2}L_1$	$\frac{1}{2}z^{-1/2}L_2$
3	$\frac{1}{2}+\frac{1}{4}z^{-1/2}(1-z)L_1$	$\frac{1}{2}+\frac{1}{4}z^{-1/2}(1-z)L_2$
5	$\frac{5}{8}-3z/8+3/16z^{-1/2}(1-z)^2L_1$	$\frac{5}{8}-3z/8+3/16z^{-1/2}(1-z)^2L_2$

TABLE II. Boundary values of the susceptibility. Here $\tilde{\chi}'$ and $\tilde{\chi}''$ are, respectively, the first and second derivatives of $\tilde{\chi}(z,D)$ with respect to z and evaluated at $z = 1$. These undesignated ∞'s are divergent as $(1 - z^{-1})^{-1}$ for D even and $(1 - z^{-1/2})^{-1}$ for D odd. In the unfilled regions, the appropriate formulas are given.

D	$\tilde{\chi}_1$	$\tilde{\chi}_2$	$\tilde{\chi}_1'$	$\tilde{\chi}_2'$	$\tilde{\chi}_1''$	$\tilde{\chi}_2''$
1	∞ (log)	∞ (log)	∞	$-\infty$	∞	∞
2	1	1	0	$-\infty$	0	∞
3	$\frac{1}{2}$	$\frac{1}{2}$	$-\infty$ (log)	$-\infty$ (log)	∞	∞
4	$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{3}{2}$	$-\frac{3}{2}$	0	∞
5			$-\frac{3}{8}$	$-\frac{3}{8}$	∞ (log)	∞ (log)
6					$\frac{16}{15}$	$\frac{16}{15}$
.	$\frac{1}{(D-1)}$		$\frac{-(D-2)}{(D-1)(D-3)}$			
.					$\frac{2(D-2)(D-4)}{(D-1)(D-3)(D-5)}$	
.						
∞	0	0	0	0	0	0

Thus, where convergent, we see that the high- and low-k sides of the boundary have the same first and second derivatives. In Table II, we have given the boundary values.

We can also examine the behavior of the susceptibility along the boundary itself, that is, the $z = 1$ constant line in, say, the Dz plane. From (26) we see that the behavior is simpler, e.g., $(\partial/\partial D)\tilde{\chi}(1,D) = -(D-1)^{-2}$, etc., than the behavior in the direction perpendicular to the boundary.

We shall use other properties of the hgf to establish additional properties of the susceptibility at the boundary. First of all, the hgf $F(a,b;c;t)$ has two branch points, one at $t = 1$ and the other at infinity if a or b is not a negative integer. Hence, except when D is an even integer on the low-k side, the boundary is a line of branch points.

Also the hgf $F(a,b;c;t)$ is defined by a power series [see Eq. (18)] for t complex when $|t| < 1$. It is certainly regular in this domain. Hence, the susceptibility is defined even for noninteger values of D. The hgf is also defined by analytic continuation when $|t| > 1$. It suggests, therefore, that $\tilde{\chi}_2(z,D)$ may be an analytic continuation of $\tilde{\chi}_1(z,D)$ into the high-k region.

It is known that when t lies in the part of the cut plane for which $|t| > 1$, $|\arg(-t) < \pi$ (see Ref. 21),

$$F(a,b;c;t) = B(a,b,c)(-t)^{-a}F(a,1-c+a;1-b+a;t^{-1}) + B(b,a,c)(-t)^{-b} \times F(b,1-c+b;1-a+b;t^{-1}), \quad (30)$$

where

$$B(a,b,c) = \Gamma(c)\Gamma(b-a)/\Gamma(b)\Gamma(c-a). \quad (31)$$

Hence, $F(a,b;c;t)$, when it has a meaning, is a one-valued analytic function, regular in the whole plane of t, cut along

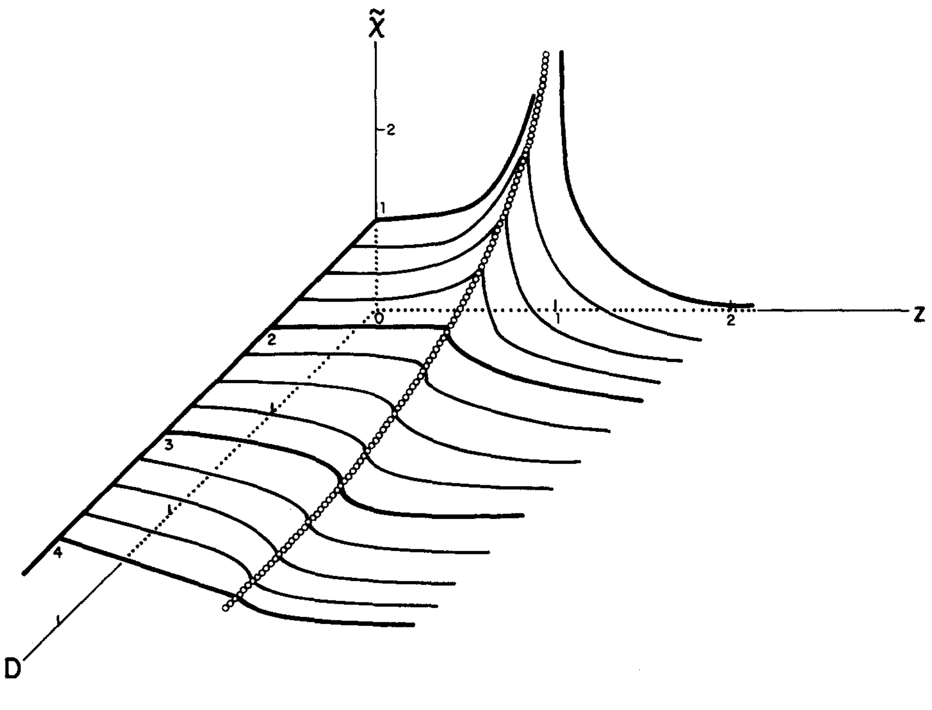

FIG. 2. The susceptibility as a function of z and D. Small circles form a line of branch points, and $z = (k/2)^2$.

the real axis from $t = 1$ to ∞.

Let the domains $|z| < 1$ and $|z| \geqslant 1$ be denoted, respectively, by $\mathscr{D}_1$ and $\mathscr{D}_2$. By our definition $\tilde{\chi}_1$, which is analytic in $\mathscr{D}_1$, is $\tilde{\chi}$ in $\mathscr{D}_1$ and similarly $\tilde{\chi}_2$ is the analytic function $\tilde{\chi}$ in $\mathscr{D}_2$. Then by the above-stated properties of the hgf, $\mathscr{F} \equiv F(1,1-\frac{1}{2}D;\frac{3}{2};z)$ is analytic in $\mathscr{D}_1 \cup \mathscr{D}_2$. If for $\operatorname{Re} z$, $\operatorname{Re} \mathscr{F} = \tilde{\chi}_1$ in $\mathscr{D}_1$ and $\operatorname{Re} \mathscr{F} = \tilde{\chi}_2$ in $\mathscr{D}_2$, then $\tilde{\chi}_2$ is the analytic continuation of $\tilde{\chi}_1$ into $\mathscr{D}_2$ (see Ref. 22). By (31),

$$\mathscr{F} = D^{-1} z^{-1} F(1,\tfrac{1}{2};1+\tfrac{1}{2}D;z^{-1}) + \Gamma(\tfrac{3}{2})\Gamma(\tfrac{1}{2}+\tfrac{1}{2}D) \times(-z)^{-1+\frac{1}{2}D}(1-z^{-1})^{-\frac{1}{2}+\frac{1}{2}D}. \tag{32}$$

Consider $\mathscr{F}$ when D is an odd integer first. It is sufficient to take $D = 1$ since $\mathscr{F}$ of other odd integers can be generated from it. Then, for $D = 1$,

$$\mathscr{F} = \tfrac{1}{2} z^{-1/2} \ln(1+z^{-1/2})/(1-z^{-1/2}) + \tfrac{1}{2}(-z)^{-1/2}. \tag{33}$$

In the domain $\mathscr{D}_1$ the above logarithmic argument is negative. It can be resolved into real and imaginary parts, the latter of which cancels the second term of (33) exactly, leaving

$$\mathscr{F} = \tfrac{1}{2} z^{-1/2} \ln(1+z^{1/2})/(1-z^{1/2}). \tag{34}$$

Hence, $\operatorname{Re} \mathscr{F} = \tilde{\chi}_1$ in $\mathscr{D}_1$. It follows directly tht $\operatorname{Re} \mathscr{F} = \tilde{\chi}_2$ in $\mathscr{D}_2$. One can similarly show that whenever D is an odd integer, $\tilde{\chi}_2$ is the analytic continuation of $\tilde{\chi}_1$ into $\mathscr{D}_2$.

We next consider $\mathscr{F}$ when D is an even integer. Then, $b = 1 - \frac{1}{2}D$ is either zero or a negative integer and $\tilde{\chi}_1 = F(1,b;\frac{3}{2};z)$ is an entire function being a polynomial. But $\tilde{\chi}_2$ is not a polynomial. The two functions are thus not related although continuous at $z = 1$. One can, in fact, show by (32) that, for $\operatorname{Re} z$, *Re* $\mathscr{F} = \tilde{\chi}_1$ in $\mathscr{D}_1$, but $\operatorname{Re} \mathscr{F} \neq \tilde{\chi}_2$ in $\mathscr{D}_2$. With this analysis we conclude that when D is an odd integer, $\tilde{\chi}_2$ is the analytic continuation of $\tilde{\chi}_1$; but when D is an even integer, it is not. When D is not an integer, the relationship established for D odd integers is expected to hold.

VII. DISCUSSION

Our results for the susceptibility obtained as a function of z and D are embodied in Fig. 2, which gives a three-dimensional projection of $\tilde{\chi} zD$. It is reminiscent of the PVT diagram of a homogeneous fluid. The surface represents the susceptibility that is physically accessible as z and D are varied. The shape of the surface is distinguished by an unbroken "ridge" (marked in the figure by small circles). It is a line of branch points, a $z = 1$ constant line. The ridge separates the surface into two sides (high k and low k). The low-k side of the surface is further subdivided by the $D = 2$ constant line into an area of rising curvature and an area of falling curvature. The high-k side of the surface is not divided further. Hence, the ridge is folded upward for $1 < D < 2$ and folded downward for $2 < D < \infty$.

The ridge itself shows very smooth behavior, becoming singular at one end ($D = 1$) and vanishing at the other end ($D = \infty$). Other z-constant lines, e.g., $z = 0$, are less interesting. More interesting are D-constant lines that intersect the ridge. They look much like the familiar isotherms in the PV diagram (also compare Fig. 1). When D is an integer, the intersection is a point of singularity, either open as in $D = 1$ or hidden as in $D = 2$. The ridge is punctuated with these intersections throughout.

Other finer details of the susceptibility surface are possible to give. Except on the D-constant lines of *even* integers, one can move across the ridge via analytic continuation. These excepted lines are demarcated by the ridge. That is, on these excepted lines, the knowledge of one side is insufficient to describe the other side.[23] The singularity at $z = D = 1$ is weaker when approached perpendicular to the ridge than when approached along the ridge. To some extent our picture is applicable to an interacting electron gas by virtue of the RPA theories.[24]

ACKNOWLEDGMENTS

We wish to thank Professor John E. Rives for useful suggestions, especially for constructing our Fig. 2.

This work was supported in part by the Department of Energy, Office of Basic Energy Science, and the Office of Naval Research.

[1] See, e.g., S. W. Lovesey, *Condensed Matter Physics* (Benjamin, Reading, MA, 1980). The susceptibility is also known as the polarizability, compressibility, density–density response function, etc., under different contexts.

[2] J. Lindhard, Kgl. Danske, Vidensk. Selskab, Mat. Fys. Medd. **28**(8), 1 (1954).

[3] F. Stern, Phys. Rev. Lett. **18**, 546 (1967). Also see M. H. Lee and J. Hong, Phys. Rev. Lett. **48**, 634 (1982); Phys. Rev. B **32**, 7734 (1985).

[4] P. F. Maldague, Surf. Sci. **73**, 296 (1978).

[5] M. Apostol, A. Corciovei, and S. Stoica, Phys. Status Solidi B **103**, 411 (1981).

[6] W. L. Friesen and B. Bergersen, J. Phys. C **13**, 6627 (1980).

[7] J. M. Ziman, *Principles of the Theory of Solids* (Cambridge U. P., London, 1972), 2nd ed., p. 155.

[8] See, e.g., G. D. Mahan, *Many Particle Physics* (Plenum, New York, 1981). Also see J. Hong and M. H. Lee, Phys. Rev. Lett. **55**, 2375 (1985).

[9] R. Kubo, Rep. Prog. Phys. **29**, 255 (1966).

[10] M. H. Lee, Phys. Rev. Lett. **49**, 1072 (1982).

[11] M. H. Lee, Phys. Rev. B **26**, 2547 (1982).

[12] N. L. Sharma, S. K. Oh, and M. H. Lee (to be published).

[13] M. H. Lee, I. M. Kim, and R. Dekeyser, Phys. Rev. Lett. **52**, 1579 (1984).

[14] B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, San Francisco, 1982).

[15] K. G. Wilson, Rev. Mod. Phys. **47**, 773 (1975).

[16] See, e.g., *Kinetics of Aggregation and Gelation*, edited by F. Family and D. P. Landau (North-Holland, Amsterdam, 1984).

[17] S. Narison, Phys. Rep. **84**, 263 (1982).

[18] See Eq. 1.79e of Ref. 17.

[19] The upper limits on the sums (14) and (15) at first appear troublesome if one takes $s \to 0$. A moment of reflection assures us that clearly there must be an infinite number of terms in that limit.

[20] By definition the homogeneous static susceptibility $\chi(k=0) = \partial\rho/\partial\epsilon_F$, where $\rho = 2(k_F/2\pi)^D S_D/D$, the density of electrons in D dimensions.

[21] See, e.g., G. Sansone and J. Gerretsen, *Lectures on the Theory of Functions of a Complex Variable* (Wolters–Noordhoff, Gronigen, 1969), Chap. 16.

[22] Since $\chi(z,D)$ is real for $\operatorname{Re} z$, we shall add this reality requirement.

[23] A similar situation occurs in the Yang–Lee theory of phase transitions in which a line of zeros of the grand partition function separates high- and low-temperature regions. Although continuous, pressure in one region in general cannot be determined from it in another region. See, e.g., G. E. Uhlenbeck and G. Ford, *Lectures in Statistical Mechanics* (Am. Math. Soc., Providence, RI, 1963), pp. 66–71.

[24] In this regard, also see Y. R. Wang, M. Ashraf, and A. W. Overhauser, Phys. Rev. B **30**, 5580 (1984).

On the zeros of the dispersion function in particle transport theory

R. L. Bowden

Department of Physics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

(Received 9 July 1985; accepted for publication 12 February 1986)

The zeros of the dispersion function that arise in particle transport with anisotropic scattering are studied. An algebraic test for the number of zeros is presented.

I. INTRODUCTION

In treating particle transport in plane geometry with azimuthal symmetry, the transport equation of the particle density $\Psi(x,\mu)$ is often written in the form[1]

$$\mu\frac{\partial\Psi}{\partial x}+\Psi(x,\mu)=\frac{c}{2}\int_{-1}^{+1}f(\mu,\mu')\Psi(x,\mu')d\mu',\tag{1.1}$$

where c is the mean number of secondary particles per collision, x is the distance measured in mean free paths, and μ is the direction cosine of the angle between the x axis and the particle velocity. Here it is assumed that the scattering law is such that $f(\mu,\mu')$ can be adequately represented by a finite Legendre expansion, viz.,

$$f(\mu,\mu')=\sum_{n=0}^{N}(2n+1)f_nP_n(\mu)P_n(\mu'),\tag{1.2}$$

where $P_n(\mu)$ is the Legendre polynomial of order n and physical considerations require that $f_0=1$ and $|f_n|\leqslant 1$, $n>1$. For definitiveness it will be assumed that $f_N\neq 0$. The purpose of this paper is to reexamine the zeros of the dispersion function that arises in the solution to Eq. (1.1). In particular, Mika[2] showed over two decades ago that solutions of the form $\varphi_\nu(\mu)\exp(-x/\nu)$ yield the eigenvalue equation

$$(\nu-\mu)\varphi_\nu(\mu)=\frac{c}{2}\sum_{n=0}^{N}\nu P_n(\mu)h_{n,c}(\nu),\tag{1.3}$$

where

$$h_{n,c}(\nu)=\int_{-1}^{+1}P_n(\mu)\varphi_\nu(\mu)d\mu.\tag{1.4}$$

Further, Mika showed, using the orthogonality and recursion properties of Legendre polynomials, that $h_{n,c}(\mu)$ is a polynomial uniquely determined by the recursion formula

$$(n+1)h_{n+1,c}(\nu)+nh_{n-1,c}(\nu)$$
$$=(2n+1)(1-cf_n)\nu h_{n,c}(\nu),\tag{1.5}$$

and the nonrestrictive requirement that $h_{-1,c}(\nu)=0$ and

$$h_{0,c}(\nu)=1.\tag{1.6}$$

The so-called discrete solutions of Eq. (1.1) are obtained by solving Eq. (1.3) for $\varphi_\nu(\mu)$ and using the normalization given by Eq. (1.6). The result is that discrete solutions occur for those values of ν in the complex plane $\mathbb{C}\setminus[-1,+1]$ that are zeros of the dispersion function

$$\Lambda_c(\nu)=1+\frac{c}{2}\int_{-1}^{+1}\frac{\nu g(\mu,\nu)}{\mu-\nu}d\mu,\tag{1.7}$$

where

$$g(\mu,\nu)=\sum_{n=0}^{N}(2n+1)f_nP_n(\mu)h_{n,c}(\nu).\tag{1.8}$$

The dispersion function obviously has a cut in the complex ν plane along $(-1,+1)$. The limit values $\Lambda_c^+(\mu)$ and $\Lambda_c^-(\mu)$ of $\Lambda_c(\nu)$ as ν approaches a value $\mu\in(-1,+1)$ from the upper and lower half complex planes, respectively, are given by

$$\Lambda_c^\pm(\mu)=1+\frac{c}{2}P\int_{-1}^{+1}\frac{\nu g(\eta,\mu)}{\eta-\mu}d\eta\pm\frac{i\pi c\mu\gamma_c(\mu)}{2},\tag{1.9}$$

where P indicates the Cauchy principal value and

$$\gamma_c(\mu)=g(\mu,\mu).\tag{1.10}$$

Case[3] and Hangelbrook[4] have shown that $\Lambda_c^\pm(\mu)$ does not vanish for $-1<\mu<+1$ and Lekkerkerker[5] has shown that the same result is true for the end points ± 1. The limit value of $\Lambda_c(\nu)$ as $\nu\to\infty$ is given by[6]

$$\Lambda_c(\infty)=\prod_{n=0}^{N}(1-cf_n).\tag{1.11}$$

Other statements about the location and character of the zeros can be made. It is readily seen that the roots must occur in ± pairs. Further, Case[3] showed that if $c<1$ that the zeros of $\Lambda_c(\nu)$ are real. Moreover, Case showed that if $1-cf_n\geqslant 0$ for $n=1,3,5,\ldots$, the zeros are all simple and are either real or purely imaginary. However, the determination of the *number* of zeros of the dispersion function remains relatively primitive. The number of zeros $2M$ of $\Lambda_c(\nu)$ can be obtained from the argument principle. The contour C in Fig. 1 and a contour at infinity encloses the cut plane. Because $\Lambda_c(\infty)$ is a constant, the number of zeros of the dispersion function is given by the change in the argument of $\Lambda_c(\nu)$ along C as the contour is collapsed (with $\rho\to 0$) onto the real interval $(-1,+1)$. This procedure yields

$$M=(1/\pi)\Delta_C\operatorname{Arg}\Lambda_c^+(\mu),\tag{1.12}$$

where $\Delta_C\operatorname{Arg}\Lambda_c^+(\mu)$ represents the change in the argument of $\Lambda_c^+(\mu)$ as μ varies along the directed line from -1 to $+1$. Since the imaginary part of $\Lambda_c^+(\mu)$, $\mu\in(-1,+1)$, is a polynomial of at most degree $N+1$, then $M\leqslant N+1$. For linear anisotropic scattering ($N=1$), the number of pairs of zeros of $\Lambda_c(\nu)$ can be shown to be either one or two depending on the values of c and f_1. The proof of this last statement is essentially an algebraic one. As will be seen below, the enumeration of the pairs of zeros of

0022-2488/86/061624-09$02.50

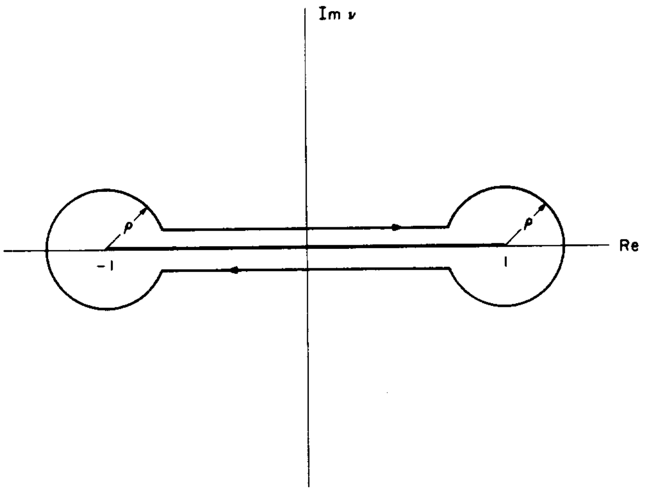

FIG. 1. Contour C.

$\Lambda_c(\nu)$ becomes more difficult as the order of the scattering increases. For $N > 4$ and for a given value of c and a set of $\{f_n\}$, the enumeration of the pairs of zeros of the dispersion function in some kind of "closed form" is an unlikely possibility and resort to some sort of numerics is inevitable. The big problem with a numerical evaluation of the change in the argument of a function is that it is easy to lose track as the argument unfolds. Thus an independent evaluation of the number of pairs of zeros would be useful.

The main result of this paper provides such an algebraic test for the number of pairs of zeros of the dispersion function. The proof of this test is based in part on the observation that the function $\gamma_c(1)$ can be regarded as a polynomial in c of order $N^* = N - K$, where K is the number of f_n, $0 < n < N$, that are zero. It will be shown below that the N^* zeros of $\gamma_c(1)$ are all simple and real. Denote the nonpositive zeros of $\gamma_c(1)$ by c_p^-, $p = 1,...,P$, and the positive zeros by c_q^+, $q = 1,...,Q$, with $P + Q = N^*$. Order these zeros according to

$$c_P^- < c_{P-1}^- < \cdots < c_1^- < c_1^+ < c_2^+ < \cdots < c_Q^+. \qquad (1.13)$$

If $0 < c_{k-1}^+ < c < c_{k+1}^+$ for a given set of $\{f_n\}$, then the number of pairs of zeros of $\Lambda_c(\nu)$ is $k+1$. A similar idea was proposed by Dawn and Chen[7] but their analysis is not as complete as the one presented here.

The proof of the preceding test is contained in the remaining sections of this paper. It proves convenient in that proof to make the change of variables $c \rightarrow 1/s$. This change is made in Sec. II. The essential points of a mapping between the s plane and the ν plane are also made in that section. The proof of the test given above is contained in the main theorem proved in Sec. III. Concluding ancillary remarks about the character of the zeros of the dispersion function are made in Sec. IV.

II. MAPPING BETWEEN THE ν PLANE AND THE *s* PLANE

The dispersion function can also be written in the form

$$\Lambda_c(\nu) = R_c(\nu) - c\nu\gamma_c(\nu)Q_0(\nu), \qquad (2.1)$$

where here and in the subsequent analysis $Q_n(\nu)$ is the nth-order Legendre function of the second kind and $R_c(\nu)$ is a polynomial in ν and c. With the change of variables $c = 1/s$ an auxiliary dispersion function $\Lambda(\nu,s)$ is defined by

$$\begin{aligned}\Lambda(\nu,s) &= s^{N^*+1}\Lambda_{1/s}(\nu)\\ &= R(\nu,s) - \nu\gamma(\nu,s)Q_0(\nu)\\ &= s^{N^*+1} + s^{N^*}A_1(\nu) + \cdots + A_{N^*}(\nu), \qquad (2.2)\end{aligned}$$

where

$$\begin{aligned}R(\nu,s) &= s^{N^*+1}R_{1/s}(\nu)\\ &= s^{N^*+1} + s^{N^*}b_1(\nu) + \cdots + b_{N^*+1}(\nu), \qquad (2.3)\end{aligned}$$

and

$$\begin{aligned}\gamma(\nu,s) &= s^{N^*}\gamma_{1/s}(\nu)\\ &= s^{N^*}a_0(\nu) + s^{N^*-1}a_1(\nu) + \cdots + a_{N^*}(\nu). \qquad (2.4)\end{aligned}$$

Here $b_j(\nu)$ and $a_j(\nu)$ are even polynomials in ν only and $A_j(\nu)$ is an analytic function on $\nu\,\mathbb{C}\setminus[-1,+1]$. Obviously, $\Lambda(\nu,s)$ and $\Lambda_{1/s}(\nu)$ have the same zeros in the ν plane for $s \neq 0$ ($c \neq \infty$). The object is to consider $\Lambda(\nu,s)$ as a complex function of two complex variables and use the implicit function theorem to study $\Lambda(\nu,s) = 0$.

In particular, $\Lambda(\nu,s)$ for fixed ν can be regarded as a polynomial in s and its zeros can be investigated. For example, with $\nu = \infty$, Eq. (1.11) can be written in the present notation as

$$\Lambda(\infty,s) = \prod_{n=0}^{N} G_n(s), \qquad (2.5)$$

where

$$\begin{aligned}G_n(s) &= (s - f_n), \quad \text{if} f_n \neq 0,\\ &= 1, \qquad\quad \text{if} f_n = 0.\end{aligned} \qquad (2.6)$$

Thus the point $\nu = \infty$ maps by $\Lambda(\infty,s) = 0$ into $N^* + 1$ real points, the nonzero f_n in the s plane. These points are, of course, distinct if the f_n are all different. Consequently, it will be assumed for simplicity that all the nonzero f_n are distinct. However, since $\Lambda(\nu,s)$ and $\gamma(\nu,s)$ are also polynomials in the f_n, the main results obtained here also follow for nondistinct f_n by continuity. Other points in the ν plane also map into real points in the s plane. To this point consider the following.

Lemma 1: If $\nu_0 \in \mathbb{R}\setminus[-1,+1]$, then the roots of $\Lambda(\nu_0,s) = 0$ are all real.

Proof: The proof of this lemma follows from using the dispersion function in a form written by Inönü,[8]

$$\Lambda_c(\nu) = (N+1)[Q_{N+1}h_{N,c}(\nu) - Q_N(\nu)h_{N+1,c}(\nu)]. \qquad (2.7)$$

If $\kappa_n < n$ of the f_n are zero and

$$h_n(\nu,s) = s^{\kappa_n}h_{n,1/s}(\nu), \qquad (2.8)$$

the recursion formula for the $h_n(\nu,s)$ can be written as

$$\begin{aligned}(n+1)h_{n+1}(\nu,s) &+ ns^{\delta_n}h_{n-1}(\nu,s)\\ &= (2n+1)G_n(s)\nu h_n(\nu,s), \qquad (2.9)\end{aligned}$$

with

$$h_{-1}(\nu,s) = 0, \; h_0(\nu,s) = 1,$$

and

$$\delta_n = \kappa_{n+1} - \kappa_{n-1} \geqslant 0.$$

Thus the auxiliary dispersion function $\Lambda(\nu,s)$ takes the form

$$\Lambda(\nu,s) = (N+1)[Q_{N+1}sh_n(\nu,s) - Q_N(\nu)h_{N+1}(\nu,s)]. \tag{2.10}$$

Let $\nu_0 \in \mathbf{R} \setminus [-1,+1]$ be fixed and consider

$$\Lambda(\nu_0,s) = Q_{N+1}(\nu_0)sh_N(\nu_0,s) - Q_N(\nu_0)h_{N+1}(\nu_0,s) = 0. \tag{2.11}$$

Note that $h_N(\nu_0,s)$ and $h_{N+1}(\nu_0,s)$ cannot vanish for the same value of s, for if they did, then the recursion formula would yield $h_{N-1}(\nu_0,s) = 0$, which would imply $h_{N-2}(\nu_0,s) = 0$, etc. This would eventually lead to the contradiction $h_0(\nu_0,s) = 0$. It can be easily shown that

$$h_n(\nu,0) = \frac{1}{n!}\prod_{j=0}^{n-1}(2n+1)[-G_n(0)]\nu^j. \tag{2.12}$$

Thus $\Lambda(\nu_0,0)$ does not vanish for $\nu_0 \in \mathbf{R} \setminus [-1,+1]$. Now let s_1 and s_2 be nonzero roots of $\Lambda(\nu_0,s) = 0$. Equation (2.11) then yields

$$s_2h_N(\nu_0,s_2)h_{N+1}(\nu_0,s_1) = s_1h_N(\nu_0,s_1)h_{N+1}(\nu_0,s_2). \tag{2.13}$$

Rewriting Eq. (2.9) for $s = s_1$, $\nu = \nu_0$, and then for $s = s_2$, $\nu = \nu_0$, and combining the results in a familiar fashion yields

$$\begin{aligned}(N+1)[s_2h_N(\nu_0,s_2)&h_{N+1}(\nu_0,s_1) - s_1h_N(\nu_0,s_1)h_{N+1}(\nu_0,s_2)]/(s_1s_2)^{N^*+1}\\ &= (s_1 - s_2)\sum_{n=0}^{N}\frac{\nu_0(2n+1)f_nh_n(\nu_0,s_1)h_n(\nu_0,s_2)}{(s_1s_2)^{\kappa_n-1}}.\end{aligned} \tag{2.14}$$

Because $h_n(\nu_0,s)$ for fixed $\nu_0 \in \mathbf{R} \setminus [-1,+1]$ is a polynomial in s with real coefficients, if s_1 is a zero of $\Lambda(\nu_0,s)$, then so is $\bar{s}_1$. Thus let $s_2 = \bar{s}_1$ and employ Eq. (2.13) to obtain

$$\operatorname{Im} s_1 \sum_{n=0}^{N}(2n+1)f_n\left|\frac{h_n(\nu_0,s_1)}{s_1^{\kappa_n}}\right|^2 = 0. \tag{2.15}$$

Hence, for example, if all of the f_n are non-negative, then the sum in the last expression is positive and therefore s_1 real.

To pin down the general situation consider the relation given by Bowden *et al.*,[9]

$$\begin{aligned}\Lambda_{1/s}(\nu)P_n(\nu) &= \frac{\nu}{2}\int_{-1}^{+1}\frac{P_n(\mu)}{\nu-\mu}\times\sum_{m=0}^{N}\frac{(2m+1)f_mP_m(\mu)h_m(\nu,s)}{s^{\kappa_m+1}}\,d\mu\\ &\quad + h_n(\nu,s)/s^{\kappa_n}.\end{aligned} \tag{2.16}$$

Now let $\nu = \nu_0$ and $s = s_1$ be defined as above. Multiplying Eq. (2.16) by $(2n+1)f_nh_n(\nu_0,\bar{s}_1)/\bar{s}^{\kappa_n}$ and summing on n yields

$$\frac{\nu_0}{2s_1}\int_{-1}^{+1}\left|\sum_{n=0}^{N}\frac{(2n+1)f_nP_n(\mu)h_n(\nu_0,s_1)}{s_1^{\kappa_n}}\right|^2\frac{d\mu}{\mu-\nu_0} + \sum_{n=0}^{N}(2n+1)f_n\left|\frac{h_n(\nu_0,s_1)}{s_1^{\kappa_n}}\right|^2 = 0. \tag{2.17}$$

Here the fact that $h_n(\nu_0,\bar{s}_1) = \bar{h}_n(\nu_0,s_1)$ for ν_0 real has been used. For $\nu_0 \in \mathbf{R} \setminus [-1,+1]$ the integral term in Eq. (2.17) will not vanish; thus Eqs. (2.15) and (2.17) state that $\operatorname{Im} s_1 = 0$, i.e., s_1 is real. This completes the proof of the lemma. To show that these zeros (for fixed ν_0) are simple, consider the following.

Lemma 2: If $\nu_0 \in \mathbf{R} \setminus [-1,+1]$ and $\Lambda(\nu_0,s_0) = 0$, then $\partial\Lambda(\nu_0,s_0)/\partial s \neq 0$.

Proof: Let $H_n(\nu,s) = sh_n(\nu,s)$. It follows from Eq. (2.10) that

$$\frac{\partial\Lambda(\nu,s)}{\partial s} = (N+1)\left[Q_{N+1}(\nu)\frac{\partial H_n(\nu,s)}{\partial s} - Q_N(\nu)\frac{h_{N+1}(\nu,s)}{\partial s}\right]. \tag{2.18}$$

If now both $\Lambda(\nu_0,s_0) = 0$ and $\partial\Lambda(\nu_0,s_0)/\partial s = 0$, then Eqs. (2.11) and (2.18) imply that

$$h_{N+1}(\nu_0,s_0)\frac{\partial H_N(\nu_0,s_0)}{\partial s} = H_N(\nu_0,s_0)\frac{\partial h_{N+1}(\nu_0,s_0)}{\partial s}. \tag{2.19}$$

Dividing both sides of Eq. (2.14) by $(s_1 - s_2)$ and taking the limit $s_2 \to s_1 = s_0$, where s_0 is a zero of $\Lambda(\nu_0,s)$, give

$$\frac{(N+1)}{s_0^{2N^*}}\left[H_N(\nu_0,s_0)\frac{\partial h_{N+1}(\nu_0,s_0)}{\partial s} - h_N(\nu_0,s_0)\frac{\partial H_N(\nu_0,s_0)}{\partial s}\right] = \sum_{n=0}^{N}(2n+1)f_n\left|\frac{h_n(\nu_0,s_0)}{s_0^{\kappa_n}}\right|^2. \tag{2.20}$$

Therefore from Eq. (2.19) a necessary condition for $\Lambda(\nu_0,s_0)$ and $\partial\Lambda(\nu_0,s_0)/\partial s$ to vanish is that the right-hand side of Eq. (2.20) also vanish. The proof of the lemma is completed by recalling from Lemma 1 that the right-hand side of Eq. (2.20) does not vanish for $\nu_0 \in \mathbf{R} \setminus [-1,+1]$.

There are $N^* + 1 = N + 1 - K$ nonvanishing roots of $\Lambda(\nu_0,s) = 0$ for $\nu_0 \in \mathbf{R} \setminus [-1,+1]$ that are real and simple. Denote these roots by $s_0^{(0)},s_0^{(2)},\ldots,s_0^{(N^*)}$. From the implicit function theorem there are neighborhoods, say $N(\nu_0)$ and $N_j(s_0^{(j)})$, such that the equation $\Lambda(\nu,s) = 0$ has a unique root $S_j(\nu)$ in $N_j(s_0^{(j)})$ for any ν in $N(\nu_0)$. Further, each function $S_j(\nu)$ is single valued and analytic on $N(\nu_0)$ and satisfies the condition $S_j(\nu_0) = s_0^{(j)}$.

The immediate objective now is to continue the $S_j(\nu)$ to the right (left) half complex plane cut as described below. That each of these functions can be continued along any line in the ν plane that avoids the cut $[-1,+1]$ and zeros of the discriminant of Eq. (2.2) is clear. The discriminant of Eq. (2.2) can be written in the form

$$D(\nu) = \sum_{n=0}^{M_0}\beta_n(\nu)[\nu Q_0(\nu)]^n, \tag{2.21}$$

where M_0 is finite and $\beta_n(\nu)$ is an even polynomial with real coefficients. Thus $D(\nu)$ is analytic on the complex plane cut along $[-1,+1]$, has at most a finite-order pole at infinity, and has the limits

$$D^{\pm}(\mu)=\sum_{n=0}^{M_0}\beta_n(\mu)\left[\mu\tanh^{-1}\mu\mp\frac{i\pi\mu}{2}\right]^n \quad (2.22)$$

on the cut $(-1,+1)$. It is readily seen that the real and imaginary parts of $D^{\pm}(\mu)$ have only a finite number of zeros for $\mu\in(-1,+1)$. A straightforward argument principle calculation similar to the one about the contour C mentioned in Sec. I shows that the number of zeros of $D(\nu)$ is finite. Because of the assumption that nonzero f_n are distinct, $D(\nu)$ does not vanish at infinity. Further, since $D(\nu)=D(-\nu)$ and $D(\bar{\nu})=\bar{D}(\nu)$, if $\nu=\nu'$ is a zero of $D(\nu)$ so are $\nu=-\nu'$ and $\nu=\bar{\nu}$. Let

$$\mathscr{D}=\{\zeta_i|D(\zeta_i)=0\}, \quad (2.23)$$

where $\pm\zeta_0, \pm\zeta_1,\dots, \zeta_p=0$ are points on the imaginary axis with $|\zeta_0|>|\zeta_1|>\cdots>|\zeta_p|$ and $\pm\zeta_{p+1},\dots, \pm\zeta_{p+q}, \pm\bar{\zeta}_{p+q}$ are the rest of the points of $\mathscr{D}$. (Note that p could be equal to zero.) Now cut the ν plane by joining $+\zeta_0, +\zeta_1,\dots,\zeta_p$ in the upper half plane with a straight line, similarly joining $-\zeta_0, -\zeta_1,\dots,\zeta_p$ in the lower half plane, joining $\zeta_p=0$, $\zeta_{p+1},\dots,\zeta_{p+q}$ with a series of straight lines in the first quadrant, making similar joinings in the remaining quadrants, and finally adding the original cut along $(-1,+1)$.

Each of the $S_j(\nu)$ can be analytically continued to the right (left) half complex plane cut as described above so that, according to the monodromy theorem, each function will be single valued and analytic in the right (left) cut plane. Each function $S_j(\nu)$ can be continued from the right half plane to the left half plane by considering the regions $|\text{Im}\,\nu|>|\zeta_0|$. Thus each $S_j(\nu)$ so continued has the property that $S_j(\nu)=S_j(-\nu)$. Since $S_j(\nu)$ is real for $\nu\in\mathbb{R}\setminus[-1,+1]$, the reflection principle yields the additional property that $S_j(\bar{\nu})=\bar{S}_j(\nu)$. Since $S_j(\nu)$ is continuous across the imaginary axis for $|\text{Im}\,\nu|>|\zeta_0|$, the two properties listed show that $S_j(\nu)$ is real if ν lies on the imaginary axis and $|\text{Im}\,\nu|>|\zeta_0|$. Most importantly, of course, is the property that $\Lambda[\nu,S_j(\nu)]=0$ for every ν in the plane cut as described. The functions $S_j(\nu)$ will be labeled according to $\lim_{\nu\to\infty}S_j(\nu)=f_{n_j}$, where $f_{n_0}=f_0=1$ and f_{n_j}, $j=1,2,\dots,N^*$, are the nonvanishing expansion coefficients.

To look at the behavior of the $S_j(\nu)$ on $[-1,+]$ it is helpful to consider the following lemmas.

Lemma 3: If $\nu_0\in\mathbb{R}\setminus(-1,+1)$, then the roots of $\gamma(\nu_0,s)=0$ are all real.

Proof: As demonstrated by Inönü,[8] the recursion formula for $P_n(\nu)$ and $h_n(\nu,s)$ can be used to write

$$\nu\gamma(\nu,s)=(N+1)[P_{N+1}(\nu)sh_N(\nu,s) - P_N(\nu)h_{N+1}(\nu,s)]. \quad (2.24)$$

This expression is entirely analogous to Eq. (2.10) with $Q_n(\nu)$ replaced by $P_n(\nu)$. Thus letting $\nu_0\in\mathbb{R}\setminus(-1,+1)$ be fixed, letting s_0 be a nonzero root of $\gamma(\nu_0,s)=0$, and following the proof of Lemma 1 yields

$$\text{Im}\,s_0\sum_{n=0}^{N}(2n+1)f_n\left|\frac{h_n(\nu_0,s_0)}{s_0^{\kappa_n}}\right|^2=0. \quad (2.25)$$

Substituting $\nu=\nu_0$ and $s=s_0$ defined as above into Eq. (2.16), multiplying the resulting equation by $(2n+1)f_nh_n(\nu_0,\bar{s}_0)/\bar{s}_0^{\kappa_n}$, and summing on n gives

$$\frac{\nu_0}{2s_0}\int_{-1}^{+1}\left|\sum_{n=0}^{N}\frac{(2n+1)f_nP_n(\mu)h_n(\nu_0,s_0)}{s_0^{\kappa_n}}\right|^2\frac{d\mu}{\mu-\nu_0} + \sum_{n=0}^{N}(2n+1)f_n\left|\frac{h_n(\nu_0,s_0)}{s_0^{\kappa_n}}\right|^2=0. \quad (2.26)$$

Note that the integral in Eq. (2.26) is well defined for $\nu_0=1$ and for $\nu_0\in\mathbb{R}\setminus(-1,+1)$ the integral term does not vanish. Thus Eqs. (2.25) and (2.26) state that $\text{Im}\,s_0=0$, i.e., s_0 is real.

Lemma 4: If $\nu_0\mathbb{R}\setminus(-1,+1)$ and s_0 is a nonzero root of $\Lambda(\nu_0,s_0)=0$, then $\partial\Lambda(\nu_0,s_0)/\partial s\neq0$.

The proof of this lemma is completely analogous to that of Lemma 3 and the details will be omitted.

It can be shown that $a_0(\nu)$ in Eq. (2.4) can be written as $\Sigma_{n=0}^N(2n+1)f_n[P_n(\nu)]^2$. It will be assumed that $a_0(1)\neq0$; this is equivalent to $f(1,1)>0$ in Eq. (1.2). However, again $a_0(\nu)$ is a polynomial in the f_n and the case $a_0(1)=0$ can be included by continuity.

The equation $\gamma(\nu_0,s)=0$ for $\nu_0\in\mathbb{R}\setminus(-1,+1)$ has N^* simple real nonvanishing roots. In particular denote the roots of $\gamma(1,s)=0$ by $\xi_1,\xi_2,\dots,\xi_{N^*}$ with the ordering of the roots given by the following.

Lemma 5:

$$\lim_{\nu\to1}S_j(\nu)=\xi_j,\quad j=1,2,\dots,N^*. \quad (2.27)$$

Proof: Let

$$\Lambda'(\nu,s)=R(\nu,s)/[\nu Q_0(\nu)]-\gamma(\nu,s). \quad (2.28)$$

For fixed $\nu\neq1$ the zeros of $\Lambda(\nu,s)$ and $\Lambda'(\nu,s)$ coincide. For $\nu=1$, it is obvious that $\Lambda'(\nu,s)$ vanishes at the zeros of $\gamma(1,s)$. Thus if $S'_j(\nu)$ is a zero of $\Lambda'(\nu,s)$ then

$$\begin{aligned}|S'_j(\nu)-S_j(\nu)|&=|S'_j(\nu)-\xi_j+\xi_j-S_j(\nu)|\\&=0,\quad \nu\neq1.\end{aligned} \quad (2.29)$$

Therefore

$$|S_j(\nu)-\xi_j|=|S'_j(\nu)-\xi_j|,\quad \nu\neq1. \quad (2.30)$$

The proof is completed by noting that the right-hand side of the last equation vanishes in the limit $\nu\to1$. A similar calculation leads to the following.

Lemma 6:

$$\lim_{\nu\to1}S_0(\nu)=\lim_{\nu\to1}\xi_0(\nu), \quad (2.31)$$

where

$$\xi_0(\nu)=-b_1(\nu)-a_0(\nu)\nu Q_0(\nu) + a_1(\nu)/a_0(\nu). \quad (2.32)$$

with the polynomials $a_n(\nu)$ and $b_n(\nu)$ given by Eqs. (2.3) and (2.4).

Proof: Let

$$\Lambda''(\nu,s)=s+\sum_{n=0}^{N^*}\frac{b_{n+1}(\nu)a_n(\nu)\nu Q_0(\nu)}{s^n}, \quad (2.33)$$

and note that for $\nu\neq1$, the zeros of $\Lambda(\nu,s)$ and $\Lambda''(\nu,s)$ coincide. Let $S''_0(\nu)$ be a zero of $\Lambda''(\nu,s)$, i.e., $\Lambda''[\nu,S''_0(\nu)]=0$, and note that

$$
\begin{aligned}
&|S_0(\nu) - S_0''(\nu)| \\
&\quad = |S_0(\nu) - \xi_0(\nu) + \xi_0(\nu) - S_0''(\nu)| \\
&\quad = 0. \qquad (2.34)
\end{aligned}
$$

Using the same argument as in Lemma 5 completes the proof. Note, for example, that as $\nu \to 1$ along the real axis that $\xi_0(\nu) \to \infty$. Further if $\xi_0^+(\mu)$ and $\xi_0^-(\mu)$ are the limits of $\xi_0(\nu)$ as $\nu \to \mu \in (-1, +1)$ from the upper and lower complex plane, respectively, then

$$\lim_{\mu \to 1} \xi_0^{\pm}(\mu) = +\infty \mp i\pi a_0(1)/2 \equiv \xi_0^{\pm}. \qquad (2.35)$$

Let $\Lambda^+(\mu,s)$ and $\Lambda^-(\mu,s)$ be the limits of $\Lambda(\nu,s)$ as $\nu \to \mu \in (-1, +1)$ from the upper and lower half complex ν plane, respectively, and consider for fixed μ the roots of $\Lambda^{\pm}(\mu,s) = 0$. There are $N^* + 1$ such roots, some of which may be multiple roots if μ is a zero of the discriminant of $\Lambda^{\pm}(\mu,s)$. Let $S_j^{\pm}(\mu)$, $j = 0,1,\ldots,N^*$, be the functions generated by such roots as μ takes on values along $(-1, +1)$.

Lemma 7: Each function $S_j^{\pm}$ is continuous on $(-1, +1)$.

Proof: The proof will be illustrated for $S_j^+(\mu)$. The proof for $S_j^-(\mu)$ follows in an analogous manner. Let s_0 be a root of $\Lambda^+(\mu_0,s) = 0$, where $\mu_0 \in (-1, +)$ is not a zero of the discriminant of $\Lambda^+(\mu,s)$. Further, let K_ϵ be a circle of radius $\epsilon > 0$ centered on s_0 so small that $\Lambda^+(\mu_0,s)$ contains no zero except at the point s_0 itself. Since $\Lambda^+(\mu_0,s)$ is analytic inside of K_ϵ, let $\eta > 0$ be the minimum of $|\Lambda^+(\mu_0,s)|$ on K_ϵ. For fixed s, $\Lambda^+(\mu,s)$ is a continuous function of μ on $(-1, +1)$. Therefore, choose a real interval Δ so small that $|\Lambda^+(\mu_0,s) - \Lambda^+(\mu,s)| < \eta$ for all $\mu \in \Delta$. Thus according to Rouche's theorem

$$\Lambda^+(\mu,s) = \Lambda^+(\mu_0,s) + [\Lambda^+(\mu,s) - \Lambda^+(\mu_0,s)] \qquad (2.36)$$

has only one zero inside K_ϵ for any fixed but arbitrary $\mu \in \Delta$. If $\Lambda^+(\mu_0,s) = 0$ has a k-fold multiple root, then repeating the argument above shows that the circle K_ϵ encloses k zeros of $\Lambda^+(\mu,s)$ for $\mu \in \Delta$. Thus each $S_j^+(\mu)$ is continuous on $(-1, +1)$ and at each zero of the discriminant of $\Lambda^+(\mu,s)$ that corresponds to a k-fold multiple root of $\Lambda^+(\mu,s) = 0$ (e.g., $\mu = 0$) k of the functions $S_j^+(\mu)$ take on the same value. The labeling of the functions $S_j^+(\mu)$ is given by the following.

Lemma 8: The limits of $S_j(\nu)$, $j = 0,\ldots,N^*$, as $\nu \to \mu \in (-1, +1)$ from the upper and lower complex plane are $S_j^+(\mu)$ and $S_j^-(\mu)$, respectively.

Proof: The proof of this lemma is similar to Lemma 7 and again the proof will be illustrated for $S_j^+(\mu)$. The proof for $S_j^-(\mu)$ follows in an analogous manner. As in Lemma 7, let s_0 be a root of $\Lambda^+(\mu_0,s) = 0$, where $\mu_0 \in (-1, +1)$ is not a zero of the discriminant of $\Lambda^+(\mu,s)$. Again let K_ϵ be a circle of radius $\epsilon > 0$ centered on s_0 so small that $\Lambda^+(\mu_0,s)$ encloses only the zero at s_0 itself. Let $\eta > 0$ be the minimum of $|\Lambda^+(\mu_0,s)|$ on K_ϵ. Finally let K_δ be a circle centered on μ_0 so small that $|\Lambda^+(\mu_0,s) - \Lambda(\nu,s)| < \eta$ for any ν with $\operatorname{Re} \nu > 0$ inside K_δ. Thus again from Rouche's theorem

$$\Lambda(\nu,s) = \Lambda^+(\mu_0,s) + [\Lambda(\nu,s) - \Lambda^+(\mu_0,s)] \qquad (2.37)$$

has only one zero inside K_ϵ for any fixed but arbitrary point ν in K_δ with $\operatorname{Re} \nu > 0$. If $\Lambda^+(\mu_0,s) = 0$ has a k-fold multiple root at s_0, then the circle K_ϵ will contain k roots of $\Lambda(\nu,s)$.

III. MAIN THEOREM

Consider the contours generated by $s = S_j(\nu)$, $j = 0,1,\ldots,N^*$, as ν varies along the contour C of Fig. 1 as that contour is collapsed (with $\rho \to 0$) onto the real interval $(-1, +1)$. These contours are in fact the contours Γ_j generated parametrically by $s = S_j^{\pm}(\mu)$, $j = 0,1,\ldots,N^*$, as μ varies along the real interval $(-1, +1)$. Note that $S_j^+(-\mu) = \overline{S}_j^+(\mu)$, $S_j^+(\mu) = S_j^-(-\mu)$, and that each of the contours begins and ends at the limit points given by Lemmas 5 and 6. Thus the contour Γ_j starts, say, at ξ_j, varies continuously in the s plane as μ varies from -1 to 0 along the top of the cut, passes through zero at $\mu = 0$, traces out its complex conjugate as μ continues to vary from 0 to $+1$ along the top of the cut, and finally retraces itself as μ varies from $+1$ to -1 along the bottom of the cut. That the contours do not cross the real s plane axis except at $s = 0$ and $s = \xi_j$, $j > 0$, is clear. For if $S_j^+(\mu_0) = s_0 \in \mathbb{R}$ for some value of $\mu_0 \in (-1, +1)$, that would imply that $\Lambda^+(\mu_0,s_0) = 0$ in contradiction to the results cited in Sec. I.

The contours Γ_j, $j = 1,\ldots,N^*$, are closed. (The contour Γ_0 can be regarded as closed if it is regarded as being closed at infinity.) The contours Γ_j have positive (counterclockwise) orientation. Since $\operatorname{Im} S_j^+(\mu) \neq 0$ for $0 < |\mu| < 1$, it is sufficient to show positive orientation of the Γ_j by demonstrating for some μ_0 with $0 < \mu_0 < 1$ that

$$
\begin{aligned}
\operatorname{Im} S_j^+(\mu_0) &< 0, \quad \text{if } \xi_j > 0, \\
&> 0, \quad \text{if } \xi_j < 0.
\end{aligned}
\qquad (3.1)
$$

Note first that $\operatorname{Im} S_0^+(\mu) > 0$, since

$$\lim_{\mu \to 1} \operatorname{Im} S_0^+(\mu) = -a_0(1)\pi/2, \qquad (3.2)$$

with $a_0(1) > 0$. If $\Lambda^+(\mu,s)$ is evaluated from Eq. (2.2), it is easy to see that

$$\lim_{\mu \to 1} \operatorname{Im} \Lambda^+(\mu,s) = \gamma(1,s)\pi/2. \qquad (3.3)$$

Now order the zeros of $\gamma(1,s)$ according to

$$\xi_{m_1} > \xi_{m_2} > \cdots > \xi_{m_Q} > 0 > \xi_{m_{Q+1}} > \cdots > \xi_{m_{N^*}}, \qquad (3.4)$$

and choose $1 > \mu_0 > 0$ so that either

$$
\begin{aligned}
&\xi_{m_{q+1}} < \operatorname{Re} S_{m_q}^+(\mu_0) < \xi_{m_q}, \quad \text{if } 1<q<Q, \\
&0 < \operatorname{Re} S_{m_q}^+(\mu_0) < \xi_{m_Q}, \quad \text{if } q = Q, \\
&\xi_{m_{Q+1}} < \operatorname{Re} S_{m_q}^+(\mu_0) < 0, \quad \text{if } q = Q+1, \text{ or} \\
&\xi_{m_q} < \operatorname{Re} S_{m_q}^+(\mu_0) < \xi_{m_{q-1}}, \quad \text{if } N^* > q > Q+1.
\end{aligned}
\qquad (3.5)
$$

If $s = \operatorname{Re} S_{m_q}^+(\mu_0)$, $q = 1,\ldots,N^*$, then

$$\Lambda^+[\mu,\mathrm{Re}\,S^+_{m_q}(\mu_0)] = -\mathrm{Im}\,S^+_{m_q}(\mu_0)[\mathrm{Re}\,S^+_{m_q}(\mu_0)-\mathrm{Re}\,S_0^+(\mu)] \times\left\{\prod_{\substack{j=1\\ j\neq m_q}}^{N^*}[\mathrm{Re}\,S^+_{m_q}(\mu_0)-\mathrm{Re}\,S_j^+(\mu)]+T(\mu,\mu_0)\right\}, \tag{3.6}$$

where $T(\mu,\mu_0)$ is a function such that $T(\mu,\mu_0)\to 0$ as $\mu\to 1$. Thus for μ sufficiently close to 1, Eqs. (3.3) and (3.6) yield

$$\begin{aligned}&\mathrm{sgn}(\Lambda^+[\mu,\mathrm{Re}\,S^+_{mq}(\mu_0)])\\&\quad=\mathrm{sgn}\Bigg[-\mathrm{Im}\,S^+_{m_q}(\mu_0)\\&\qquad\times\prod_{\substack{j=0\\ j\neq m_q}}^{N^*}(\mathrm{Re}\,S^+_{m_q}(\mu_0)-\mathrm{Re}\,S_j^+(\mu))\Bigg]\\&\quad=\mathrm{sgn}(\gamma[1,\mathrm{Re}\,S^+_{m_q}(\mu_0)]).\end{aligned} \tag{3.7}$$

Moreover, since $\lim_{s\to\infty}\gamma(1,s)\to\infty$, then $\mathrm{sgn}[\gamma(1,s)] = \mathrm{sgn}[(-1)^q]$ if $\xi_{m_{q+1}}<s<\xi_{m_q}$. Thus if $\mathrm{Re}\,S^+_{m_q}(\mu_0)$ is chosen by Eq. (3.5), then Eq. (3.7) gives

$$\begin{aligned}&\mathrm{sgn}[-(-1)^q\,\mathrm{Im}\,S^+_{m_q}(\mu_0)]\\&\quad=\mathrm{sgn}[(-1)^q],\quad \text{if}\ \xi_{m_q}>0,\\&\quad=\mathrm{sgn}[(-1)^{q+1}],\quad \text{if}\ \xi_{m_q}<0.\end{aligned} \tag{3.8}$$

Theorem: Let $I(\Gamma_j)$ and $E(\Gamma_j)$ represent the interior and exterior of the contours Γ_j, $j=0,1,\ldots,N^*$, respectively, and let

$$s\in\bigcap_{j=0}^{P-1} I(\Gamma_{m_j})\bigcap_{j=P}^{N^*} E(\Gamma_{m_j}). \tag{3.9}$$

In other words, let s lie in the interior of P of the contours Γ_j and the exterior to all the other Γ_j. The number of roots of $\Lambda(\nu,s)=0$ is $\Sigma_{j=0}^{P-1} N_{m_j}$, where N_{m_j} is the index of s with respect to Γ_j. Further, if s is real and satisfies Eq. (3.9) then $N_{m_j}=1$ and $M=P+1$, i.e., just equal to the number of contours Γ_j in which s lies.

Proof: As indicated in Sec. I, $\Lambda(\infty,s)$ is a constant and the number of zeros of $\Lambda(\nu,s)$ is given by the change in the argument of $\Lambda(\nu,s)$ along the contour C in Fig. 1 as the contour is collapsed (with $\rho\to 0$) onto the real interval $(-1,+1)$. This procedure yields [cf. Eq. (1.12)]

$$M=(1/\pi)\Delta_C\,\mathrm{Arg}\,\Lambda^+(\mu,s), \tag{3.10}$$

where $\Delta_C\,\mathrm{Arg}\,\Lambda^+(\mu,s)$ represents the change in the argument along the directed line from -1 to $+1$. Thus

$$\begin{aligned}M&=\Delta_C\,\mathrm{Arg}\prod_{j=0}^{N^*}[s-S_j^+(\mu)]\\&=\sum_{j=0}^{N^*}\Delta_C\,\mathrm{Arg}[s-S_j^+(\mu)]=\sum_{j=0}^{P-1}N_{m_j}.\end{aligned} \tag{3.11}$$

If $s\in\mathbf{R}\subset I(\Gamma_{m_j})$ then $N_{m_j}=1$ since Γ_{m_j} does not cross the real axis for $0<s<\xi_{m_j}$. Of course if s does not lie inside of any

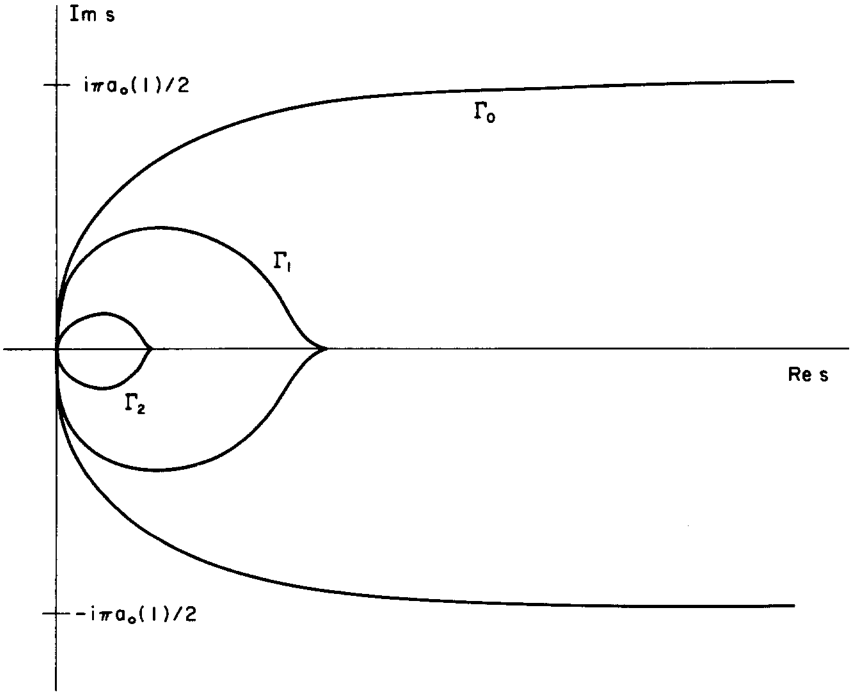

FIG. 2. Contours Γ_0, Γ_1, and Γ_2 for $f_1=0.2$ and $f_2=0.05$. Scale of Γ_0 reduced by factor of 6 and scale of Γ_2 enlarged by factor of 2.

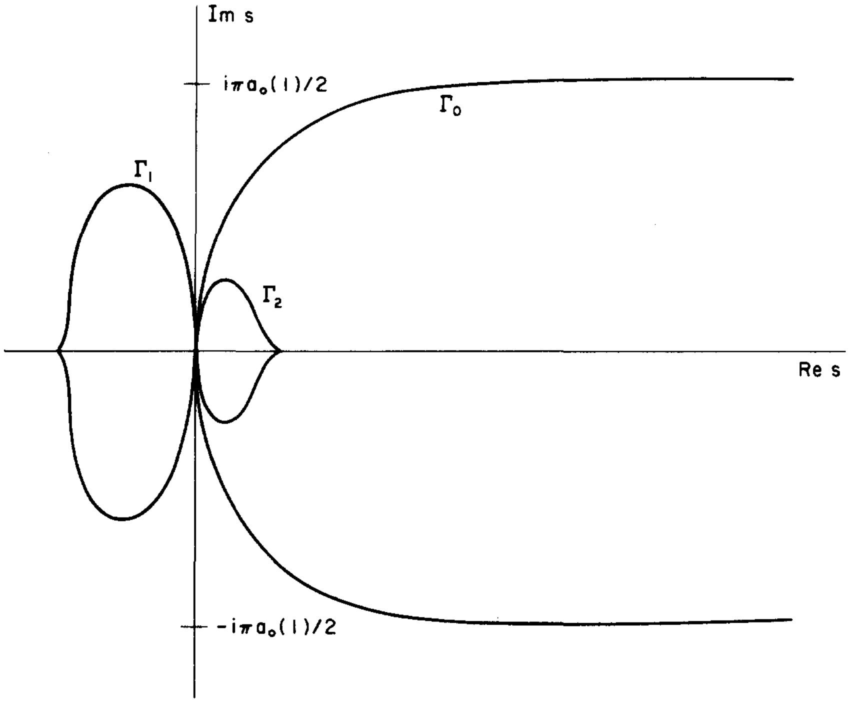

FIG. 3. Contours Γ_0, Γ_1, and Γ_2 for $f_1 = -0.1$ and $f_2 = 0.05$. Scale of Γ_0 reduced by factor of 5 and scale of Γ_2 enlarged by factor of 2.

of the contours then $M = 0$, i.e., $\Lambda(\nu,s)$ has no zeros. The number of zeros of $\Lambda(\nu,s)$ is intimately connected to the zeros of $\gamma(1,s)$, i.e., to the ξ_j, $j = 1,\dots,N^*$.

Corollary: If $s\in\mathbf{R}$ satisfies $0<\xi_{m_{j+1}}<s<\xi_{m_j}$, where the ξ_{m_j} are ordered according to Eq. (3.4), then the number of pairs of zeros of $\Lambda(\nu,s)$ is $j+1$. Further, if $c_q^+ = 1/\xi_{m_j}$, then the test of Sec. I follows directly.

For a numerical illustration of the mappings $s = S_j^+(\mu)$, $j = 0,1,\dots,N^*$, consider Figs. 2–5. These curves were generated for the case $N = 2$ by solving $\Lambda^+(\mu,s) = 0$ for s as μ varies from 0 to $+1$. For a numerical illustration of the roots of $\gamma(1,s)$, consider Table I. In this calculation

$$f(\mu,\mu') = \sum_{n=0}^{N} (2n+1) f_n^J P_n(\mu) P_n(\mu'), \qquad (3.12)$$

where the expansion coefficients are given by the recursion relation[9]

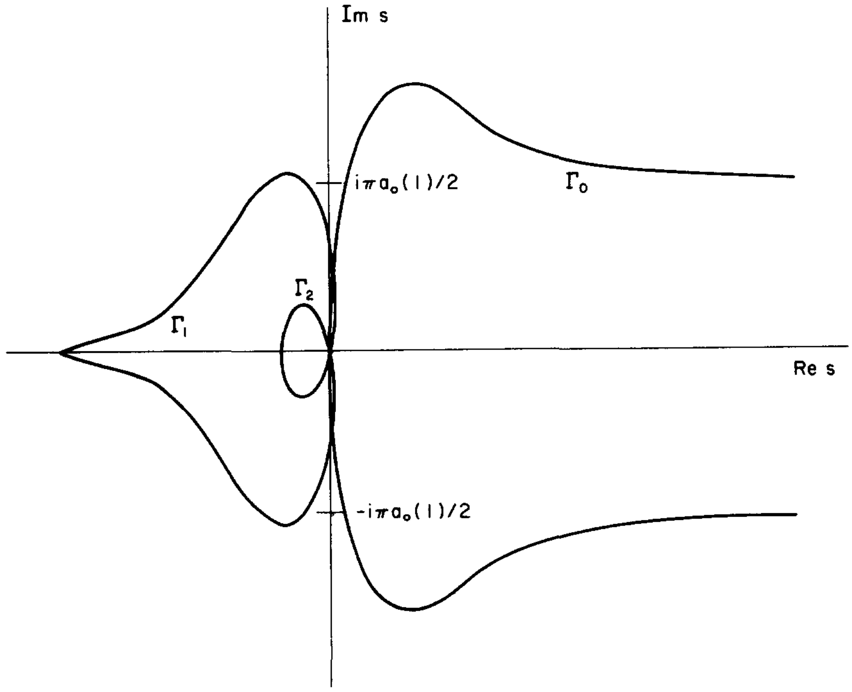

FIG. 4. Contours of Γ_0, Γ_1, and Γ_2 for $f_1 = -0.1$ and $f_2 = -0.05$. Scale of Γ_0 reduced by factor of 3 and scale of Γ_2 enlarged by factor of 4.

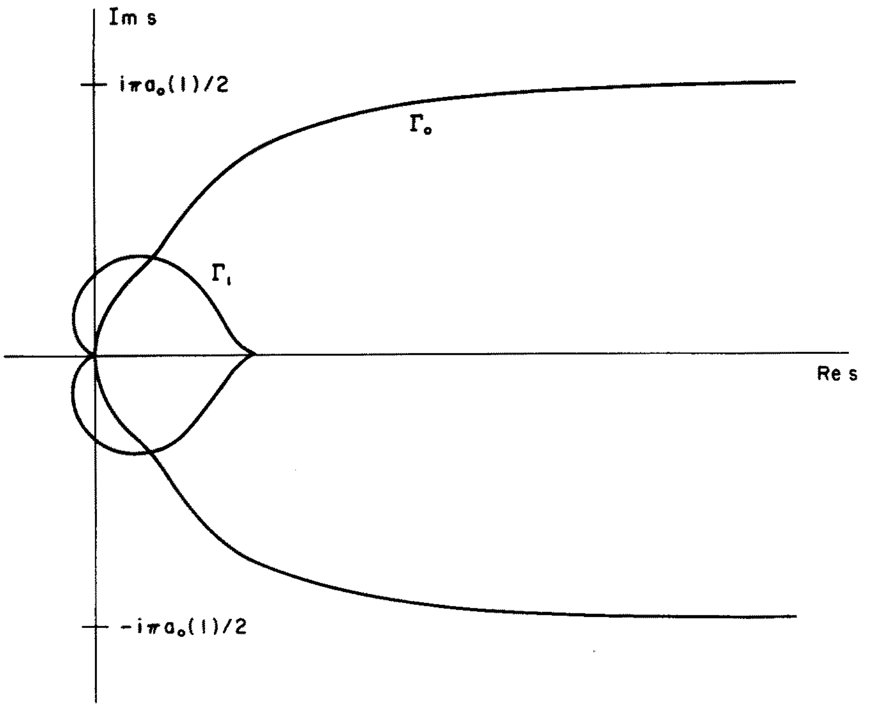

FIG. 5. Contours Γ_0 and Γ_1 for $f_1 = 0$ and $f_2 = 0.1$. Scale of Γ_0 reduced by factor of 3.

$$f_n^j = \frac{j+1}{2j(2n+1)}\left[\frac{n}{(2n-1)} f_{n-1}^{j-1} + f_n^{j-1} + \frac{n+1}{2n+3} f_{n+1}^{j-1}\right], \tag{3.13}$$

with $f_0^j = 1$, $j = 0,1,\ldots$, and $f_n^j = 0$ if $n > j$. The calculations in Table I were made with $J = 50$. Other numerical results agree with the azimuthally symmetric results reported by Shultis and Hill.[11]

IV. CONCLUDING REMARKS

It seems appropriate to conclude with a couple of remarks about the nature of the zeros of $\Lambda(\nu,s)$ and $\gamma(\nu,s)$. Several years ago Kuščer[12] pointed out that for the case $N = 2$ that the zeros of $\Lambda_c(\nu)$ could be complex. The advantage of the present analysis is that it points out that the zeros of $\Lambda(\nu,s)$ become complex (even for real s) whenever the discriminant $D(\nu)$ has zeros on the imaginary axis. Stated somewhat differently, the zeros of $\Lambda(\nu,s)$ are mapped via the $S_j(\nu)$ from the ν plane to the s plane and that map is conformal as long as the path in the ν plane avoids the cuts as described in Sec. II. In particular, the imaginary axis in the ν plane is conformally mapped to the real axis in the s plane as ν marches in from infinity. This conformal mapping is broken if a zero in the discriminant of $\Lambda(\nu,s)$ is encountered, resulting with complex zeros of $\Lambda(\nu,s)$. One can quickly show that this is just the situation for the special case considered by Kuščer.

TABLE I. Zeros of $\gamma(1,s)$. The last row is the reported number of pairs of zeros of $\Lambda_c(\nu)$ for $c = 0.95$ (see Ref. 10).

		Order of scattering N		
4	6	8	10	15
3.0765	5.0607	6.8243	8.0824	9.2031
1.0388	1.6213	2.1600	2.5665	2.9677
0.6067	0.8576	1.1052	1.2990	1.5039
0.4645	0.5854	0.7131	0.8201	0.9402
	0.4683	0.5405	0.5986	0.6710
	0.3396	0.4422	0.4893	0.5332
		0.3265	0.3920	0.4457
		0.2057	0.2842	0.3491
			0.1860	0.2554
			0.1047	0.1762
				0.1147
				0.0701
				0.0399
				0.0206
				0.0091
2	3	4	4	4

Somewhat similar related remarks can be made about the zeros of $\gamma(\nu,s)$. It has been shown that the number of zeros of $\Lambda(\nu,s)$ are related to the zeros of $\gamma(1,s)$. If the number of pairs of zeros of $\gamma(\nu,s)$ (for fixed s) that lie in the interval $(-1,+1)$ is denoted by α, the discussion in Sec. I indicates that the number of pairs of zeros M of $\Lambda(\nu,s)$ must be bounded $M \leqslant \alpha + 1$. Further, numerical calculation with real s not too small (c not too large) suggest that M can be, in fact, just equal to $\alpha + 1$. To see the reason for this consider the fact that $\gamma(\nu,s) = 0$ generates an algebraic function, say $\nu(s)$, each branch of which conformally maps the appropriately cut s plane to the ν plane. Note that $\nu(\xi_j) = 1$, $j = 1,2,\ldots,N^*$. The number of zeros of $\gamma(\nu,s)$ must always be sufficient to satisfy the main theorem. Thus there is always a certain branch of $\nu(s)$ that maps the interval $(\xi_j,0)$ in the s plane to the real interval $(-1,+1)$ in the ν plane, and that mapping will be conformal (and thus one-to-one) if the discriminant of $\gamma(\nu,s)$ does not vanish on the interval $(\xi_j,0)$. Therefore, if the set of expansion coefficients $\{f_n\}$ is such that the discriminant of $\gamma(\nu,s)$ does not vanish on any of the intervals $(\xi_j,0)$, $j = 1,2,\ldots,$N*, in the s plane, then indeed $M = \alpha + 1$. This is certainly the case for $N = 0$ and $N = 1$. However, one can show quite easily that for the case $N = 2$, $f_1 < 0$, and $f_2 < 0$ that the discriminant does vanish for s small enough. However, it is apparent that there always exist values of s greater than the largest zero of the discriminant of $\gamma(\nu,s)$ for which the number of pairs of zeros of $\Lambda(\nu,s)$ is always given by $M = \alpha + 1$.

[1] K. M. Case and P. F. Zweifel, *Linear Transport Theory* (Addison–Wesley, Reading, MA, 1967).

[2] J. R. Mika, Nucl. Sci. Eng. **11**, 415 (1961).

[3] K. M. Case, J. Math. Phys. **15**, 974 (1974).

[4] R. J. Hangelbroek, Transp. Theory Stat. Phys. **8**, 133 (1979).

[5] C. G. Lekkerkerker, Proc. R. Soc. Edinburgh Sec. A **83**, 303 (1979).

[6] A. Leonard and T. W. Mullikin, J. Math. Phys. **5**, 399 (1964).

[7] T-Y. Dawn and I-J. Chen, Nucl. Sci. Eng. **72**, 237 (1979).

[8] E. Inönü, J. Math. Phys. **11**, 568 (1970).

[9] R. Bowden, F. J. McCrosson, and E. A. Rhodes, J. Math. Phys. **9**, 753 (1968).

[10] J. K. Shultis, J. Comput. Phys. **11**, 109 (1973).

[11] J. K. Shultis and T. R. Hill, Nucl. Sci. Eng. **59**, 53 (1976).

[12] I. Kuščer, Nucl. Sci. Eng. **38**, 175 (1969).

Multigroup transport equations with nondiagonal cross-section matrices

B. L. Willis
Center for Transport Theory and Mathematical Physics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

C. V. M. van der Mee
Department of Mathematics and Computer Science, Clarkson University, Potsdam, New York 13676

(Received 8 August 1985; accepted for publication 26 November 1985)

Multigroup transport equations with nondiagonal cross-section matrices are studied using the Wiener–Hopf method. Formulas for the solution and the exit distribution are given in terms of the factorization of the symbol of the Wiener–Hopf equation. Unlike the formulas for a diagonal cross-section matrix, these formulas involve derivatives of the H-functions. For the case of two groups, the H-functions are computed explicitly.

I. INTRODUCTION

Multigroup transport equations with nondiagonal and possibly nondiagonalizable cross-section matrices have been proposed as a model of, for example, neutron transport in reactors.[1,2] In this paper, transport equations with nondiagonalizable cross-section matrices are studied by making use of the Wiener–Hopf method. In Sec. II an integral equation equivalent to the transport equation is derived along with expressions connecting the solutions of the integral equation to the solutions of the transport equation. In Secs. III and IV we outline the Wiener–Hopf method. In Sec. V the Wiener–Hopf factorization is constructed explicitly for the two-group case. For the general N-group problem, we are not able to construct the factorization; the best that we are able to do is derive the generalized Chandrasekhar H-equations and to set up a numerical scheme for computing the H-functions. This work will be published in another paper, where we consider a more general scattering matrix. Finally, in Secs. VI and VII we determine the exit distribution and the solution in terms of the H-functions. In these two sections we do not limit ourselves to the two-group problem; instead we consider the N-group problem in anticipation of the above-mentioned generalization.

Briefly, transport equations with nondiagonal cross-section matrices occur when the energy dependence of the cross section is expanded in terms of orthogonal functions, and then the method of weighted residuals is applied to determine equations for the coefficients of the expansion. The method of weighted residuals is discussed by Stacey[1] and by Ames,[3] where different choices of the orthogonal functions and the weights are considered and the physical reasons behind the choices are given. If this procedure is followed for the problem of radiative transfer with the assumption of a uniform or picket fence model,[4] then the resulting vector equation has the form

$$\mu\,\partial_x F(x,\mu) + \Sigma F(x,\mu) = \frac{1}{2} C \int_{-1}^{+1} F(x,\mu')d\mu', \qquad (1)$$

where the matrix C is noninvertible. A derivation of these results can be found in Siewert and Zweifel,[4] the only difference being that the cross-section matrix Σ is no longer necessarily diagonal. If Σ is diagonalizable, then a similarity transformation will reduce Eq. (1) to the problem considered in Ref. 4. More generally, Eq. (1) is solvable for the case that the matrices Σ and C are simultaneously upper triangularizable. In such a case, the problem reduces to a system of uncoupled inhomogeneous scalar Wiener–Hopf equations.

In the following, the simplest equation of the form (1) that does not satisfy either one of the two above conditions will be studied. In particular, the two-group equation defined by

$$\Sigma = \begin{vmatrix} 1 & \alpha \\ 0 & 1 \end{vmatrix}, \quad C = \begin{vmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{vmatrix}, \qquad (2)$$

will be studied with $\alpha \neq 0$ and $c_{21} \neq 0$. A similarity transformation can always be applied to set $\alpha = 1$, but for bookkeeping purposes it is convenient to keep α as a parameter so that the limit $\alpha \to 0$ is apparent. A direct calculation shows that, for $\alpha \neq 0$ and $c_{21} \neq 0$, Σ and C are not simultaneously upper triangularizable, whence the conditions $\alpha \neq 0$, $c_{21} \neq 0$. In this paper we will study Eq. (1) with Σ and C defined by Eq. (2), along with half-space boundary conditions given by

$$F(0,\mu) = \Phi(\mu), \quad \mu > 0, \qquad (3a)$$

$$F(x,\mu) \to 0, \quad x \to \infty. \qquad (3b)$$

Equation (3b) holds true for each component separately.

II. AN EQUIVALENT INTEGRAL EQUATION

Equation (1) is studied using the Wiener–Hopf method. To carry out the procedure, an equivalent integral equation is sought. If G is defined by

$$G(x) = \int_{-1}^{+1} F(x,\mu)d\mu, \qquad (4)$$

then an integral equation for G can be derived analogously to the one-speed case.[5] The result is

$$G(x) = U(x) + \frac{1}{2} \int_0^\infty \mathrm{Ei}_\Sigma(|x-s|)CG(s)ds, \qquad (5a)$$

where

$$U(x) = \int_0^1 e^{-x\Sigma/\mu}\Phi(\mu)d\mu. \qquad (5b)$$

The function Ei_Σ is defined in terms of the exponential integral[6] and its derivative by

$$\mathrm{Ei}_\Sigma(z) = \int_0^1 \mu^{-1} e^{-x\Sigma/\mu}\, d\mu = \begin{vmatrix} E_1(z) & \alpha z E_1'(z) \\ 0 & E_1(z) \end{vmatrix}. \tag{5c}$$

Once G is known, F can be computed using the formulas

$$F(x,\mu) = -\frac{1}{2\mu}\int_x^\infty e^{-(x-s)\Sigma/\mu} CG(s)ds, \quad \mu<0, \tag{6a}$$

and, for $\mu>0$,

$$F(x,\mu) = e^{-\Sigma x/\mu}\Phi(\mu) + \frac{1}{2\mu}\int_0^x e^{-(x-s)\Sigma/\mu} CG(s)ds. \tag{6b}$$

The matrix-valued function $e^{-\Sigma/\mu}$ is easy to compute if the Jordan decomposition of Σ given by

$$\Sigma = I + M, \quad M^2 = 0, \tag{7}$$

is used. It is easy to check that

$$e^{-x\Sigma/\mu} = \begin{vmatrix} 1 & \alpha x/\mu \\ 0 & 1 \end{vmatrix} e^{-x/\mu}. \tag{8}$$

III. THE WIENER-HOPF METHOD OF SOLUTION

Following the standard notation, define the functions $G^\pm$ and $U^\pm$ by

$$G^\pm(x) = \begin{cases} G(x), & \pm x>0, \\ 0, & \pm x<0, \end{cases} \tag{9}$$

and similarly for $U^\pm$. With these definitions, Eq. (5a) can be written as a convolution equation on $(-\infty,\infty)$, namely

$$G^+(x) + G^-(x) = U^+(x) + \frac{1}{2}\int_{-\infty}^{+\infty} \mathrm{Ei}_\Sigma(|x-s|)CG^+(s)ds. \tag{10a}$$

The Fourier transform of Eq. (10a) yields

$$W(\lambda)\hat{G}^+(\lambda) + \hat{G}^-(\lambda) = \hat{U}^+(\lambda). \tag{10b}$$

Here the Fourier transform of a function F is denoted as $\hat{F}$, where

$$\hat{F}(\lambda) = \int_{-\infty}^{+\infty} e^{i\lambda x} F(x)dx.$$

The matrix-valued function W is the symbol of the Wiener–Hopf equation (10a) and is given by

$$W(\lambda) = I - ((1/\lambda)\tan^{-1}\lambda)C + [1/(1+\lambda^2)]MC. \tag{11}$$

The nilpotent matrix M has already been introduced in Eq. (7).

IV. FACTORIZATION OF THE SYMBOL

The crucial step in the Wiener–Hopf method is the construction of the Wiener–Hopf factorization of the symbol. This paper will only consider the canonical Wiener–Hopf (WH) factorization. A canonical WH factorization is a pair of functions $W^\pm$ such that

$$W(\lambda) = W^-(\lambda)W^+(\lambda), \quad \lambda\in\mathbb{R}_\infty = \mathbb{R}\cup\{\pm\infty\}, \tag{12}$$

where the matrix function W^+ (W^-) is analytic in the open upper (lower) half-plane, and continuous and invertible in the closed upper (lower) half-plane. As in the one-speed case, a factorization of the form (12) does not exist for all possible choices of Σ and C. In fact, in the one-speed case, a canonical factorization exists only for $c<1$ (see Ref. 5). A necessary condition for the existence of $W^\pm$ is that $W(\lambda)$ is invertible for $\lambda\in\mathbb{R}_\infty$, i.e., $\det W(\lambda) \pm 0$ for $\lambda\in\mathbb{R}_\infty$. For this reason one should study the zeros of det W. Explicitly, det W is given by

$$\det W(\lambda) = 1 - \mathrm{tr}\, C[\lambda^{-1}\tan^{-1}\lambda] + \alpha c_{21}[1+\lambda^2]^{-1}. \tag{13}$$

Here, tr C denotes the trace of C, and the assumption that $\det C = 0$ has been used. Observe that the dispersion function has branch points at $\pm i$. We will always choose the branch cuts to be the lines $z = it$, $|t| \geqslant 1$. Therefore, the dispersion function is analytic in the region $\mathbb{C}\setminus\{z\in\mathbb{C}: z = it, |t|\geqslant 1, t\in\mathbb{R}\}$. Note that

$$\lim \det W(\lambda) = 1, \quad |\lambda|\to\infty, \tag{14}$$

holds inside the region of analyticity. Furthermore, det W satisfies the symmetries

$$[\det W(\lambda)]^* = \det W(\lambda^*), \tag{15a}$$

$$\det W(-\lambda) = \det W(\lambda). \tag{15b}$$

The superscript * denotes complex conjugation. These symmetries imply that λ_0 is a zero of the dispersion function if and only if both λ_0^* and $-\lambda_0$ are zeros of the dispersion function. Therefore the dispersion function must have an even number of zeros. The symmetries [Eqs. (15a) and (15b)] along with the behavior of det W at infinity [Eq. (14)] allow one to compute the number of zeros of the dispersion function by computing the change of the argument of det W along the branch cuts, the so-called Nyquist method,[5] just as is done in the one-group case.[5] We apply the

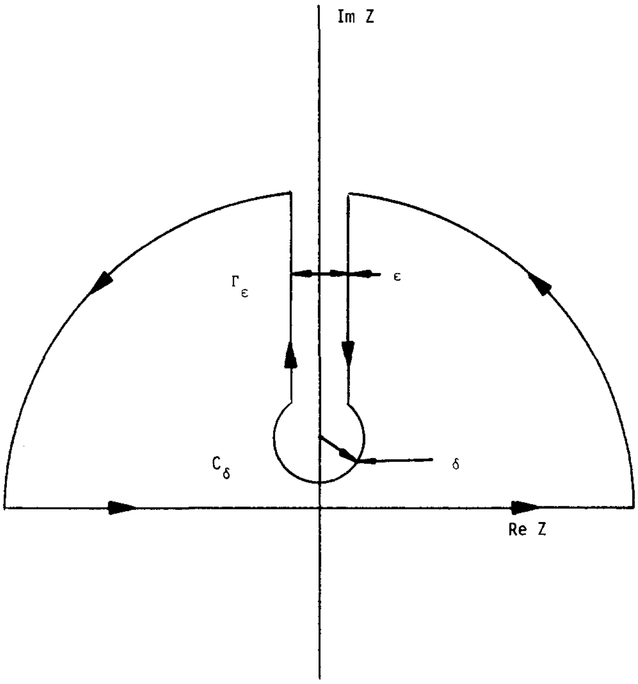

FIG. 1. Contour for computing Δ arg det W.

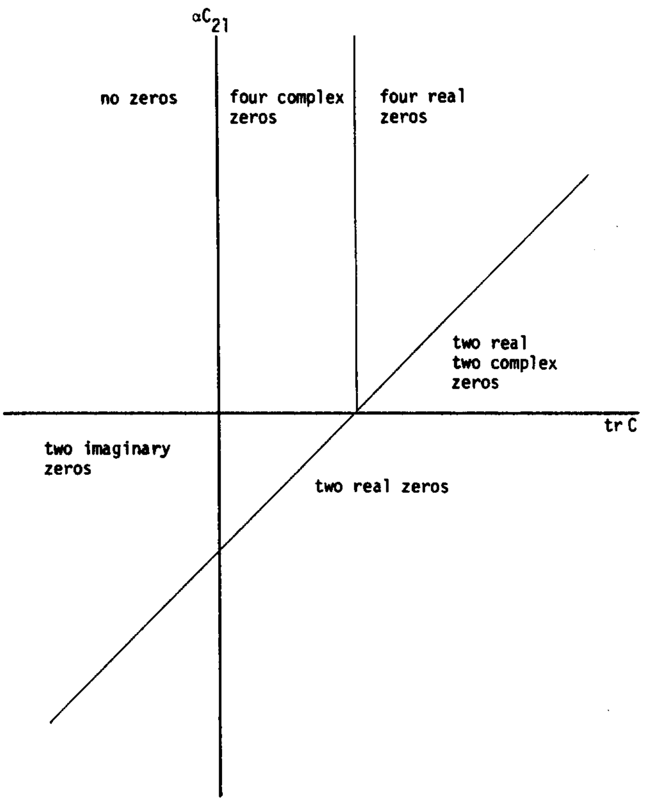

FIG. 2. The zeros of det W in the tr C, αC_{21} plane.

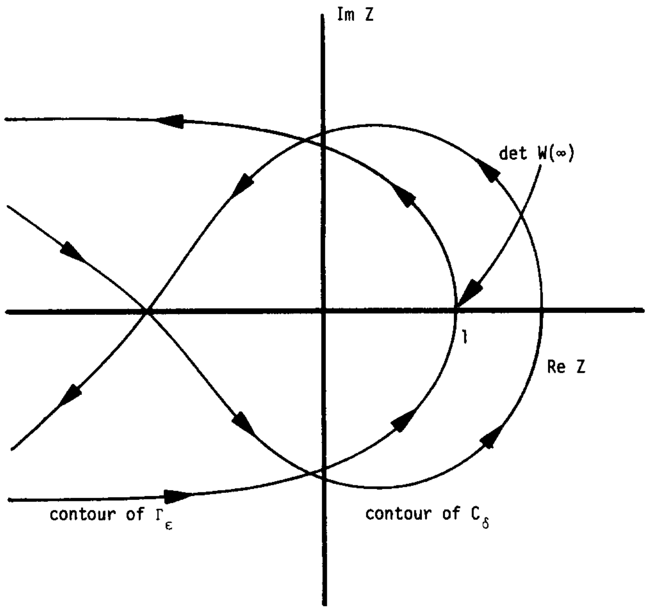

FIG. 3. Nyquist diagram for $\alpha C_{21} > 0$ and tr $C > 0$.

argument principle to the contour in Fig. 1. This problem divides into three special cases: (i) tr $C = 0$, (ii) $\alpha c_{21} = 0$, and (iii) both tr $C \neq 0$ and $\alpha c_{21} \neq 0$. The case tr $C = 0$ is solved easily by algebra, and the dispersion function for $\alpha c_{21} = 0$ is identical to the one-group dispersion function so that the number of zeros is known.[7] These results are summarized in Fig. 2. Case (iii) requires special attention. Unlike the one-group dispersion function, i.e., the case $\alpha c_{21} = 0$, the dispersion function now has poles at the branch points due to the term $\alpha c_{21}[1+\lambda^2]^{-1}$ [see Eq. (13)]. For this case, the change in the argument when rounding the branch points is now important. For this reason, the change in the argument of the dispersion function (denoted by Δ arg det W) along the contour in Fig. 1 will be considered in the limit as ϵ and $\delta \downarrow 0$. First we study Δ arg det W along the straight lines Γ_ϵ by taking the limit $\epsilon \downarrow 0$ while keeping δ a constant, then we study Δ arg det W along the circle C_δ by taking the limit as $\delta \downarrow 0$. Along the lines Γ_ϵ the real and imaginary parts of the boundary values of det W are given by

$$\operatorname{Re} \det(\pm 0 + iy) = 1 - \frac{\operatorname{tr} C}{2y} \ln \left| \frac{1+y}{1-y} \right| + \frac{\alpha c_{21}}{1-y^2}, \tag{16a}$$

$$\operatorname{Im} \det W(\pm 0 + iy) = \pm (\pi \operatorname{tr} C)/2y. \tag{16b}$$

[Note that Eq. (16b) proves that det W is nonvanishing on the contour Γ as required by the argument principle.] With these formulas, the Nyquist diagram for the contour Γ_ϵ can be sketched; for the case $\alpha c_{21} > 0$ and tr $C > 0$ the result is shown in Fig. 3. The diagrams for the other possible choices of signs of αc_{21} and tr C are similar. To complete the Nyquist diagrams, the contour C_δ must now be considered. Along the C_δ, the pole term $(1+\lambda^2)^{-1}$ dominates, and the contour approaches a circle at infinity as $\delta \downarrow 0$. With this information, the Nyquist diagrams can be sketched (see Fig. 3), and the number of zeros of the dispersion function can be deduced. Now that the number of zeros of the dispersion function is known, the remaining task is to determine whether the zeros are purely real, purely imaginary, or neither. The graphs of the real and imaginary parts of the dispersion function are easy to sketch, so it is easy to determine if the dispersion function has a real zero. These results are also summarized in Fig. 2. Thus we can conclude that $W(\lambda)$, $\lambda \in \mathbf{R}_\infty$, is invertible for $1 + \alpha c_{21} >$ tr C, and tr $C > 1$. As we previously mentioned, these conditions give a necessary condition for the existence of a WH factorization. In the next section, these conditions will be shown to be sufficient by explicit calculation of the factorization of W.

V. CONSTRUCTION OF THE WIENER–HOPF FACTORIZATION

The matrix valued function to be factorized is

$$W(\lambda) = I - (\lambda^{-1} \tan^{-1} \lambda) C + (1+\lambda^2)^{-1} MC; \tag{17}$$

the matrix M has been defined in Eq. (7). In general it is not known how to construct the Wiener–Hopf factorization of matrices, but Cebotarev[8] has shown how to factorize any upper triangular matrix. The matrix (17) can be made upper triangular by a similarity transformation with constant elements. One possible transformation is given by

$$S = \begin{vmatrix} -c_{22} & c_{21} \\ c_{21} & 0 \end{vmatrix} \begin{vmatrix} 1 & A \\ 0 & 1 \end{vmatrix}, \tag{18a}$$

where

$$A = c_{21} (\operatorname{tr} C)^{-1}, \quad \text{if tr } C \neq 0,$$

and

$$A=0, \quad \text{if} \operatorname{tr} C=0. \tag{18b}$$

The matrix S is always invertible, because $\det S = -c_{21}^2$, which is nonvanishing by assumption. The particular choice for S has been made with forethought, so that the transformed matrices MC and C are especially simple. Explicitly the transformed matrices are

$$S^{-1}(I+M)CS = \begin{vmatrix} 0 & 0 \\ 0 & \alpha c_{21} \end{vmatrix}, \quad \operatorname{tr} C = 0, \tag{19a}$$

and, for $\operatorname{tr} C \neq 0$,

$$S^{-1}(I+M)CS = \begin{vmatrix} 0 & -\alpha c_{21}^2 (\operatorname{tr} C)^{-1} \\ 0 & \operatorname{tr} C + \alpha c_{21} \end{vmatrix}. \tag{19b}$$

The transformed matrix $S^{-1}CS$ is given by the same expression, but with $\alpha = 0$. It is tempting to think that the similarity transformation [Eq. (18a)] applied to the original equation will result in a similar simplification, but this is not the case. The reason is that although C and MC are simultaneously upper triangularizable, C and Σ are not.

The Wiener–Hopf factorization now can be computed. If $S^{-1}WS$ is denoted by $\widetilde{W}$, then

$$\widetilde{W} = \begin{vmatrix} 1 & K(\lambda) \\ 0 & \det W(\lambda) \end{vmatrix}, \tag{20a}$$

where

$$K(\lambda) = -c_{21}\lambda^{-1}\tan^{-1}\lambda, \quad \operatorname{tr} C = 0, \tag{20b}$$

$$K(\lambda) = -\alpha c_{21}^2 (\operatorname{tr} C)^{-1}(1+\lambda^2)^{-1}, \quad \operatorname{tr} C \neq 0. \tag{20c}$$

The function $\widetilde{W}$ is an upper triangular matrix function of second order and the procedure for getting its Wiener–Hopf factorization when it exists has been developed by Cebotarev.[8] Here we follow the method of Ref. 9. First we note that the factors of an upper triangular matrix can be taken to be upper triangular, so we set

$$\widetilde{W}(\lambda) = X(\lambda)Y(\lambda), \tag{21}$$

with X (Y) analytic and invertible in the lower (upper) half-plane. If the elements of the matrices X and Y are denoted by X_{ij} and Y_{ij}, respectively, then the following system of equations results when Eq. (21) is substituted into Eq. (17) and the corresponding matrix elements are equated:

$$1 = X_{11}Y_{11}, \tag{22a}$$

$$1-(\operatorname{tr} C)\lambda^{-1}\tan^{-1}\lambda + \alpha c_{21}(1+\lambda^2)2^{-1} = X_{22}(\lambda)Y_{22}(\lambda), \tag{22b}$$

and

$$-(c_{21}-A\operatorname{tr} C)\lambda^{-1}\tan^{-1}\lambda - \sigma c_{21}A(1+\lambda^2)^{-1} = X_{11}(\lambda)Y_{12}(\lambda) + X_{12}Y_{22}(\lambda). \tag{22c}$$

These equations do not uniquely determine X and Y, since XU and $U^{-1}Y$ satisfy Eqs. (22a)–(22c) whenever X and Y do, where U is any invertible matrix. However, it is consistent to impose the conditions

$$X_{ij}(\infty) = Y_{ij}(\infty) = \delta_{ij}. \tag{23}$$

With these conditions, Eq. (22a) uniquely determines X_{11} and Y_{11} to be

$$X_{11}(\lambda) = Y_{11}(\lambda) = 1, \tag{24}$$

while the solution to Eq. (22b) is given by

$$X_{22}(\lambda) = \exp\left\{\frac{1}{2\pi i}\int_{-\infty+i/2}^{+\infty+i/2} \frac{B(z)}{z-\lambda}\,dz\right\}, \tag{25a}$$

where

$$B(z) = \ln[1-[(\operatorname{tr} C)/z]\tan^{-1}z + \alpha c_{21}(1+z^2)^{-1}]. \tag{25b}$$

The expression for Y_{22} is the same, except that the limits of integration are replaced by $\infty - \frac{1}{2}$ and $-\infty - i/2$. Finally, we determine Y_{12} and X_{12}. To do this, divide Eq. (22c) by Y_{22}, and define the left-hand side of Eq. (22c) to be $L(\lambda)$. Then

$$L(\lambda)/Y_{22}(\lambda) = Y_{12}(\lambda)/Y_{22}(\lambda) + X_{12}(\lambda). \tag{26}$$

The left-hand side of this equation is known, while the right-hand side is the sum of two functions, one analytic in the upper half-plane, the other one analytic in the lower half-plane. To solve for Y_{12} it is only necessary to write LY_{22}^{-1} as the sum of two functions:

$$L(\lambda)/Y_{22}(\lambda) = L^{+}(\lambda) + L^{-}(\lambda), \tag{27}$$

with L^{+} (L^{-}) analytic in the upper (lower) half-plane. Therefore

$$L^{+}(\lambda) = \frac{1}{2\pi i}\int_{-\infty-i/2}^{+\infty-i/2} \frac{L(z)/Y_{22}(z)}{z-\lambda}\,dz, \tag{28a}$$

$$L^{-}(\lambda) = \frac{1}{2\pi i}\int_{-\infty+i/2}^{+\infty+i/2} \frac{L(z)/Y_{22}(z)}{z-\lambda}\,dz. \tag{28b}$$

Now with the definitions

$$Y_{12}(\lambda) = Y_{22}(\lambda)L^{+}(\lambda), \tag{29a}$$

$$X_{12}(\lambda) = L^{-}(\lambda), \tag{29b}$$

the matrices X and Y have all the properties required of a WH factorization.

VI. THE EXIT DISTRIBUTION

Once the canonical Wiener–Hopf factorization has been computed, an expression for the exit distribution, i.e., $F(0,\mu)$ for $\mu < 0$, can be written in terms of the factors of $W(\lambda)$. Unlike for the one-speed case, the exit distribution will involve derivatives of the factors of $W(1/i\lambda)$. The method followed in this section parallels the one given by van der Mee.[10] First, the exit distribution for the two-speed problem defined by Eq. (29) will be derived; then the formulas will be generalized to the N-group problem.

Following Gohberg and Krein,[7] there exists a resolvent kernel $\gamma(\cdot,\cdot)$ so that the general solution to the Wiener–Hopf equation

$$G(x) = \int_0^\infty K(x-y)G(y)dy + U(x) \tag{30a}$$

can be written as

$$G(x) = U(x) + \int_0^\infty \gamma(x,y)U(y)dy, \tag{30b}$$

and the general solution to the transposed equation

$$G(x) = \int_0^\infty G(y)K(y-x)dx + U(x) \qquad (31a)$$

can be written as

$$G(x) = U(x) + \int_0^\infty U(y)\gamma(y,x)dy. \qquad (31b)$$

Note that the resolvent kernels for Eq. (30a) and Eq. (31a) are identical. Returning to Eq. (30a), the exit distribution can be written in terms of G by the formula

$$F(0,\mu) = -\frac{1}{2\mu}\int_0^\infty e^{y\Sigma/\mu}CG(y)dy, \quad \mu<0. \qquad (32)$$

Introducing the resolvent kernel $\gamma(\cdot,\cdot)$ this can be rewritten as

$$F(0,\mu) = -\frac{1}{2\mu}\int_0^\infty \int_0^\infty e^{y\Sigma/\mu} \times C[\delta(y-z) + \gamma(y,z)]U(z)dz\,dy. \qquad (33)$$

If the expression for $U(z)$ in terms of the incident flux is used in Eq. (33), then

$$F(0,\mu) = -\frac{1}{2\mu}\int_0^\infty \int_0^\infty \int_0^1 e^{y\Sigma/\mu}C[\delta(y-z) + \gamma(y,z)]e^{-z\Sigma/s}\Phi(s)ds\,dz\,dy. \qquad (34)$$

This equation relates the exit distribution to the incident distribution by making use of the resolvent kernel. To write Eq. (34) in terms of the factors of W, it is necessary to write

$$\int_0^\infty \int_0^\infty e^{y\Sigma/\mu}C[\delta(y-z) + \gamma(y,z)]e^{-z\Sigma/s}dz\,dy \qquad (35)$$

in terms of the factors of W. This will be accomplished in two parts. First we have the following lemma.

Lemma 1:

$$\int_0^\infty \int_0^\infty e^{y/\mu}C[\delta(y-z) + \gamma(y,z)]e^{-z\Sigma/s}dz\,dy$$
$$= H_1(-\mu)[[s\mu/(\mu-s)]H_r(s) - s(\mu/(\mu-s))^2(H_r(s) + (\mu-s)H_r'(s))M], \qquad (36)$$

where

$$W^{-1}(1/i\mu) = H_l(-\mu)H_r(\mu)$$

is a canonical factorization with H_l and H_r analytic in the open right half-plane and continuous and invertible in the closed right half-plane.

Proof: Let $G(x;s)$ be a solution to the matrix Wiener–Hopf equation

$$G(x;s) = \int_0^\infty K(|x-y|)G(y;s)dy + e^{-x\Sigma/s}. \qquad (37)$$

In this equation the variable s is considered to be a parameter. Note that the left-hand side of Eq. (36) is

$$\int_0^\infty e^{y/\mu}G(y;s)dy = \hat{G}^+(\mu;s). \qquad (38)$$

If Eq. (37) is extended to the entire real line in the usual way and the Laplace transform is defined by

$$\tilde{G}(\lambda) = \int_{-\infty}^\infty dx\, e^{x/\lambda}G(x), \quad \mathrm{Re}(\lambda)=0, \qquad (39)$$

while $Z(\lambda) = W(1/i\lambda)$, then the Laplace transform of the integral equation is

$$Z(\lambda)\tilde{G}^+(\lambda) + \tilde{G}^-(\lambda) = [s\lambda/(\lambda-s)]I - s(\lambda/(\lambda-s))^2M. \qquad (40)$$

The functions G^+ and G^- have already been defined by Eqs. (40a)–(40d), and the matrix M was introduced in Eq. (45a). Now assume that the factorization of $Z(\lambda)$ is given by

$$Z^{-1}(\mu) = H_l(-\mu)H_r(\mu),$$

where the functions H_l and H_r are analytic and invertible on the open right half-plane, and continuous and invertible on the closed right half-plane. Using the above factorization, Eq. (40) may be rewritten as

$$H_l^{-1}(-\mu)\tilde{G}^+(\mu) + H_r(\mu)\tilde{G}^-(\mu) = H_r(\mu)[[s\mu/(\mu-s)]I - s(\mu/(\mu-s))^2M]. \qquad (41)$$

If the right-hand side of Eq. (41) can be written as the sum of two terms, one analytic and invertible in the right half-plane, the other one analytic and invertible in the left half-plane, then Liouville's theorem can be invoked to solve for $\tilde{G}^+$ and $\tilde{G}^-$. Due to the second-order pole in Eq. (41), it is necessary to introduce the first derivatives of the H-functions into this splitting. By inspection, the splitting is given by the sum of

$$[s\mu/(\mu-s)][H_r(\mu) - H_l(s)] - s(\mu/(\mu-s))^2[H_r(\mu) - H_r(s) - (\mu-s)H_r'(s)]M, \qquad (42a)$$

which is analytic in the right half-plane, and the expression

$$\frac{s\mu}{\mu-s}H_r(s) - s\left(\frac{\mu}{\mu-s}\right)^2[H_r(s) + (\mu-s)H_r'(s)]M, \qquad (42b)$$

which is analytic in the left half-plane. An application of Liouville's theorem then proves Lemma 1. Note that, for $M=0$, Eq. (42b) reduces to the result given in Ref. 9. Using Lemma 1 it is now possible to write Eq. (35) in terms of the H-functions. To do this it is expedient to define

$$\Gamma(\mu,s) = \text{right-hand side of Eq. (36).} \qquad (43)$$

Now substitute the explicit formula for $\exp(-y\Sigma/\mu)$ into Eq. (35). The result is

$$\int_0^\infty \int_0^\infty e^{x/\mu}\left[I + \frac{x}{\mu}M\right]C[\delta(x-z) + \gamma(x,z)]e^{-z\Sigma/\mu}dz\,dx. \qquad (44)$$

The contribution due to the term $e^{x/\mu}$ gives $C\Gamma(\mu,s)$, while the term $(x/\mu)e^{x/\mu}$ gives rise to first derivatives of the function Γ. It is easily checked that

$$\int_0^\infty \int_0^\infty \frac{x}{\mu} e^{x/\mu} C\,[\delta(x-z)+\gamma(x,z)]e^{-z\Sigma/\mu}\,dz\,dx$$
$$= \mu C\,\partial_\mu \Gamma(\mu,s). \tag{45}$$

Therefore,

$$F(0,\mu) = -\frac{1}{2\mu}\,[I-\mu\,\partial_\mu M\,]C\int_0^1 \Gamma(\mu,s)\Phi(s)ds. \tag{46}$$

It is routine to generalize the exit distribution formula [Eq. (46)] to the N-group problem. If $\Sigma = D + M$ is the Jordan decomposition of Σ with D the diagonal matrix given by $\text{diag}\{\sigma_i\}_{i=1}^N$, then the right-hand side of Eq. (40) is replaced by

$$\sum_{m=0}^{N-1}(-1)^m \text{diag}\left\{s\left(\frac{\mu}{\mu\sigma_i - s}\right)^{m+1}\right\}_{i=1}^N M^m. \tag{47}$$

Now it is necessary to write

$$H_r(\mu)\sum_{m=0}^{N-1}(-1)^m\,\text{diag}\left\{s\left(\frac{\mu}{\mu\sigma_i - s}\right)^{2}\right\}_{i=1}^N M^m \tag{48}$$

as the sum of two terms, just as was done for the two-group case. Note that Eq. (48) has poles at $\mu = s/\sigma_i$, which are in the right half-plane. Denoting the ith column of a matrix A by $[A\,]_{(i)}$ and noting that

$$\left[H_r(\mu)\text{diag}\left\{s\left(\frac{\mu}{\mu\sigma_i - s}\right)^{m+1}\right\}\right]_{(i)}$$
$$= [H_r(\mu)]_{(i)}s(\mu/(\mu\sigma_i - s))^{m+1}, \tag{49}$$

Eq. (48) can easily be written as the sum of two terms, one analytic in the right half-plane, the other one analytic in the left half-plane. This is accomplished by writing Eq. (49) as the sum of

$$[Q_n^+(\mu,s)]$$
$$= \left[H_r(\mu) - \sum_{m=0}^{N-1}\frac{1}{m!}\left(\mu\sigma_i - \frac{\sigma_i}{s}\right)^m H_r^{(m)}\left(\frac{s}{\sigma_i}\right)\right]_{(i)}$$
$$\times s(\mu/(\mu\sigma_i - s))^{m+1}, \tag{50}$$

which is analytic in the right half-plane, and

$$[Q_n^-(\mu,s)]_{(i)} = \left[\sum_{m=0}^{N-1}\frac{1}{m!}\left(\mu\sigma_i - \frac{\sigma_i}{s}\right)^m H_r^{(m)}\left(\frac{s}{\sigma_i}\right)\right]_{(i)}$$
$$\times s(\mu/(\mu\sigma_i - s))^{m+1}, \tag{51}$$

which is analytic in the left half-plane, where $H_r^{(m)}$ and $H_l^{(m)}$ are the mth derivatives of H_r and H_l, respectively. Therefore the generalization of Lemma 1 to the N-group problem is

$$\Gamma(\mu,s) = H_l(-\mu)\sum_{m=0}^{N-1} Q_m^-(\mu,s)M^m, \tag{52}$$

and the exit distribution $[\,F(0,\mu)]_{(i)}$ is given by

$$-\frac{1}{2\mu}\sum_{m=0}^{N-1}\frac{1}{m!}\left(\frac{\mu}{\sigma_i}\right)^m(\partial_\mu)^m$$
$$\times\int_0^1\left[M^m\Gamma\left(\frac{\mu}{\sigma_i},s\right)\right]_{(i)}\Phi(s)\,ds. \tag{53}$$

Not only can the exit distribution be written in terms of the factors of the symbol, but the solution for any value of x can also be written in a similar fashion. This can be done by making use of Eq. (30b), which relates $F(x,\mu)$ to $G(x)$, and the results of this section. First we note that

$$\hat{G}^+(\mu) = \int_0^1 \Gamma(\mu,s)\Phi(s)\,ds. \tag{54}$$

From this expression it is possible to recover the function G. Now that G is known, the solution $F(x,\mu)$ for $x<0$ can be computed by making use of Eq. (30b).

VII. CONCLUSION

Formulas for the exit distribution and the solution to a multigroup transport equation with a nondiagonal cross-section matrix have been derived in terms of generalized Chandrasekhar H-functions. For the special case of two groups with a noninvertible scattering matrix, the H-functions were computed explicitly. Unfortunately, for $N>2$ we are not able to construct the factorization explicitly, so we are forced to derive a nonlinear integral equation which the H-functions satisfy and to set up a numerical scheme for solving them. This work will be published elsewhere.

ACKNOWLEDGMENTS

This work is part of a Ph.D. dissertation accepted by the faculty of Virginia Polytechnic Institute and State University by Barton L. Willis, who would like to thank his advisor P. F. Zweifel for suggesting this problem and for helpful discussions.

This work was supported by the U.S. Department of Energy under Grant No. DE-AS05-80-ER10711 and the National Science Foundation under Grants No. DMS-8312451 and No. DMS-8501337.

[1] W. M. Stacey, *Model Approximations* (M.I.T., Cambridge, MA, 1967).
[2] J. J. Duderstadt and W. R. Martin, *Transport Theory* (Wiley, New York, 1979), p. 408.
[3] W. F. Ames, *Nonlinear Partial Differential Equations in Engineering* (Academic, New York, 1972), Vol. 2, pp.146–180.
[4] C. E. Siewert and P. F. Zweifel, "An exact solution of equations of radiative transfer for local thermodynamic equilibrium in the non-gray case. Picket fence approximation," Ann. Phys. NY **36**, 61 (1966).
[5] K. M. Case and P. F. Zweifel, *Linear Transport Theory* (Addison–Wesley, Reading, MA, 1967).
[6] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).
[7] I. C. Gohberg and M. G. Krein, "Systems of Integral Equations on a Half Line with Kernel Depending on the Difference of Arguments," Am. Math. Soc. Trans. **14**, 217 (1960).
[8] C. N. Cebotarev, "Partial indices for the Riemann boundary problem for a matrix of second order," Usp. Mat. Nauk. **11**, 144 (1956).
[9] E. W. Larsen and P. F. Zweifel, "Explicit Wiener–Hopf factorizations," SIAM J. Appl. Math. **30**, 732 (1976).
[10] C. V. M. van der Mee, "Albedo operators and *H*-equations for generalized kinetic models," Transport Theory Stat. Phys. **13**, 341 (1984).

Multispinor symmetries for massless arbitrary spin Fierz–Pauli and Rarita–Schwinger wave equations[a)]

Noel A. Doughty and Graham P. Collins[b)]
Physics Department, University of Canterbury, Christchurch, New Zealand

(Received 13 November 1985; accepted for publication 30 January 1986)

Massless, $D(j,0) \oplus D(0,j)$, multispinor fields of arbitrary unmixed spin j are reduced by simple matrix-algebra methods to associated tensors and tensor–spinors. A generalized Majorana condition applied to the multispinors is seen to correspond to reality and Majorana conditions on the associated tensors and tensor–spinors, respectively. The symmetries of the latter are displayed explicitly for arbitrary spin. For spin-1, -$\frac{3}{2}$, -2, and -$\frac{5}{2}$ the free-field gauge-invariant Lagrangian wave equations of Maxwell (spin-1), Rarita–Schwinger (spin-$\frac{3}{2}$), Fierz–Pauli (spin-2), and spin-$\frac{5}{2}$ are derived directly and in a uniform manner from the simpler equations of the unmixed spin reps strongly suggesting the method is extendable to arbitrary spin. Similar features of massive fields are briefly reviewed.

I. INTRODUCTION

The only irreps of unmixed spin of the full Lorentz group $O_{1,3}$ are $D(0,0)$, the trivial scalar rep, and $D(j,0) \oplus D(0,j)$, where $D(j_1,j_2)$ are the irreps of the restricted (inversion-free) Lorentz group $SO^+_{1,3}$. In a previous paper,[2] symmetries of Weyl field tensors and tensor–spinors for arbitrary spin were considered using primarily matrix methods. We consider here the arbitrary spin case from the point of view of Dirac spinors, again using matrix methods rather than explicitly indexed spinors.

In Secs. II and III we set out the details of the $2(2j+1)$-component spinors that form the starting point of our analysis and describe the generalization of charge conjugation to arbitrary spin. In Sec. IV we establish the relations between the multispinors of spin-1, -$\frac{3}{2}$, -2, and -$\frac{5}{2}$ and the corresponding tensors or tensor–spinors. From each we derive the corresponding Lagrangian potential. We discuss the arbitrary spin case in Sec. V and the massive spin results briefly in Sec. VI.

II. ARBITRARY SPIN SPINORS AND WAVE EQUATIONS

The $D(\frac{1}{2},0) \oplus D(0,\frac{1}{2})$ rep space is the space of four-component Dirac spinors $\Psi = (\Psi^\alpha)$ $(\alpha = 1,2,3,4)$. Under a Lorentz transformation, Ψ transforms according to $\Psi \rightarrow S(\Lambda)\Psi$, where $S = \exp(-\frac{1}{4}i\omega_{\mu\nu}\gamma^{\mu\nu})$ and the Lorentz transformation Λ is parametrized by $\omega_{\mu\nu}$. (The infinitesimal generators $\gamma^{\mu\nu}$ are defined in the Appendix, which fixes our conventions and lists Dirac algebra identities.) The Dirac adjoint spinor $\overline{\Psi} = \Psi^\dagger\gamma_0$ transforms as $\overline{\Psi} \rightarrow \overline{\Psi}S^{-1}$. Note that in this role, γ_0 actually transforms according to $\gamma_0 \rightarrow S^{-1\dagger}\gamma_0 S^{-1} = \gamma_0$ consistently with $\gamma_\mu \rightarrow \Lambda^\nu{}_\mu S\gamma_\nu S^{-1} = \gamma_\mu$.

The chiral parts of a Dirac spinor are $\Psi_\pm = \frac{1}{2}(1 \pm \gamma_5)\Psi$ and the sum of these reconstructs the whole spinor. In the chiral or Weyl rep, the chiral parts may be regarded as (two-component) Weyl spinors ($\overline{\phi}$ and χ) of opposite chirality

$$\Psi = \begin{pmatrix} \overline{\phi}_{\dot{U}} \\ \chi^A \end{pmatrix}. \tag{2.1}$$

For arbitrary higher spin j, the corresponding $O_{1,3}$ irreducible $D(j,0) \oplus D(0,j)$ rep space is comprised of multiply indexed Dirac spinors Ψ, symmetric on all of their $2j$ indices (which we suppress) and with totally symmetric chiral parts $\Psi_\pm = \frac{1}{2}(1 \pm \gamma_5)\Psi$. Consequently γ_5 can act on any index of Ψ with equivalent results. In the chiral rep, such a spinor is the direct sum of two (totally symmetric) opposite chirality Weyl spinors. This is in contrast to the multispinors used in the formalism of Bargmann and Wigner,[3] where the rep space is $\otimes_1^{2j}(D(\frac{1}{2},0) \oplus D(0,\frac{1}{2}))$ ($O_{1,3}$ reducible for $j > \frac{1}{2}$) and the spinors are merely required to be symmetric. Since the $2(2j+1)$-dimensional space $D(j,0) \oplus D(0,j)$ is a subspace of the $(2j+3)!/3!2j!$-dimensional space $\otimes_1^{2j}(D(\frac{1}{2},0) \oplus D(0,\frac{1}{2}))$ it is possible to handle an unmixed spin multispinor as a Bargmann–Wigner spinor that happens to have block-diagonal form in the chiral rep. The transformation law for both types of spinors can therefore be written $\Psi \rightarrow (S \otimes S \otimes \cdots \otimes S)\Psi$, where there is one transformation matrix S for each of the $2j$ indices of Ψ. By indexing and rearranging, we see that any number of the transformation matrices may be transferred to the right-hand side of Ψ provided they are then transposed $(S \rightarrow S^T)$, namely $\Psi \rightarrow (S \otimes S \otimes \cdots \otimes S)\Psi(S^T \otimes \cdots \otimes S^T)$.

The adjoint for a spinor of spin j is $\overline{\Psi} = \Psi^\dagger(\gamma_0 \otimes \gamma_0 \otimes \cdots \otimes \gamma_0)$, where the γ_0 matrices are $2j$ in number, any of which may be transferred to the left-hand side by transposing. The complete contraction of $\overline{\Psi}$ with Ψ is a Lorentz scalar.

For massless fields, the appropriate free-field wave equations for the $D(j,0)$ and $D(0,j)$ parts of Ψ are generalized Weyl equations.[2] In the Bargmann–Wigner spinor space these can be written

$$\eth\Psi = 0, \tag{2.2}$$

where the γ matrix of $\eth = \gamma \cdot \partial$ may act on any of the $2j$ indices of Ψ. Mass irreducibility with zero mass follows by acting on (2.2) with $\eth$ to yield $\Box\Psi = 0$. The form (2.2) is in

[a)] This work contains part of the M.Sc. thesis of Collins[1] presented to the Department of Physics, Canterbury University, 1985.

[b)] Present address: Department of Physics, State University of New York at Stony Brook, Stony Brook, New York 11794.

0022-2488/86/061639-07$02.50

fact just the massless limit of the Bargmann–Wigner equations,

$$(i\partial - m)\Psi = 0, \tag{2.3}$$

which, for spin-$\frac{1}{2}$, is simply the Dirac equation. For arbitrary spin and nonzero mass, Eq. (2.3) may be derived from a $2j$th-order equation, which projects out the highest spin of Ψ.[4] For spin-$\frac{1}{2}$, the analogous equation for massive $2(2j+1)$-component spinors is again the Dirac equation (2.3). However, for massive spin $> \frac{1}{2}$ there is no simple first-order form like (2.3). Naively applying (2.3) to a massive $2(2j+1)$-component spinor ($j > \frac{1}{2}$) annihilates the spinor.

III. CHARGE CONJUGATION OF AN ARBITRARY SPIN MULTISPINOR

As usual, the charge conjugate of a spin-$\frac{1}{2}$ spinor Ψ is written

$$\Psi^c = C\overline{\Psi}^T = C\gamma_0^T\Psi^*, \tag{3.1}$$

where C is the charge conjugation matrix. For free fields we require Ψ^c to satisfy the same equation as Ψ [namely (2.2) or (2.3)]. This condition implies that positive-energy solutions of the equation of motion are carried into negative-energy solutions under charge conjugation and vice versa. (Furthermore, if Ψ is minimally coupled to the Maxwell or Yang–Mills fields, then Ψ^c will be minimally coupled but with opposite charge.) The above requirement implies that C satisfies

$$C\gamma_\mu^T C^{-1} = a\gamma_\mu \quad (m=0), \tag{3.2}$$

$$C\gamma_\mu^T C^{-1} = -\gamma_\mu \quad (m \neq 0). \tag{3.3}$$

The associated requirement that Ψ^c transform under Lorentz transformations as a Dirac spinor gives

$$C\gamma_{\mu\nu}^T C^{-1} = -\gamma_{\mu\nu}, \tag{3.4}$$

which is consistent with (3.3) and implies that $a = \pm 1$ in the massless case (3.2). Also, C allows nontrivial self-conjugate massless fields if and only if $a = -1$ in (3.2).[5] It is therefore natural to impose this condition, making the massive and massless charge conjugation matrices identical. Application of Schur's lemma[6,7] permits us to establish[5] the rep-independent properties of the charge conjugation matrix C to be

$$C\gamma_{\mu\nu}^T C^{-1} = -\gamma_{\mu\nu}, \quad C\gamma_\mu^T C^{-1} = -\gamma_\mu,$$
$$C^T = -C, \quad C^\dagger C = 1, \quad C^* C = -1. \tag{3.5}$$

In addition, in "friendly" representations, where each of the γ_μ is either symmetric or antisymmetric, the phase of C can be selected so that

$$C^{-1} = -C \Rightarrow C^* = C. \tag{3.6}$$

Although Eq. (3.6) holds in many representations (for example, the Weyl or chiral rep and the Dirac and the Majorana reps[8]), because they are not totally representation independent, no results of physical importance should depend on their use.

We define a Majorana spinor to be a charge self-conjugate spinor

$$\Psi^c = b\Psi, \tag{3.7}$$

where b is an arbitrary phase factor ($|b| = 1$). By (A23) this is a rep-independent concept. Here we will always select $b = 1$. (In a Majorana representation $C\gamma_0^T = i$ could be taken so that selecting $b = i$ would be equivalent to requiring that the components of Ψ be real in such a representation.) Clearly a Majorana Dirac spinor has four real or two complex components. A Majorana spinor may be split into chiral parts by the operators $\frac{1}{2}(1 \pm \gamma_5)$, but each part still has two complex components and Majorana conjugation links the two chiralities

$$(\Psi_\pm)^c = (\Psi^c)_\mp, \quad \text{or} \quad \Psi^c = \Psi \Leftrightarrow (\Psi_\pm)^c = \Psi_\mp. \tag{3.8}$$

We note that this implies that C has (2×2) block-diagonal form in the chiral representation. For a spinor of spin j, the natural generalization of (3.1) has $2j$ factors of C:

$$\Psi^c = (C \otimes C \otimes \cdots \otimes C)\overline{\Psi}^T$$
$$= (C \otimes \cdots \otimes C)\overline{\Psi}^T(C^T \otimes \cdots \otimes C^T). \tag{3.9}$$

As for spin-$\frac{1}{2}$, the charge conjugate will obey the same equation as Ψ, but with opposite charge when minimally coupled electromagnetically or via Yang–Mills charges. A single Majorana spinor of arbitrary spin can have no gauge-invariant couplings to a spin-1 boson and in this sense is neutral.

IV. FIELDS STRENGTHS AND LAGRANGIAN WAVE EQUATIONS FOR SPIN-1, -$\frac{3}{2}$, -2, AND -$\frac{5}{2}$

We will first review the well-known relationships between the unmixed and mixed irreps of massless spin-1 to assist in establishing a pattern for the higher spin cases.

A. Helicity 1: Maxwell field

From the spin-1 spinor Ψ we may construct a corresponding complex field strength tensor by defining

$$F_{\mu\nu} = \frac{1}{4}\,\mathrm{Tr}(\Psi C^{-1}\gamma_{\mu\nu}). \tag{4.1}$$

Because of the symmetry of Ψ, it is not necessary to specify which of the two indices of Ψ are contracted in the matrix multiplication with $C^{-1}\gamma_{\mu\nu}$ and which is traced with the remaining free matrix index. It is also readily verified that (4.1) is representation-independent and $F_{[\mu\nu]} = F_{\mu\nu}$ follows from the properties of $\gamma_{\mu\nu}$. Hence $F_{\mu\nu}$ has six complex components, the same number as Ψ. Now ΨC^{-1} has 2×2 block-diagonal form in the chiral rep and is traceless because of the symmetry of Ψ and the antisymmetry of C^{-1}. Hence using identity (A17) the spinor Ψ can be recovered from $F_{\mu\nu}$ via

$$\Psi = \frac{1}{2}F_{\mu\nu}\gamma^{\mu\nu}C. \tag{4.2}$$

Given any antisymmetric $F_{\mu\nu}$, (4.2) yields a $D(1,0) \oplus D(0,1)$ spinor, from which $F_{\mu\nu}$ may be recovered via Eq. (4.1) due to identity (A15). Hence (4.1) and (4.2) are reciprocal and define an isomorphism between the spaces of $D(1,0) \oplus D(0,1)$ spinors and antisymmetric second-order tensors. That is, $F_{\mu\nu}$ and Ψ contain the same information.

One can define chiral field strengths $F_{\mu\nu}^\pm$ using $\Psi_\pm$ in place of Ψ in Eq. (4.1). These can be shown to be self-dual ($i\widetilde{F}_{\mu\nu}^+ = F_{\mu\nu}^+$) and anti-self-dual ($i\widetilde{F}_{\mu\nu}^- = -F_{\mu\nu}^-$) from the duality property (A7) of $\gamma_{\mu\nu}$. Any antisymmetric second-order tensor can be split into the sum of a self-dual and an anti-self-dual part, and this could serve as an alternative definition of the chiral parts of $F_{\mu\nu}$. Equations (4.1) and (4.2) also provide isomorphisms between the respective chiral parts of $F_{\mu\nu}$ and Ψ.

Using relations (3.5) and (A5) to evaluate $F^*_{\mu\nu}$ gives

$$F^*_{\mu\nu} = \tfrac{1}{4}\,\mathrm{Tr}(\gamma_{\mu\nu} C^T \overline{\Psi}), \tag{4.3}$$

demonstrating a close relation between the Dirac adjoint at the spinorial level and the complex conjugate at the tensorial level. Indeed, Eq. (4.3) could be taken as the definition of an isomorphism between Dirac adjoint spinors and tensorial field strengths analogous to (4.1). Equation (4.3) can also be put in the form

$$F^*_{\mu\nu} = \tfrac{1}{4}\,\mathrm{Tr}(\Psi^c C^{-1} \gamma_{\mu\nu}), \tag{4.4}$$

and clearly $F_{\mu\nu}$ will be real if and only if it is derived from a Majorana spinor. Note that, by (3.8), a real $F_{\mu\nu}$ cannot be separated into real chiral parts. The tracelessness of ΨC^{-1} and the wave equation (2.2) imply

$$\partial^{\mu} F_{\mu\nu} = 0. \tag{4.5}$$

Equation (4.5) also holds for the chiral parts of $F_{\mu\nu}$ and hence for the dual of $F_{\mu\nu}$. The dual equation acts as an integrability condition for $F_{\mu\nu}$ so the Poincaré lemma[9] implies that a field A_μ (which may be the Hermitian electromagnetic potential if Ψ is Majorana) exists such that

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{4.6}$$

Substituting into (4.5) now gives the usual gauge-invariant Maxwell equation

$$\Box A_\mu - \partial_\mu \partial\cdot A = 0, \tag{4.7}$$

with an identically divergence-free left-hand side. This latter condition is, of course, a necessary condition (see, for example, Doughty and Wiltshire[2]) for the equation to be derivable from a Lagrangian. The usual helicity-1 gauge-invariant Hermitian Lagrangian can then be defined using this potential A_μ as the coordinate field.

B. Helicity $\frac{3}{2}$: Rarita–Schwinger field

Analogous to (4.1), define a complex field-strength tensor–spinor $f_{\mu\nu}$ from the spin-$\frac{3}{2}$ spinor Ψ:

$$f_{\mu\nu} = \tfrac{1}{4}\,\mathrm{Tr}(\Psi C^{-1} \gamma_{\mu\nu}). \tag{4.8}$$

As before, the symmetry of Ψ makes it unnecessary to specify which of its indices are used for matrix multiplication or tracing, and which is left free to act as the (suppressed) spinorial index of $f_{\mu\nu}$. Clearly $f_{\mu\nu}$ is antisymmetric:

$$f_{[\mu\nu]} = f_{\mu\nu}. \tag{4.9}$$

The chiral parts of $f_{\mu\nu}$ can be defined by applying (4.8) to the chiral parts of Ψ. By applying the chiral projector to that free index of Ψ, which becomes the spinorial index of $f_{\mu\nu}$, it is clear that the chiral parts of $f_{\mu\nu}$ can also be written

$$f^{\pm}_{\mu\nu} = \tfrac{1}{2}(1 \pm \gamma_5) f_{\mu\nu}. \tag{4.10}$$

As for helicity 1 these chiral parts are (anti-)self-dual:

$$i\tilde{f}^{\pm}_{\mu\nu} = \pm f^{\pm}_{\mu\nu}. \tag{4.11}$$

Hence the full tensor–spinor has the duality property

$$i\tilde{f}_{\mu\nu} = \gamma_5 f_{\mu\nu}. \tag{4.12}$$

In the chiral representation, for each value of the index of Ψ that was left free in (4.8), there is a symmetric block-diagonal matrix. Hence, as with the helicity-1 case, Ψ can be recovered from $f_{\mu\nu}$ by the use of (A17):

$$\Psi = \tfrac{1}{2} f_{\mu\nu} \otimes \gamma^{\mu\nu} C. \tag{4.13}$$

From this it is readily shown that

$$\gamma^{\mu\nu} f_{\mu\nu} = 0. \tag{4.14}$$

These symmetries (4.9), (4.12), and (4.14) reduce $f_{\mu\nu}$ to just eight complex components, the same number as Ψ. It is therefore clear that Eq. (4.13) is the inverse of Eq. (4.8) so that these equations define an isomorphism between the spaces of spin-$\frac{3}{2}$ multispinors and tensor–spinors. From the above symmetries we establish the dependent symmetry

$$\gamma^{\mu} f_{\mu\nu} = 0. \tag{4.15}$$

This symmetry combines with (4.9) to form a complete set of symmetries for $f_{\mu\nu}$ equivalent to the set (4.9), (4.12), and (4.14).[2,10] For helicity $\frac{3}{2}$, the Dirac adjoint of Ψ is related to the Dirac adjoint of $f_{\mu\nu}$ analogously with Eq. (4.3):

$$\bar{f}_{\mu\nu} = \tfrac{1}{4}\,\mathrm{Tr}(\gamma_{\mu\nu} C^T \overline{\Psi}). \tag{4.16}$$

Similarly, the Majorana conjugate of Ψ is related to $f_{\mu\nu}$ by

$$C\bar{f}^{T}_{\mu\nu} = \tfrac{1}{4}\,\mathrm{Tr}(\Psi^c C^{-1} \gamma_{\mu\nu}). \tag{4.17}$$

Consequently, applying the generalized Majorana condition to a multispinor Ψ is equivalent to simply applying the usual Majorana condition to the spinor part of the associated tensor–spinor. As for helicity 1, Eq. (2.2) implies

$$\partial^{\mu} f_{\mu\nu} = 0, \quad \partial^{\mu} \tilde{f}_{\mu\nu} = 0. \tag{4.18}$$

Hence the Poincaré lemma ensures the existence of a potential Ψ_μ (which will be a Majorana vector–spinor if the original multispinor Ψ, and hence $f_{\mu\nu}$, were Majorana), which satisfies

$$f_{\mu\nu} = \partial_\mu \Psi_\nu - \partial_\nu \Psi_\mu. \tag{4.19}$$

However, the field equation for Ψ_μ does not follow from the first equation of (4.18), which yields a second-order differential equation for Ψ_μ, but from the symmetry (4.15), which directly yields the simplest form of the gauge-invariant massless spin-$\frac{3}{2}$ Rarita–Schwinger equation,

$$\not{\partial}\Psi_\mu - \partial_\mu \gamma\cdot\Psi = 0. \tag{4.20}$$

The second-order equation can also be derived from (4.20) and constitutes a differential condition on the Ψ_μ field. Equation (4.20) is invariant[11] under the gauge transformation $\delta\Psi_\mu = \partial_\mu \epsilon$ (for an arbitrary spin-$\frac{1}{2}$ field ϵ) but its left-hand side is not divergence-free off-shell and the equation cannot therefore be derived from a Hermitian Lagrangian, nor coupled directly to a conserved vector–spinor source. However, by adding a multiple of its spinorial trace, a divergence-free form is readily obtained,

$$\not{\partial}\Psi_\mu - \partial_\mu \gamma\cdot\Psi - \gamma_\mu \partial\cdot\Psi + \gamma_\mu \not{\partial}\gamma\cdot\Psi = 0, \tag{4.21}$$

and this equation is equivalent to (4.20) except that it is derivable from a Lagrangian. By using the Dirac identity (A8), this equation can also be written compactly as

$$\epsilon^{\mu\nu\lambda\rho} \gamma_5 \gamma_\nu \partial_\lambda \Psi_\rho = 0. \tag{4.22}$$

For both of these, the Hermitian gauge-invariant Lagrangian may be taken to be

$$L = -\tfrac{1}{2}\overline{\Psi}_\mu \epsilon^{\mu\nu\lambda\rho}\gamma_5\gamma_\nu \overleftrightarrow{\partial}_\lambda \Psi_\rho . \tag{4.23}$$

C. Helicity 2: Fierz–Pauli field

We form the complex tensor

$$C_{\mu\nu\lambda\rho} = \tfrac{1}{16}\,\mathrm{Tr}(\mathrm{Tr}(\Psi C^{-1}\gamma_{\mu\nu})C^{-1}\gamma_{\lambda\rho}), \tag{4.24}$$

with the inverse relation following from a double application of (A17):

$$\Psi = \tfrac{1}{4}C_{\mu\nu\lambda\rho}(\gamma^{\mu\nu}C \otimes \gamma^{\lambda\rho}C). \tag{4.25}$$

The symmetries

$$C_{[\mu\nu]\lambda\rho} = C_{\mu\nu[\lambda\rho]} = C_{\mu\nu\lambda\rho} = C_{\lambda\rho\mu\nu} \tag{4.26}$$

leave $C_{\mu\nu\lambda\rho}$ with 21 complex components. The independent symmetry

$$C_{[\mu\nu\lambda\rho]} = 0 \tag{4.27}$$

eliminates another component. The ten further relations needed for $C_{\mu\nu\lambda\rho}$ to have the same number of components as Ψ are supplied by the tracelessness relation[10,12]

$$C^{\lambda}{}_{\mu\lambda\nu} = 0. \tag{4.28}$$

Two dependent relations are

$$C_{\widetilde{\mu\nu}\lambda\rho} = C_{\mu\nu\widetilde{\lambda\rho}}, \tag{4.29}$$

$$C_{\mu[\nu\lambda\rho]} = 0. \tag{4.30}$$

As with helicity 1, the spinorial chiral parts of Ψ can be used to define tensorial chiral parts $C^{\pm}_{\mu\nu\lambda\rho}$. These can be shown to be (anti-)self-dual on both their first and second pair of indices. Since any fourth-order tensor obeying the symmetries (4.26)–(4.28) [and hence (4.29)] can be split into an anti-self-dual and a self-dual part, the chiral parts of $C_{\mu\nu\lambda\rho}$ could equivalently be defined by this property.

By double application of the same procedure as used for helicity 1, an analogous relation to (4.3) between the Dirac adjoint of Ψ and the complex conjugate of the tensor can be established.

$$C^{*}_{\mu\nu\lambda\rho} = \tfrac{1}{16}\,\mathrm{Tr}(\mathrm{Tr}(\gamma_{\mu\nu}C^{T}(\gamma_{\lambda\rho}C^{T}\overline{\Psi}))). \tag{4.31}$$

Double application of Eq. (4.4) also gives the relation between the complex conjugate of the tensor $C_{\mu\nu\lambda\rho}$ and the charge conjugate of Ψ:

$$C^{*}_{\mu\nu\lambda\rho} = \tfrac{1}{16}\,\mathrm{Tr}(\mathrm{Tr}(\Psi^{c}C^{-1}\gamma_{\mu\nu})C^{-1}\gamma_{\lambda\rho}). \tag{4.32}$$

Hence, as before, a Majorana spinor will give rise to a real tensor.

As for the lower spins, Eq. (2.2) leads to an integrability condition for $C_{\mu\nu\lambda\rho}$,

$$C_{\mu\nu[\lambda\rho,k]} = 0, \tag{4.33}$$

where the comma is used as a convenient alternative notation for the partial derivative. Following Pirani[13] we use condition (4.33), the Poincaré lemma, and (4.26) to obtain a tensor $Q_{\mu\nu\lambda}$ such that

$$C_{\mu\nu\lambda\rho} = Q_{\mu\nu[\lambda,\rho]}. \tag{4.34}$$

We may vary $Q_{\mu\nu\lambda}$ by an arbitrary divergence without altering $C_{\mu\nu\lambda\rho}$:

$$\delta Q_{\mu\nu\lambda} = P_{\mu\nu,\lambda}. \tag{4.35}$$

Symmetry (4.30) implies an integrability condition for $Q_{\mu\nu\lambda}$,

$$Q_{\mu[\nu\lambda,\rho]} = 0, \tag{4.36}$$

which in turn implies the existence of a tensor $N_{\mu\nu}$, such that

$$Q_{\mu[\nu\lambda]} = N_{\mu[\nu,\lambda]}. \tag{4.37}$$

The freedom (4.35) allows $N_{\mu\nu,\lambda}$ to be subtracted from $Q_{\mu\nu\lambda}$ so that we may take

$$Q_{\mu\nu\lambda} = Q_{\mu(\nu\lambda)}. \tag{4.38}$$

Now

$$0 = C_{(\mu\nu)\lambda\rho} = Q_{(\mu\nu)[\lambda,\rho]} \tag{4.39}$$

implies the existence of a symmetric tensor $h_{\mu\nu}$, such that

$$Q_{(\mu\nu)\lambda} = h_{\mu\nu,\lambda}. \tag{4.40}$$

Symmetry (4.38) allows $Q_{\mu\nu\lambda}$ to be expanded according to

$$\begin{aligned} Q_{\mu\nu\lambda} &= Q_{(\mu\nu)\lambda} + Q_{(\mu\lambda)\nu} - Q_{(\nu\lambda)\mu} \\ &= h_{\mu\nu,\lambda} + h_{\mu\lambda,\nu} - h_{\nu\lambda,\mu}, \end{aligned} \tag{4.41}$$

yielding

$$C_{\mu\nu\lambda\rho} = -2\partial_{[\mu}h_{\nu][\lambda,\rho]}. \tag{4.42}$$

A real tensor $C_{\mu\nu\lambda\rho}$ (arising from a Majorana multispinor Ψ) corresponds, of course, to a real potential $h_{\mu\nu}$.

The second-order massless Fierz–Pauli field equation[14] follows from substitution of (4.42) into the symmetry (4.28):

$$\Box h_{\mu\nu} - 2\partial^{\lambda}\partial_{(\mu}h_{\nu)\lambda} + \partial_\mu\partial_\nu h = 0, \tag{4.43}$$

where $h = h^{\mu}{}_{\mu}$. As with the helicity-$\frac{3}{2}$ case, the field equations for Ψ and $C_{\mu\nu\lambda\rho}$ constitute higher-order differential conditions on (4.43). Furthermore, (4.43) is gauge-invariant under $\delta h_{\mu\nu} = \partial_\mu\xi_\nu + \partial_\nu\xi_\mu$ (where ξ_μ is an arbitrary vector field), but the left-hand side is not divergence-free off-shell, and hence not derivable from a Hermitian Lagrangian. Again an identically divergence-free equation is obtained by adding to (4.43) a multiple of its own trace giving

$$\begin{aligned} &\Box h_{\mu\nu} - 2\partial^{\lambda}\partial_{(\mu}h_{\nu)\lambda} + \partial_\mu\partial_\nu h \\ &\quad - \eta_{\mu\nu}\Box h + \eta_{\mu\nu}\partial^{\lambda}\partial^{\rho}h_{\lambda\rho} = 0. \end{aligned} \tag{4.44}$$

These equations are, of course, derivable from the linearized Einstein Lagrangian.[15]

D. Helicity $\frac{5}{2}$

Analogously to helicity 2, we define a complex tensor spinor

$$f_{\mu\nu\lambda\rho} = \tfrac{1}{16}\,\mathrm{Tr}(\mathrm{Tr}(\Psi C^{-1}\gamma_{\mu\nu})C^{-1}\gamma_{\lambda\rho}), \tag{4.45}$$

with the inverse relation

$$\Psi = \tfrac{1}{4}f_{\mu\nu\lambda\rho}(\gamma^{\mu\nu}C \otimes \gamma^{\lambda\rho}C). \tag{4.46}$$

As for $C_{\mu\nu\lambda\rho}$, the tensor–spinor $f_{\mu\nu\lambda\rho}$ obeys the symmetries

$$\begin{aligned} &f_{[\mu\nu]\lambda\rho} = f_{\mu\nu[\lambda\rho]} = f_{\mu\nu\lambda\rho} = f_{\lambda\rho\mu\nu}, \\ &f_{[\mu\nu\lambda\rho]} = 0. \end{aligned} \tag{4.47}$$

In addition to these, $f_{\mu\nu\lambda\rho}$ obeys the symmetry

$$\gamma^{\mu}f_{\mu\nu\lambda\rho} = 0. \tag{4.48}$$

This implies the vanishing trace relation

$$f^{\lambda}{}_{\mu\lambda\nu} = 0. \tag{4.49}$$

Dependent relations include

$$f_{\widetilde{\mu\nu}\lambda\rho} = f_{\mu\nu\widetilde{\lambda\rho}}, \tag{4.50}$$

$$f_{\mu[\nu\lambda\rho]} = 0, \quad \gamma^{\mu\nu} f_{\mu\nu\lambda\rho} = 0. \tag{4.51}$$

As for lower spin, the spinorial chiral parts of the Ψ can be used to define tensor–spinor chiral parts $f^{\pm}_{\mu\nu\lambda\rho}$. These obey

$$f^{\pm}_{\mu\nu\lambda\rho} = \tfrac{1}{2}(1 \pm \gamma_5) f_{\mu\nu\lambda\rho}, \quad i\tilde{f}^{\pm}_{\mu\nu\lambda\rho} = \pm f^{\pm}_{\mu\nu\lambda\rho}, \tag{4.52}$$

where the dual operation may act on either the first or second pair of indices of $f_{\mu\nu\lambda\rho}$. The full tensor–spinor therefore obeys

$$i\tilde{f}_{\mu\nu\lambda\rho} = \gamma_5 f_{\mu\nu\lambda\rho}. \tag{4.53}$$

The Majorana conjugate of $f_{\mu\nu\lambda\rho}$ is related to the Majorana conjugate of Ψ

$$f^{c}_{\mu\nu\lambda\rho} = \tfrac{1}{16}\,\mathrm{Tr}(\mathrm{Tr}(\Psi^{c} C^{-1}\gamma_{\mu\nu}) C^{-1}\gamma_{\lambda\rho}). \tag{4.54}$$

Equation (2.2) leads to an integrability condition for $f_{\mu\nu\lambda\rho}$ and the Poincaré lemma may be applied as for spin-2. The only differences are that the tensors are replaced by tensor–spinors and the tensor–spinor $q_{\mu\nu\lambda}$, corresponding to $Q_{\mu\nu\lambda}$, can be chosen to satisfy

$$\gamma^{\mu} q_{\mu\nu\lambda} = 0. \tag{4.55}$$

One obtains a symmetric tensor–spinor potential $\Psi_{\mu\nu}$, such that

$$q_{\mu\nu\lambda} = \Psi_{\mu\nu,\lambda} + \Psi_{\mu\lambda,\nu} - \Psi_{\nu\lambda,\mu}, \tag{4.56}$$

$$f_{\mu\nu\lambda\rho} = -2\partial_{[\mu}\Psi_{\nu][\lambda,\rho]}. \tag{4.57}$$

The potential $\Psi_{\mu\nu}$ is a Majorana tensor–spinor if the original multispinor Ψ (and hence $f_{\mu\nu\lambda\rho}$) was Majorana.

Equation (4.55) implies the simplest form of the massless spin-$\frac{5}{2}$ field equation[16,17] namely

$$\eth\Psi_{\mu\nu} - 2\gamma^{\lambda}\partial_{(\mu}\Psi_{\nu)\lambda} = 0. \tag{4.58}$$

The $f_{\mu\nu\lambda\rho}$ field equation and (4.49) imply differential conditions for $\Psi_{\mu\nu}$, which are also derivable from the $\Psi_{\mu\nu}$ field equation (4.58). Equation (4.58) is gauge-invariant under $\delta\Psi_{\mu\nu} = \partial_{\mu}\epsilon_{\nu} + \partial_{\nu}\epsilon_{\mu}$, where ϵ_{μ} is an arbitrary γ-traceless vector–spinor field (that is, arbitrary up to satisfying $\gamma\cdot\epsilon = 0$). Equation (4.58) is equivalent to the following equation (in which $\Psi = \Psi^{\mu}{}_{\mu}$):

$$\begin{aligned}\eth\Psi_{\mu\nu} &- 2\gamma^{\lambda}\partial_{(\mu}\Psi_{\nu)\lambda} - 2\partial^{\lambda}\gamma_{(\mu}\Psi_{\nu)\lambda} + 2\gamma_{(\mu}\eth\gamma^{\lambda}\Psi_{\nu)\lambda} \\ &- \tfrac{1}{2}\eta_{\mu\nu}\eth\Psi + \eta_{\mu\nu}\partial^{\lambda}\gamma^{\rho}\Psi_{\lambda\rho} + \gamma_{(\mu}\partial_{\nu)}\Psi = 0.\end{aligned} \tag{4.59}$$

Denoting the left-hand side of (4.59) by $\chi_{\mu\nu}$, the γ-traceless part of $\chi_{\mu\nu}$ is identically divergence-free,[16,17]

$$\partial^{\mu}(\chi_{\mu\nu} - \tfrac{1}{4}\gamma_{\nu}\gamma^{\lambda}\chi_{\mu\lambda}) = 0, \tag{4.60}$$

as is necessary for the equation to be derivable from a Lagrangian. One such Lagrangian was displayed recently as part of an analysis of arbitrary spin by Berends *et al.*[16] Earlier papers with massless spin-$\frac{5}{2}$ Lagrangians include Berends *et al.*,[18–20] while Fronsdal,[21] Fang and Fronsdal,[22] and de Wit and Freedman[23] give Lagrangians for massless fields or arbitrary spin.

V. ARBITRARY HELICITY

The cases of spin-1 through -$\frac{5}{2}$ establish the pattern for the relations between spinors of unmixed spin and corresponding tensors and tensor-spinors. No new features occur for higher spins: there are simply more space-time indices in the equations. For integer spin j and half-odd-integer spin $j = n + \frac{1}{2}$, respectively, one may define

$$F_{\mu_1\nu_1\cdots\mu_j\nu_j} = (\tfrac{1}{4})^{j}\,\mathrm{Tr}(\cdots\mathrm{Tr}(\Psi C^{-1}\gamma_{\mu_1\nu_1})\cdots C^{-1}\gamma_{\mu_j\nu_j}), \tag{5.1}$$

$$f_{\mu_1\nu_1\cdots\mu_n\nu_n} = (\tfrac{1}{4})^{n}\,\mathrm{Tr}(\cdots\mathrm{Tr}(\Psi C^{-1}\gamma_{\mu_1\nu_1})\cdots C^{-1}\gamma_{\mu_n\nu_n}). \tag{5.2}$$

These have inverses

$$\Psi = (\tfrac{1}{2})^{j} F_{\mu_1\nu_1\cdots\mu_j\nu_j}(\gamma^{\mu_1\nu_1}C \otimes \cdots \otimes \gamma^{\mu_j\nu_j}C), \tag{5.3}$$

$$\Psi = (\tfrac{1}{2})^{n} f_{\mu_1\nu_1\cdots\mu_n\nu_n}(\gamma^{\mu_1\nu_1}C \otimes \cdots \otimes \gamma^{\mu_n\nu_n}C). \tag{5.4}$$

Both F and f are antisymmetric on each pair of indices $[\mu_i\nu_i]$ and are symmetric with respect to permutations of such pairs of indices. Analogously with Eq. (4.28), the trace on any two indices in different pairs vanish. Analogously with (4.27) and (4.30), the parts of F and f that are totally antisymmetric on two pairs of indices or on three indices including a pair vanish. For integer spin, these reduce F to $2(2j+1)$ complex components. For half-odd-integer spin there is the further relation

$$\gamma^{\mu\nu} f_{\mu\nu\cdots} = 0, \tag{5.5}$$

or equivalently

$$\gamma^{\mu} f_{\mu\nu\cdots} = 0, \tag{5.6}$$

which reduces f to $2(2j+1)$ components. The symmetries and equations obeyed by our F and f are identical to all the corresponding equations for the tensor and tensor–spinor field strengths R constructed by Berends *et al.*[24] from the Lagrangian potentials. Chiral parts for both f and F may be defined by using the chiral parts of Ψ in definitions (5.1) and (5.2) and these are (anti-)self-dual on each successive pair of antisymmetric indices. Majorana multispinors continue to correspond to Majorana tensor–spinors and real tensors.

As for lower spins, the generalized Weyl equation (2.2) yields mass irreducibility and an integrability condition for F and f, allowing the Poincaré lemma to be applied. The uniformity of the direct procedures used here to derive the forms of the Lagrangian wave equations for spins-1, -$\frac{3}{2}$, -2, and -$\frac{5}{2}$ from the simpler unmixed spin reps strongly suggests that the method is extendable to obtain the Lagrangians for arbitrary spin.[16,17,21–23] Further new features occur beyond spin-$\frac{5}{2}$, such as the vanishing of the double trace of the Lagrangian field for spin $\geqslant 4$, and one would expect these to arise naturally in the transition from the unmixed to the mixed spin irreps rather than requiring *ad hoc* assumptions.

Nevertheless, beyond spin-$\frac{5}{2}$ one would wish to find the general procedure to carry out the j or j-$\frac{1}{2}$ integrations to obtain the Lagrangian fields, unique up to the appropriate gauge freedoms. The extension to arbitrary spin of the procedures developed here will be considered in subsequent papers.[25]

VI. MASSIVE FIELDS

The isomorphism in Eq. (4.1) and (4.2) is possible for massless spin-1 because the matrices $\gamma_{\mu\nu}C$ span the space of $D(1,0)\oplus D(0,1)$ spinors. For massive spin-1 fields it is not possible to use these $D(1,0)\oplus D(0,1)$ fields alone since a simple first-order wave equation is not then derivable. The same is true for all $j\geqslant 1$ with nonzero mass. Bargmann–Wigner multispinors are used instead. For massive spin-1 the space of symmetric spinors of rank two is spanned by the matrices $(\gamma_{\mu\nu}C,\gamma_\mu C)$ so a vector field A_μ must be introduced[26] in addition to $F_{\mu\nu}$. One expands Ψ as

$$\Psi=\tfrac{1}{2}F_{\mu\nu}\gamma^{\mu\nu}C+A_\mu\gamma^\mu C, \qquad (6.1)$$

with

$$F_{\mu\nu}=\tfrac{1}{4}\operatorname{Tr}(\Psi C^{-1}\gamma_{\mu\nu}),\quad A_\mu=\tfrac{1}{4}\operatorname{Tr}(\Psi C^{-1}\gamma_\mu). \qquad (6.2)$$

As is well known,[10,26] applying the Bargmann–Wigner equations to Ψ implies $F_{\mu\nu}$ is dependent on A_μ,

$$F_{\mu\nu}=(1/m)(\partial_\mu A_\nu-\partial_\nu A_\mu), \qquad (6.3)$$

and A_μ obeys the usual Proca equations

$$(\Box+m^2)A_\mu=0,\quad \partial^\mu A_\mu=0. \qquad (6.4)$$

As for the massless fields, the complex conjugates of A_μ and $F_{\mu\nu}$ are related to the Majorana conjugate of Ψ by (6.2) so that a Majorana multispinor Ψ corresponds to real fields A_μ and $F_{\mu\nu}$.

Similar results hold for spin-$\frac{3}{2}$ and higher with the general Rarita–Schwinger equations arising for half-odd-integer spin and the general Fierz–Pauli equations arising for integer spin.[10,26,27] At all levels, Majorana multispinors correspond to Majorana tensor–spinors and real tensors.

VII. CONCLUSION

We have demonstrated how massless multispinor fields of unmixed spin may be used to derive the standard massless Lagrangian field equations of helicity 1, $\frac{3}{2}$, 2, and $\frac{5}{2}$. We have also displayed explicitly the relationship between reality or Majorana conditions on the Lagrangian fields and a generalized Majorana condition on the underlying multispinor fields. The correspondence between the multispinor fields and associated tensor and tensor–spinor fields is essentially uniform for arbitrary spins. The method used here for deriving the standard Lagrangian field equations ought to be extendable, in a uniform manner, to arbitrary spin.

APPENDIX: CONVENTIONS AND IDENTITIES

Matrix operations: superscripts on matrices have the following meanings:

T = transpose,

$*$ = complex conjugate,

$\dagger$ = Hermitian conjugate = $*T$.

Our metric and Levi-Civita conventions are $(\eta_{\mu\nu})=\operatorname{diag}(+1,-1,-1,-1)$, $\epsilon^{0123}=+1$. Indices enclosed in (square) round brackets are (anti)symmetrized according to

$$T_{(\mu_1\cdots\mu_m)}=\frac{1}{m!}\sum T_{\mu_{\pi 1}\cdots\mu_{\pi m}},$$

$$T_{[\mu_1\cdots\mu_m]}=\frac{1}{m!}\sum(-1)^\pi T_{\mu_{\pi_1}\cdots\mu_{\pi m}},$$

where the sum is taken over all the permutations $\pi(n)$ of the numbers $1,\dots,m$ and $(-1)^\pi=+1$ if π is even, or -1 if π is odd.

Dual tensors: for an antisymmetric tensor $F_{[\mu\nu]}=F_{\mu\nu}$, the dual is defined to be $\widetilde{F}_{\mu\nu}=\frac{1}{2}\epsilon_{\mu\nu\lambda\rho}F^{\lambda\rho}$. The dual can also be defined on any pair of antisymmetric indices of any tensor and is then denoted by placing the tilde (∼) over the pair of indices

$$A_{\cdots\mu\nu\cdots}=\tfrac{1}{2}\epsilon_{\mu\nu\lambda\rho}A_{\dots}{}^{\lambda\rho}{}_{\dots}\,.$$

The dual satisfies $\widetilde{\widetilde{F}}_{\mu\nu}=-F_{\mu\nu}$.

Dirac algebra: The defining relations are

$$\{\gamma^\mu,\gamma^\nu\}=2\eta^{\mu\nu},\quad \gamma_0^\dagger=\gamma_0,\quad \gamma_k^\dagger=-\gamma_k. \qquad (A1)$$

Some other definitions and relations follow:

$$\gamma^{\mu\nu}=\tfrac{1}{2}i[\gamma^\mu,\gamma^\nu],\quad \gamma_5=i\gamma^0\gamma^1\gamma^2\gamma^3, \qquad (A2)$$

$$\{\gamma_5,\gamma_\mu\}=0,\quad [\gamma_5,\gamma_{\mu\nu}]=0, \qquad (A3)$$

$$\gamma_5\gamma_5=1,\quad \gamma_5^\dagger=\gamma_5, \qquad (A4)$$

$$\gamma_0\gamma_\mu^\dagger\gamma_0=\gamma_\mu,\quad \gamma_0\gamma_{\mu\nu}^\dagger\gamma_0=\gamma_{\mu\nu}, \qquad (A5)$$

$$\gamma_0\gamma_5^\dagger\gamma_0=\gamma_5,\quad \gamma_0(\gamma_\mu\gamma_5)^\dagger\gamma_0=\gamma_\mu\gamma_5, \qquad (A6)$$

$$i\widetilde{\gamma}_{\mu\nu}=\gamma_5\gamma_{\mu\nu}, \qquad (A7)$$

$$\gamma^\mu\gamma^{\nu\lambda}=\epsilon^{\mu\nu\lambda\rho}\gamma_5\gamma_\rho+i\eta^{\mu\nu}\gamma^\lambda-i\eta^{\mu\lambda}\gamma^\nu, \qquad (A8)$$

$$[\gamma^\mu,\gamma^{\nu\lambda}]=2i(\eta^{\mu\nu}\gamma^\lambda-\eta^{\mu\lambda}\gamma^\nu), \qquad (A9)$$

$$[\gamma^{\mu\nu},\gamma^{\lambda\rho}]=2i(\eta^{\nu\lambda}\gamma^{\mu\rho}+\eta^{\mu\rho}\gamma^{\nu\lambda}-\eta^{\mu\lambda}\gamma^{\nu\rho}-\eta^{\nu\rho}\gamma^{\mu\lambda}), \qquad (A10)$$

$$\gamma^\mu\gamma_\mu=4,\quad \gamma^{\mu\nu}\gamma_{\mu\nu}=12, \qquad (A11)$$

$$\operatorname{Tr}(\gamma^{\mu_1}\cdots\gamma^{\mu_k})=\operatorname{Tr}(\gamma^{\mu_1}\cdots\gamma^{\mu_k}\gamma_5)=0\quad (k\ \text{odd}), \qquad (A12)$$

$$\operatorname{Tr}(\gamma^\mu\gamma^\nu)=4\eta^{\mu\nu}, \qquad (A13)$$

$$\operatorname{Tr}(\gamma_5)=\operatorname{Tr}(\gamma_\mu)=\operatorname{Tr}(\gamma_{\mu\nu})=\operatorname{Tr}(\gamma_{\mu\nu}\gamma_5)=0, \qquad (A14)$$

$$\operatorname{Tr}(\gamma^{\mu\nu}\gamma^{\lambda\rho})=4(\eta^{\mu\lambda}\eta^{\nu\rho}-\eta^{\mu\rho}\eta^{\nu\lambda}). \qquad (A15)$$

Regarding the nonzero components of $\gamma_{\mu\nu}$ as six independent matrices, the following collection forms a linearly independent set and is a basis for the set of 4×4 matrices over the complex numbers:

$$\{1,\gamma_\mu,\gamma_{\mu\nu},\gamma_\mu\gamma_5,\gamma_5\}. \qquad (A16)$$

In particular, for any traceless 4×4 matrix W that decomposes in the chiral representation into diagonal 2×2 block form, one has the completeness relation

$$W=\tfrac{1}{8}\operatorname{Tr}(W\gamma_{\mu\nu})\gamma^{\mu\nu}. \qquad (A17)$$

Another basis for the 4×4 complex matrices is given by

$$\{C,\gamma_\mu C,\gamma_{\mu\nu}C,\gamma_\mu\gamma_5 C,\gamma_5 C\}, \qquad (A18)$$

where C is the charge conjugation matrix discussed in the text. Note that $\gamma_\mu C$ and $\gamma_{\mu\nu}C$ are symmetric, while the others are antisymmetric.

Any two representations γ_μ and $\hat{\gamma}_\mu$ of the Dirac algebra are related by a similarity transformation

$$\hat{\gamma}_\mu=U\gamma_\mu U^{-1}, \qquad (A19)$$

where U may be taken to be unitary due to the unitarity (A1) of the γ elements. The $D(\frac{1}{2},0) \oplus D(0,\frac{1}{2})$ spinors are hence transformed between representations according to

$$\Psi \to \hat{\Psi} = U\Psi, \quad \bar{\Psi} \to \hat{\bar{\Psi}} = \bar{\Psi} U^{-1}. \tag{A20}$$

Similarly, spin j multispinors transform according to

$$\Psi \to \hat{\Psi} = (U \otimes U \otimes \cdots \otimes U)\Psi, \tag{A21}$$

where there are $2j$ matrices U, any number of which may be transferred to the right-hand side by transposing. The appropriate transformation for C is

$$\hat{C} = UCU^T. \tag{A22}$$

Under (A22), the properties (3.5) also hold for $\hat{C}$, while Eq. (3.6) are not preserved under (A22) with an arbitrary unitary matrix U. Furthermore, a charge-conjugate spinor correctly transforms between representations as a spinor

$$\Psi^c = (C\gamma_0^T \otimes C\gamma_0^T \otimes \cdots C\gamma_0^T)\Psi^* \to (U \otimes U \otimes \cdots \otimes U)\Psi^c. \tag{A23}$$

[1] G. P. Collins, M.Sc. thesis, University of Canterbury, 1985.

[2] N. A. Doughty and D. L. Wiltshire, "Weyl field-strength symmetries for arbitrary helicity and gauge-invariant Fierz–Pauli and Rarita–Schwinger wave equations," accepted for publication in J. Phys. A.

[3] V. Bargmann and E. P. Wigner, Proc. Natl. Acad. Sci. USA **34**, 211 (1946).

[4] A. O. Barut and R. Rączka, *Theory of Group Representations and Applications* (Polish Scientific, Warsaw, 1980).

[5] P. van Nieuwenhuizen, Phys. Rep. **68**, 189 (1981), see especially pp. 363–365.

[6] E. P. Wigner, *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra* (Academic, New York, 1959).

[7] P. H. Butler, *Point Group Symmetry Applications* (Plenum, New York, 1981).

[8] C. Itzykson and J.-B. Zuber, *Quantum Field Theory* (McGraw–Hill, New York, 1981), p. 693.

[9] W. E. Brittin, W. R. Smythe, and W. Wyss, Am. J. Phys. **50**, 693 (1982).

[10] D. L. Wiltshire, M.Sc. thesis University of Canterbury, 1983.

[11] W. Rarita and J. Schwinger, Phys. Rev. **60**, 61 (1941).

[12] E. A. Lord, *Tensors, Relativity and Cosmology* (McGraw–Hill, New Delhi, 1976).

[13] F. A. E. Pirani, "Introduction to gravitational radiation theory," in *Lectures on General Relativity: 1964 Brandeis Summer Institute in Theoretical Physics*, edited by S. Deser and K. Ford (Prentice–Hall, Englewood Cliffs, NJ, 1964), p. 278.

[14] M. Fierz and W. Pauli, Proc. Roy. Soc. London Ser. A **173**, 211 (1939).

[15] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).

[16] F. A. Berends, G. J. H. Burgers, and H. Van Dam, Preprint IFP234-UNC, University of North Carolina, 1984.

[17] G. J. H. Burgers, Ph. D. thesis, Rijksuniversiteit te Leiden, 1985.

[18] F. A. Berends, J. W. van Holten, P. van Nieuwenhuizen, and B. de Wit, Phys. Lett. B **83**, 188 (1979).

[19] F. A. Berends, J. W. van Holten, P. van Nieuwenhuizen, and B. de Wit, Nucl. Phys. B **154**, 261 (1979).

[20] F. A. Berends, J. W. van Holten, B de Wit, and P. van Nieuwenhuizen, J. Phys. A **13**, 1643 (1980).

[21] J. Fronsdal, Phys. Rev. D **18**, 3624 (1978).

[22] J. Fang and C. Fronsdal, Phys. Rev. D **18**, 3630 (1978).

[23] B. de Wit and D. Z. Freedman, Phys. Rev. D **21**, 358 (1980).

[24] F. A. Berends, G. J. H. Burgers, and H. van Dam, "Explicit construction of conserved currents for massless fields of arbitrary spin," Rijksuniversiteit te Leiden preprint, 1985.

[25] N. A. Doughty and G. P. Collins, "Gauge-invariant Lagrangian wave equations of arbitrary helicity" (to be published); G. P. Collins and N. A. Doughty, "Systematics of arbitrary-helicity Lagrangian wave equations" (to be published).

[26] M. A. Rodriguez and M. Lorente, J. Math. Phys. **22**, 1283 (1981).

[27] S. N. Gupta and W. W. Repko, Phys. Rev. **165**, 1415 (1968).

Coordinate-independent formulation of the Langevin equation

G. G. Batrouni, H. Kawai, and Pietro Rossi
Newman Laboratory of Nuclear Studies, Cornell University, Ithaca, New York 14853

(Received 2 December 1985; accepted for publication 21 February 1986)

A diffusion process on a compact Riemannian manifold is considered, and a coordinate-invariant Fokker–Planck equation is formulated. A covariant form of the Langevin equation is also derived, and the formalism is applied to the stochastic quantization of lattice gauge theories.

I. INTRODUCTION

Variables in lattice gauge theories take values in Lie groups; the problem of describing the stochastic quantization of these theories led us to consider diffusion on compact Riemannian manifolds with metric $g_{\mu\nu}$ without boundaries. Stochastic processes and their use in the quantization of field theories in flat manifolds is an extensively studied subject.[1,2]

Diffusion on Riemannian manifolds was discussed in detail in Ref. 3, where a coordinate-invariant Fokker–Planck equation was presented. In this paper we will discuss the same equation but our approach differs from that of Ref. 3 in that our underlying Langevin equation is simpler. The reason for this is that the authors of Ref. 3 explicitly avoided using metric-dependent noise terms that resulted in two rather unconventional coordinate-invariant Langevin equations. As we shall see below, using metric-dependent noise averages results in a very simple form for the Langevin equation.

We propose here to give a presentation of the topic in a form helpful for the purpose of numerical simulation of lattice gauge theories. In practice this calls for simulation of the relevant diffusion process by a discrete step Langevin equation; it will be our purpose to derive this.

II. COORDINATE-INVARIANT FOKKER–PLANCK EQUATION

If $P(x,t)$ is a time-dependent probability distribution, defined on a manifold M, P will satisfy the equation

$$\frac{d}{dt}\int_M d^D x\sqrt{g}P(x,t) = 0\,. \tag{2.1}$$

Let $\rho(x,t) = \sqrt{g}P$, if we limit ourselves to boundaryless manifolds, Eq. (2.1) is guaranteed if

$$\frac{\partial}{\partial t}\rho = -\partial_\mu j^\mu\,, \tag{2.2}$$

where j^μ is a current density. A popular choice for j^μ is given by

$$j^\mu = F^\mu\rho - kT\sqrt{g}g^{\mu\nu}\partial_\nu((1/\sqrt{g})\,\rho)\,. \tag{2.3}$$

The first term on the right-hand side of (2.3) is the so-called drift force. The second term is the diffusion term,[1] F^μ is a force field tangent vector defined on the manifold M, and T is the temperature. The diffusion term is also a vector density, since $(1/\sqrt{g})$ is a scalar. The form (2.3) for j^μ guarantees that if F^μ is a gradient, that is

$$F^\mu = -g^{\mu\nu}\partial_\nu V\,, \tag{2.4}$$

then a stationary solution to Eq. (2.2) is given by

$$\rho = \sqrt{g}e^{-V/kT}\,. \tag{2.5}$$

Thus the general coordinate-invariant diffusion equation on a curved manifold is given by

$$\frac{\partial\rho}{\partial t} = -\partial_\mu\left\{F^\mu\rho - kT\sqrt{g}g^{\mu\nu}\partial_\nu\left(\frac{1}{\sqrt{g}}\rho\right)\right\} \tag{2.6}$$

$$\begin{aligned}&= \partial_\mu\{[-F^\mu - kT(1/\sqrt{g})\partial_\nu(\sqrt{g}g^{\mu\nu})]\,\rho\}\\ &\quad + \partial_\mu\,\partial_\nu(kTg^{\mu\nu}\rho)\,,\end{aligned} \tag{2.7}$$

where all differential operators are moved to the left.

III. THE GENERALIZED LANGEVIN EQUATION

We shall try to define now a stochastic process that is coordinate independent, in the sense that the probability distribution must satisfy the coordinate invariant diffusion equation (2.6).

We make use of the following asymptotic expansion.

$$\begin{aligned}&\left(\frac{1}{\sqrt{\Delta t}}\right)^d W\left(\frac{z}{\sqrt{\Delta t}}\right)\\ &\quad = f_{(0)}\delta^d(z) - f^i_{(1)}\,\partial_i\,\delta^d(z)\sqrt{\Delta t}\\ &\qquad + \tfrac{1}{2}f^{ij}_{(2)}\,\partial_i\,\partial_j\,\delta^d(z)(\sqrt{\Delta t}\,)^2\\ &\qquad - (1/3!)\,f^{ijk}_{(3)}\,\partial_i\,\partial_j\,\partial_k\,\delta^d(z)(\sqrt{\Delta t}\,)^3 + \cdots\,.\end{aligned} \tag{3.1}$$

Here $W(x)$ is a smooth function from R^d to R and the numbers $f_{(0)}, f^i_{(1)},\ldots,f^{i\ldots in}_{(n)}$ are defined by the following equations:

$$\begin{aligned}&\int d^dx\;W(x) = f_{(0)}\,,\\ &\int d^dx\,x^iW(x) = f^i_{(1)}\,,\\ &\int d^dx\,x^{i_1}\cdots x^{i_n}W(x) = f^{i_1\ldots i_n}_{(n)}\,.\end{aligned} \tag{3.2}$$

The asymptotic expansion can easily be proved multiplying both sides of Eq. (3.1) by a monomial $z^{i_1}\cdots z^{i_r}$ and integrating over z.

Next we consider the following generalized form of the discrete Langevin equation

$$\mathbf{x}_{n+1} - \mathbf{x}_n = \sqrt{\Delta t}\,\boldsymbol{\eta}_n\,, \tag{3.3}$$

where the $\boldsymbol{\eta}_n$ are independent stochastic variables whose moments are fixed to the following values:

0022-2488/86/061646-03$02.50

$$\overline{\eta^i_n} = \sqrt{\Delta t}\, f^i(x_n)\,,$$
$$\overline{\eta^i_m\, \eta^j_n} = f^{ij}(x_n)\delta_{mn}\,. \tag{3.4}$$

We further demand that the higher moments are not too large, or more precisely

$$\overline{\eta^{i_1}_n \dots \eta^{i_M}_n} \sim o(1/\sqrt{\Delta t})^{M-2} \quad (M \geqslant 3)\,. \tag{3.5}$$

If we write the probability distribution of $\boldsymbol{\eta}_n$ as $W_{x_n}(\boldsymbol{\eta}_n)$, the Fokker–Planck kernel for a time interval Δt can be written as follows:

$$K(x,y) = \int d^d\eta\, \delta(x - y - \eta\sqrt{\Delta t})\, W_y(\eta)$$
$$= \left(\frac{1}{\sqrt{\Delta t}}\right)^d W_y\left(\frac{x-y}{\sqrt{\Delta t}}\right). \tag{3.6}$$

Applying the general formula (3.1) we obtain

$$K(x,y) = \delta^d(x-y) + \Bigg\{ -f^i(y)\frac{\partial}{\partial x^i}\delta^d(x-y)$$
$$+ \frac{f^{ij}(y)}{2}\frac{\partial}{\partial x^i}\frac{\partial}{\partial x^j}\delta^d(x-y)\Bigg\}\Delta t + O(\Delta t^{3/2})\,. \tag{3.7}$$

Thus we see that the discretized Fokker–Planck equation

$$P(x,t+\Delta t) = \int d^d y\, K(x,y)P(y,t) \tag{3.8}$$

has a well-defined $\Delta t \to 0$ limit if the higher moments of the stochastic noise do not violate (3.5):

$$\dot{P}(x,t) = \partial i[f^i(x)P] + \tfrac{1}{2}\partial i\, \partial j[f^{ij}(x)P]\,. \tag{3.9}$$

By comparing (3.9) and (2.7) we find easily that the Langevin equations

$$x^\mu_{n+1} = x^\mu_n + \sqrt{\Delta t}\,\eta^\mu_n\,, \tag{3.10a}$$

$$\overline{\eta^\mu_n} = \sqrt{\Delta t}\left[F^\mu(x_n) + \frac{kT}{\sqrt{g(x_n)}}\partial_\nu(\sqrt{g(x_n)}g^{\mu\nu}(x_n)\right], \tag{3.10b}$$

$$\overline{\eta^\mu_m \eta^\nu_n} = 2kTg^{\mu\nu}(x_n)\delta_{mn}\,, \tag{3.10c}$$

$$\overline{\eta^{\mu_1}_m \dots \eta^{\mu_M}_m} \sim o(1/\sqrt{\Delta t})^{M-2}$$

describe the process (2.6). Equations (3.10) can be rewritten by shifting the random variables η^μ_m by their mean values

$$x^\mu_{m+1} = x^\mu_m + \Delta t\left\{F^\mu(x_m) + \frac{kT}{\sqrt{g(x_m)}}\partial_\nu(\sqrt{g}g^{\mu\nu})\right\}$$
$$+ \sqrt{\Delta t}\;\eta^\mu_m\,, \tag{3.11a}$$

$$\overline{\eta^\mu_m} = 0\,, \tag{3.11b}$$

$$\overline{\eta^\mu_m \eta^\nu_n} = 2kTg^{\mu\nu}(x_n)\delta_{mn}\,. \tag{3.11c}$$

Here the corrections to the second moment (3.11c) are irrelevant because they are of higher order.

Equations (3.11) describe a covariant Langevin equation. This covariance property allows us to perform a change of coordinates at every step, and in particular we can always choose an inertial frame, where $g^{\mu\nu}(x) = \delta^{\mu\nu}$ and $\partial_\rho\, g^{\mu\nu}(x) = 0$, as will be the case in Sec. IV.

IV. STEP BY STEP COVARIANCE OF THE LANGEVIN PROCESS

It is of course possible to check explicitly that the Langevin process described by Eq. (2.10) is covariant.

Let us consider the average jump taken at the mth step. From (3.10) we have

$$\overline{x^\mu_{m+1} - x^\mu_m}$$
$$= \sqrt{\Delta t}\;\; \overline{\eta^\mu_m}$$
$$= \Delta t\left[F^\mu(x_m) + \frac{1}{\sqrt{g(x_m)}}\partial_\nu(g^{\mu\nu}(x_m)\sqrt{g(x_m)})\right], \tag{4.1}$$

$$\overline{(x^\mu_{m+1} - x^\mu_m)(x^\nu_{m+1} - x^\nu_m)} = \Delta t\, 2kTg^{\mu\nu}(x_m)\,. \tag{4.2}$$

If we now perform a change of coordinates $y^\mu = y^\mu(x)$ then the μth step will be described in the y coordinate system by

$$\overline{y^\mu_{m+1} - y^\mu_m} = \overline{y^\mu(x_{m+1}) - y^\mu(x_m)}\,. \tag{4.3}$$

Expanding the function $y^\mu(x_{m+1})$ around x_m we obtain

$$\overline{y^\mu_{m+1} - y^\mu_m}$$
$$= \partial_\nu\, y^\mu(x_m)\;\overline{(x^\nu_{m+1} - x^\nu_m)} + \tfrac{1}{2}\partial_\nu\, \partial_\sigma\, y^\mu(x_m)$$
$$\times\; \overline{(x^\nu_{m+1} - x^\nu_m)(x^\sigma_{m+1} - x^\sigma_m)} + \cdots$$
$$= \Delta t\, \partial_\nu\, y^\mu(x_m)\Bigg[-F^\nu(x_m)$$
$$+ \frac{1}{\sqrt{g(x_m)}}\partial_\sigma(g^{\nu\sigma}(x_m)\sqrt{g(x_m)})\Bigg]$$
$$+ \Delta t\, \partial_\nu\, \partial_\sigma\, y^\mu(x_m)[kTg^{\nu\sigma}(x_m)]\,. \tag{4.4}$$

The last two terms combine in the covariant Laplacian and we have

$$\overline{y^\mu_{m+1} - y^\mu_m} = \Delta t F^\nu(x_m)\frac{\partial y^\mu(x_m)}{\partial x^\nu_m} + \Delta t \nabla y^\mu(x_m)\,. \tag{4.5}$$

The first term is clearly the force expressed in the y-coordinate system, and the second term can be written as

$$\frac{\Delta t}{\sqrt{g'(y)}}\frac{\partial}{\partial y^\sigma}\left[g'^{\sigma\nu}(y)\sqrt{g'(y)}\frac{\partial y^\mu}{\partial y^\nu}\right]$$
$$= \frac{\Delta t}{\sqrt{g'(y)}}\partial_\sigma[g'^{\sigma\mu}(y)\sqrt{g'(y)}]\,. \tag{4.6}$$

The expression for the width in the y system is even simpler:

$$\overline{(y^\mu_{m+1} - y^\mu_m)(y^\nu_{m+1} - y^\nu_m)}$$
$$= \partial_\sigma\, y^\mu(x_m)\partial_\rho\, y^\nu(x_m)\;\overline{(x^\sigma_{m+1} - x^\sigma_m)(x^\rho_{m+1} - x^\rho_m)}$$
$$= 2kTg'^{\mu\nu}(y_m)\,. \tag{4.7}$$

V. LATTICE GAUGE THEORIES: AN EXAMPLE

We consider the Langevin equation (3.11) on a group manifold. At every step we parametrize the manifold in the following way:

$$U = e^{x\cdot T}U_N\,, \tag{5.1}$$

where U_N is the instantaneous position in group space. We want to compute $g^{ij}(x_N)$ and $\partial_k g^{ij}(x)/x = x_N$, with $x_N = 0$. The line element generated by the Haar measure is given by

$$ds^2 = \tfrac{1}{2}\,\mathrm{Tr}\,(dUU^{-1})^2 = g_{ij}(x)dx^i dx^j . \tag{5.2}$$

Next we express dUU^{-1} as a power series in x, where

$$dUU^{-1} = (dx\cdot T) + \tfrac{1}{2}[x\cdot T, dx\cdot T] + o(x^2) . \tag{5.3}$$

From (5.2) it is easy to show that

$$g_{ij}(0) = \delta^{ij} , \tag{5.4a}$$

$$\partial_k\, g_{ij}(0) = 0 . \tag{5.4b}$$

Using (5.4) to (5.5) the Langevin equation (3.11) becomes

$$x^i_{N+1} = x^i_N - \Delta t F^{\,i}(x_N) + \sqrt{\Delta t}\,\eta^i_N(x_N) , \tag{5.5a}$$

$$\overline{\eta^i_N}(x_N) = 0 , \tag{5.5b}$$

$$\overline{\eta^i_N\,\eta^j_M} = 2kT\delta^{ij}\delta_{NM} , \tag{5.5c}$$

but $x_N = 0$ and the $N+1$ step on the group manifold is described by

$$U_{N+1} = \exp[\,-\Delta t F^i(x_N) + \sqrt{\Delta t}\,\eta^i_N T^i\,]\,U_N , \tag{5.6}$$

where

$$F^i = \delta^{ij}\frac{\partial S}{\partial x^j}\bigg|_{x=x_N} = (T^iU)_{ab}\left(\frac{\partial S}{\partial U_{ab}}\right) . \tag{5.7}$$

This Langevin equation for lattice gauge theory has been studied by many authors.[2,4]

ACKNOWLEDGMENTS

The authors wish to thank D. Zwanziger for providing motivation for this work and sharing with us many useful insights.

[1] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1981).

[2] G. Parisi and Y. Wu, Sci. Sin. **24**, 483 (1981).

[3] M. Claudson and M. B. Halpern, Phys. Rev. D **31**, 3310 (1985); M. Claudson and M. B. Halpern, preprint No. UCB-PTH-84/28, to appear in Ann. Phys. (NY).

[4] M. B. Halpern, Nucl. Phys. B **228** (1983); I. T. Drummond, S. Duane, and R. R. Horgan, Nucl. Phys. B **220** 119 (1983); A. Guha and S. C. Lee, Phys. Rev. D **27**, 2412 (1983); H. Hamber and U. M. Heller, Phys. Rev. D **29**, 928 (1984); G. G. Batrouni, G. R. Katz, A. S. Kronfeld, G. P. Lepage, B. Svetitsky, and K. G. Wilson, Phys. Rev. D **32**, 2736 (1985).

A simple Grassmannian path integral representation of the Dirac propagator

Helmut Rumpf
Institut für Theoretische Physik, Universität Wien, A-1090 Vienna, Austria

(Received 14 June 1985; accepted for publication 30 January 1986)

Starting from the affinely parametrized supersymmetric Dirac particle model a Grassmannian path integral expression for the propagator of the Dirac equation, minimally coupled to an external electromagnetic field, is derived. A purely "bosonic" path integral representation of the propagator of the iterated minimally coupled Dirac equation is also obtained. It appears that a Nicolai mapping exists, even in a formal sense, only in the case of a constant external field.

I. INTRODUCTION

The history of attempts to construct a classical model of the Dirac electron is a long and intriguing one (see Refs. 1 and 2 for early references). Three different reasons for this continued interest can be identified: (i) aesthetic ones, (ii) the search for a toy model of field theory (mainly in the context of supersymmetry), and, last but not least, (iii) the desire to learn something about the Dirac particle itself. A minimal requirement for considering a classical electron model as "correct" is that when quantized it should give rise to the Dirac equation. Until recently, the only models that fulfilled this requirement involved anticommuting numbers. Out of these we mention the model of Berezin and Marinov[1] whose action is invariant under arbitrary reparametrizations of the evolution parameter and exhibits a gauge supersymmetry, and the model of Di Vecchia and Ravndal,[3] where the parameter is fixed up to affine transformations and there exists a *global* supersymmetry. The most recent upsurge of interest in this area is due to the fact that models involving only complex numbers have been found. We refer to the work of Barut and collaborators,[4,5] who succeeded in constructing a classical action with the desired properties, and to the work of Jacobson *et al.*,[6,7] who extended Feynman's checkerboard path integral[8] to $3+1$ dimensions and also constructed the stochastic process related to this path integral via analytic continuation. (Strictly speaking the latter approach does not define a model since there is no classical action involved in the checkerboard prescription.) One might take these results as an indication that "anticommuting *c*-numbers are an unnecessary addition to mathematical physics."[9] We feel, however, that simplicity and predictive power are also relevant criteria for the selection of a model, and in this respect the globally supersymmetric model[3] is unsurpassed. It is the only model in which so far the minimal coupling to an external electromagnetic field *and* to an external gravitational field including torsion could be incorporated.[2] Most remarkably, these couplings are uniquely determined by the supersymmetry. It seemed therefore interesting to derive a path integral expression of the Dirac propagator from this model.

In Sec. II of this paper we briefly review the model of Di Vecchia and Ravndal and introduce a pseudo-Schrödinger representation for its canonically quantized version, which is more general than the Dirac representation. Based on this representation, path integral expressions for the Dirac propagator and iterated Dirac propagator in an external electromagnetic field are derived in Sec. III. Finally in Sec. IV we address the question of whether the fact that the path integral for the iterated propagator can be reduced to a "bosonic" one can be interpreted in terms of a Nicolai mapping. We find that a rigorous version of the mapping exists only for the free particle and even a formal version appears to exist only in the case of a constant external field.

II. SUPERSYMMETRIC QUANTUM MECHANICS OF A SPIN-$\frac{1}{2}$ PARTICLE

In the model of Di Vecchia and Ravndal[3] the classical spin degrees of freedom are described by a four-vector with real anticommuting components ξ^a. The free-particle Lagrangian is

$$L = (\dot{x}^2/2) - (i/2)\xi\dot{\xi}\,. \tag{2.1}$$

The dot denotes differentiation with respect to a parameter λ, which is equal to s/m if the equations of motion are fulfilled, where s is the proper time and m the mass of the particle; λ is well defined even if $m=0$, in which case it is an "affine parameter." The Lagrangian (2.1) is invariant (up to a time derivative) under the supersymmetry transformations

$$\delta x^a = i\epsilon\xi^a\,, \tag{2.2}$$

$$\delta\xi^a = \epsilon\dot{x}^a\,. \tag{2.3}$$

These transformations can be implemented as "supertranslations" in a superspace formulation of the model, where in addition to λ there appears an anticommuting evolution parameter θ. Note that in contrast to the supersymmetric relativistic field theories no spinors appear here at the classical level. The "simple" supersymmetry (to be consistent with current terminology in supersymmetric nonlinear σ models, one should call it $N=\frac{1}{2}$ supersymmetry) embodied in (2.2) and (2.3) may be generalized to an "extended" one; the resulting theory in the $N=2$ (alias $N=1$) extended case is a relativistic extension of Witten's version of supersymmetric quantum mechanics[10] and provides a classical model of the photon.[11]

Canonical quantization of the model yields, in a straightforward manner, the Dirac theory: The pseudoclassical Dirac bracket[2,12]

$$[\xi^a, \xi^b\} = i\eta^{ab} \tag{2.4}$$

is replaced at the quantum level by the Clifford algebra rela-

0022-2488/86/061649-06$02.50

tions

$$\{\hat{\xi}^a, \hat{\xi}^b\} = \hbar\eta^{ab}. \tag{2.5}$$

Thus the quantum spin variables $\hat{\xi}^a$ may be represented by $(\hbar/2)^{1/2}\gamma^a$, where γ^a are the Dirac matrices, and this representation is essentially unique. However for our purposes it will be useful to transcend the standard notion of an algebra representation and to consider as a representation of (2.4) a space of functions defined on a Grassmann algebra. This Grassmann algebra will be an analog of the configuration space of the x variables. We stress that the ξ^a have to be considered as phase-space variables and that no natural configuration space exists for the spin variables.[1] Nonetheless, due to the even dimensionality of space-time, it is possible to identify "position" and "momentum" variables by exhibiting the pseudosymplectic structure defined by (2.4) in its canonical form. We therefore define Grassmannian "coordinates" $\zeta^{\rm I}, \zeta^{\rm II}$ and "momenta" $\beta_{\rm I}$, $\beta_{\rm II}$ by

$$\zeta^{\rm I} = 2^{-1/2}(\xi^0 + \xi^3), \qquad \zeta^{\rm II} = 2^{-1/2}(\xi^1 + i\xi^2), \tag{2.6}$$

$$\beta_{\rm I} = 2^{-1/2}(\xi^0 - \xi^3), \qquad \beta_{\rm II} = 2^{-1/2}(-\xi^1 + i\xi^2), \tag{2.7}$$

obeying the bracket relations

$$[\zeta^A, \zeta^B\} = 0 = [\beta_A, \beta_B\}, \tag{2.8}$$

$$[\zeta^A, \beta_B\} = i\delta^A{}_B \tag{2.9}$$

($A, B = {\rm I}$ or II). These relations become Heisenberg-type anticommutation relations upon canonical quantization and we are going to construct the (pseudo-) "Schrödinger" representation of them. There is a certain amount of arbitrariness in the definition of the ζ^A and β_A: Any set of null vectors (two of them necessarily complex) $\{k, l, m, n\}$ with $k \cdot l = 1 = m \cdot n$ and all other scalar products vanishing provides a possible set $\{\xi \cdot k, \xi \cdot l, \xi \cdot m, \xi \cdot n\}$ of coordinates and momenta. It will be seen shortly that all choices lead to equivalent (though not identical) representations. The following representation of the quantum variables suggests itself:

$$\hat{\zeta}^A: f(\zeta) \rightarrow \hbar^{1/2}\zeta^A f(\zeta), \tag{2.10}$$

$$\hat{\beta}_A: f(\zeta) \rightarrow \hbar^{1/2}\frac{\partial}{\partial \zeta^A} f(\zeta). \tag{2.11}$$

Here f is an arbitrary analytic function of $\zeta^{\rm I}$ and $\zeta^{\rm II}$. The most general one is of the form

$$f(\zeta) = a_1 + \zeta^{\rm I}a_2 + \zeta^{\rm II}a_3 + \zeta^{\rm I}\zeta^{\rm II}a_4, \tag{2.12}$$

where the a_i are arbitrary Grassmann numbers. The Dirac representation is obtained from the ζ representation by specializing to complex a_i and having them comprise the spinor $\psi = (a_1, a_2, a_3, a_4)$. The definitions (2.10) and (2.11) imply, then,

$$\hat{\xi}^a: \psi \rightarrow (\hbar/2)^{1/2}\gamma^a\psi, \tag{2.13}$$

where the γ^a are a particular realization of the Dirac matrices.

In order to complete the quantum kinematics for the spin variables we have to define a scalar product in our representation space. We note that $\zeta^{\rm I}$ and $\beta_{\rm I}$ are real and that $\beta_{\rm II} = -\zeta^{\rm II*}$. Therefore $\partial/\partial\zeta^{\rm I}$ must be Hermitian and

$$\frac{\partial}{\partial \zeta^{\rm II}} = -\bar{\hat{\zeta}}^{\rm II}, \tag{2.14}$$

the bar denoting the adjoint. These properties determine the scalar product up to a constant factor

$$\langle f_1, f_2\rangle = \int d\zeta^{\rm I}\, d\zeta^{\rm II*}\, d\zeta^{\rm II}\, e^{-\zeta^{\rm II*}\zeta^{\rm II}} f_1^* f_2$$

$$\equiv \int dG(\zeta) f_1^* f_2. \tag{2.15}$$

This scalar product involves Berezin integration with respect to the anticommuting variables and the Grassmann involution * defined such that $(ab)^* = b^*a^*$ for arbitrary Grassmann numbers a and b (the ξ^a are constrained to be real, $\xi^{a*} = \xi^a$). For the spinors ψ corresponding to the pseudowave functions f the scalar product (2.15) implies exactly the Lorentz invariant

$$\bar{\psi}_1\psi_2 = \psi_1^\dagger\beta\psi_2, \tag{2.16}$$

$$\beta = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} = \gamma^0. \tag{2.17}$$

So far we have considered only the quantum representation of the spin variables. Of course the full representation space is the tensor product of the space just defined with a standard representation space for the translational degrees of freedom.

The dynamics of the particle is governed by the Hamiltonian H generating translations in the parameter λ. But H is determined by the generator Q of the supersymmetry transformations (2.2) and (2.3) via

$$[Q, Q\} = 2iH, \tag{2.18}$$

and therefore all the information about the dynamics is already contained in Q. In particular the quantum-mechanical mass-shell condition

$$\hat{H}|\,\rangle = (m^2/2)|\,\rangle \tag{2.19}$$

is a consequence of the eigenvalue condition

$$\hat{Q}|\,\rangle = m(\hbar/2)^{1/2}|\,\rangle. \tag{2.20}$$

The latter is just the Dirac equation.

It is amusing to observe (although this will not be needed in the rest of the paper) that the "bosonic" and "fermionic" states with respect to the supersymmetry are just the states $|\pm\rangle$ of definite chirality,

$$\hat{\xi}_5|\pm\rangle = \pm(\hbar^2/4)|\pm\rangle, \tag{2.21}$$

$$\hat{\xi}_5 := i\hat{\xi}_0\hat{\xi}_1\hat{\xi}_2\hat{\xi}_3, \tag{2.22}$$

as

$$\{\hat{Q}, \hat{\xi}_5\} = 0 \tag{2.23}$$

implies

$$\hat{Q}|+\rangle \sim |-\rangle, \qquad \hat{Q}|-\rangle \sim |+\rangle. \tag{2.24}$$

Consequently the Witten index[13] in the Dirac representation,

$$\mathrm{Tr}(-1)^F = \mathrm{Tr}\,\gamma_5 \tag{2.25}$$

(F denoting the fermion number operator), just coincides with the usual notion of the index of the Dirac operator. This fact lies at the basis of an alternative proof of the Atiyah–Singer index theorem for the Dirac operator on a Riemannian manifold.[14]

III. DERIVATION OF THE PATH INTEGRAL EXPRESSIONS

In this paper we are interested in the inverses (putting $\hbar = 1$ from now on)

$$(2\hat{H} - m^2 + i0)^{-1} = -\frac{i}{2}\int_0^\infty d\lambda\, e^{-im^2\lambda/2} e^{i\hat{H}\lambda}, \tag{3.1}$$

$$(2^{1/2}\hat{Q} - m + i0)^{-1} = \int d\theta\, e^{m\theta} e^{2^{1/2}\theta\hat{Q}} (2\hat{H} - m^2 + i0)^{-1}, \tag{3.2}$$

whose integral kernels in the Dirac representation are the propagators of the iterated and simple Dirac equation, respectively. The integration variable θ in (3.2) is anticommuting, and $\theta\hat{Q} = -\hat{Q}\theta$. In Eqs. (3.1) and (3.2) we have already indicated the manner in which these propagators will be represented, namely via path integral expressions for the kernels of $\exp(i\hat{H}\lambda)$ and $\exp(\theta\hat{Q} + i\hat{H}\lambda)$. These kernels are matrix-valued two-point functions, and it is this matrix character that presents a challenge to the construction of a path integral representation. In the context of the model of Berezin and Marinov,[1] Ogielsky and Sobczyk[15] had to use the notion of symbol of an operator (introduced in Ref. 1) for this purpose. Due to the even dimensionality of phase space, in our case we can dispense with this tool and work in the ζ representation instead.

We introduce the states $|\zeta_0\rangle = |\zeta_0^{\rm I}, \zeta_0^{\rm II*}\rangle$ represented by

$$\delta_{\zeta_0}(\zeta) = (\zeta^{\rm I} - \zeta_0^{\rm I})(1 - \zeta^{\rm II}\zeta_0^{\rm II*}), \tag{3.3}$$

and obeying

$$\hat{\zeta}^{\rm I}|\zeta_0\rangle = \zeta_0^{\rm I}|\zeta_0\rangle, \tag{3.4}$$

$$\bar{\hat{\zeta}}^{\rm II}|\zeta_0\rangle = \zeta_0^{\rm II*}|\zeta_0\rangle, \tag{3.5}$$

$$\langle\zeta_0|f\rangle = f(\zeta_0). \tag{3.6}$$

One easily checks the completeness relation

$$\mathbb{1} = \int |\zeta\rangle\, dG(\zeta) \langle\zeta|. \tag{3.7}$$

Therefore

$$\bar{\psi}_1\hat{O}\psi_2 = \langle f_1|\hat{O}|f_2\rangle = \int\int \langle f_1|\zeta_1\rangle dG(\zeta_1)\langle\zeta_1|\hat{O}|\zeta_2\rangle dG(\zeta_2)\langle\zeta_2|f_2\rangle. \tag{3.8}$$

Equation (3.8) shows that the matrix elements $\langle\zeta_1|\hat{O}|\zeta_2\rangle$ of a spin "observable" in the ζ representation indeed contain all the information necessary to construct the corresponding spinorial matrix elements $\bar{\psi}_1\hat{O}\psi_2$ (with ψ_1 and ψ_2 the spinors corresponding to f_1 and f_2) in the Dirac representation.

Consider now the matrix element

$$\langle x'', \zeta'';\lambda\,|x', \zeta';0\rangle \equiv \langle x'', \zeta''|e^{i\hat{H}\lambda}|x', \zeta'\rangle. \tag{3.9}$$

Using (3.7) we may write this as

$$\langle x'', \zeta'';\lambda\,|x', \zeta';0\rangle = \int d^4x_1 \cdots d^4x_n \langle x'', \zeta''|e^{i\hat{H}\epsilon}|x_n, \zeta_n\rangle dG(\zeta_n) \times \langle x_n, \zeta_n|e^{i\hat{H}\epsilon}|x_{n-1}, \zeta_{n-1}\rangle \cdots \langle x_2, \zeta_2|e^{i\hat{H}\epsilon}|x_1, \zeta_1\rangle dG(\zeta_1)\langle x_1, \zeta_1|e^{i\hat{H}\epsilon}|x', \zeta'\rangle, \tag{3.10}$$

$$\epsilon = \lambda/(n+1). \tag{3.11}$$

We restrict ourselves first to the free-particle case. Then

$$H = p^2/2, \tag{3.12}$$

$$\langle x_{k+1}, \zeta_{k+1}|e^{i\hat{H}\epsilon}|x_k, \zeta_k\rangle = \langle x_{k+1}|e^{i\hat{H}\epsilon}|x_k\rangle\langle\zeta_{k+1}|\zeta_k\rangle. \tag{3.13}$$

As usual we express the first factor on the right-hand side (rhs) of (3.13) as

$$\langle x_{k+1}|e^{i\hat{H}\epsilon}|x_k\rangle = \int dp_{k+1}\, dp_k \langle x_{k+1}|p_{k+1}\rangle\langle p_{k+1}|e^{i\hat{H}\epsilon}|p_k\rangle\langle p_k|x_k\rangle = \int \frac{d^4p}{(2\pi)^4} \exp[ip(x_k - x_{k+1})]\exp[i(p^2/2)\epsilon] = -(1/2\pi^2\epsilon^2)\exp[-(i/2\epsilon)(x_{k+1} - x_k)^2]. \tag{3.14}$$

Similarly we can write the second factor as

$$\langle\zeta_{k+1}|\zeta_k\rangle = -\int d\bar{G}(\beta)\exp[-\beta_{\rm I}(\zeta_{k+1}^{\rm I} - \zeta_k^{\rm I})]\exp(\zeta_{k+1}^{\rm II}\beta_{\rm II} + \beta_{\rm II}^*\zeta_k^{\rm II*}), \tag{3.15}$$

$$d\bar{G}(\beta) := d\beta_{\rm I}\, d\beta_{\rm II}\, d\beta_{\rm II}^*\, e^{-\beta_{\rm II}\beta_{\rm II}^*}. \tag{3.16}$$

Now

$$
\begin{aligned}
&-\tfrac{1}{2}(\zeta_k^{\mathrm{II}*}\zeta_k^{\mathrm{II}}+\zeta_{k+1}^{\mathrm{II}*}\zeta_{k+1}^{\mathrm{II}})-\beta_{\mathrm{II}}\beta_{\mathrm{II}}^{*}-\beta_{\mathrm{I}}(\zeta_{k+1}^{\mathrm{I}}-\zeta_k^{\mathrm{I}})+\zeta_{k+1}^{\mathrm{II}}\beta_{\mathrm{II}}+\beta_{\mathrm{II}}^{*}\zeta_k^{\mathrm{II}*}\\
&\quad=-\tfrac{1}{2}\eta_{ab}\xi^a(\xi_{k+1}^b-\xi_k^b)-\tfrac{1}{2}\xi^0(\xi_{k+1}^3-\xi_k^3)+\tfrac{1}{2}\xi^3(\xi_{k+1}^0-\xi_k^0)+(i/2)\xi^1(\xi_k^2+\xi_{k+1}^2)\\
&\qquad+(i/2)(\xi_k^1+\xi_{k+1}^1)\xi^2-i\xi^1\xi^2-(i/2)\xi_k^1\xi_k^2-(i/2)\xi_{k+1}^1\xi_{k+1}^2\,.
\end{aligned}
\tag{3.17}
$$

If we assume for the moment that $\xi_k=\xi(k\epsilon)$ with $\xi(\lambda)$ differentiable and that $\xi=\xi[(k+\frac{1}{2})\epsilon]$, then for small ϵ (3.17) becomes

$$
-(\epsilon/2)(\eta_{ab}\xi^a\dot{\xi}^b+\xi^0\dot{\xi}^3+\dot{\xi}^0\xi^3)+O(\epsilon^2)\,,
\tag{3.18}
$$

i.e., $-i\epsilon$ times the spin part of the Lagrangian (2.1) up to a total derivative. Therefore in a formal sense

$$
\langle x'',\zeta''|e^{i\hat{H}\lambda}|x',\zeta'\rangle\sim\int\mathscr{D}[x]\,\mathscr{D}[\xi]\exp\left(-i\int_0^\lambda L\,d\lambda'\right).
\tag{3.19}
$$

The rigorous result is

$$
\begin{aligned}
\langle x'',\zeta'';\lambda|x',\zeta';0\rangle=&-\int\left(\prod_{k=1}^{n}\frac{d^4x_k}{2\pi^2\epsilon^2}\right)d\overline{G}(\beta_{1/2})\left[\prod_{k=1}^{n}dG(\zeta_k)d\overline{G}(\beta_{k+1/2})\right]\exp\left\{\sum_{k=0}^{n}\left[-\frac{i}{2}\epsilon\left(\frac{x_{k+1}-x_k}{\epsilon}\right)^2\right.\right.\\
&\left.\left.-\beta_{\mathrm{I},k+1/2}(\zeta_{k+1}^{\mathrm{I}}-\zeta_k^{\mathrm{I}})+\zeta_{k+1}^{\mathrm{II}}\beta_{\mathrm{II},k+1/2}+\beta_{\mathrm{II},k+1/2}^{*}\zeta_k^{\mathrm{II}*}\right]\right\},
\end{aligned}
\tag{3.20}
$$

where $x_0=x'$, $\zeta_0=\zeta'$, $x_{n+1}=x''$, and $\zeta_{n+1}=\zeta''$. Note that there is one β integration more than there are ζ integrations, since the result is an odd Grassmann number. Of course the Grassmannian part of the integration can be performed trivially, and in virtue of (3.7), (3.10), and (3.13) we have

$$
\langle x'',\zeta'';\lambda|x',\zeta';0\rangle=K_0(x'',x';\lambda)\langle\zeta''|\zeta'\rangle\,,
\tag{3.21}
$$

where $K_0(x'',x';\lambda)$ is the proper time evolution kernel of the Klein–Gordon equation. Note that the kernel

$$
\langle\zeta''|\zeta'\rangle=(\zeta''^{\mathrm{I}}-\zeta'^{\mathrm{I}})(1-\zeta''^{\mathrm{II}}\zeta'^{\mathrm{II}*})\,,
\tag{3.22}
$$

corresponds exactly to the unit matrix in the Dirac representation via (3.8) and that we therefore indeed obtain the iterated Dirac propagator by combining (3.20) and (3.1).

In order to obtain the Dirac propagator itself we need the matrix element

$$
\langle x'',\zeta'';\lambda,\theta|x',\zeta';0,0\rangle\equiv\langle x'',\zeta''|e^{i\hat{H}\lambda+\theta\hat{Q}}|x',\zeta'\rangle\,.
\tag{3.23}
$$

Since $\hat{Q}$ is a constant of motion we can represent this matrix element by an expression like (3.10) with the only difference that *one* of the factors $\langle x_{k+1},\zeta_{k+1}|e^{i\hat{H}\epsilon}|x_k,\zeta_k\rangle$ has to be replaced by

$$
\langle x_{k+1},\zeta_{k+1}|e^{i\hat{H}\epsilon+\theta\hat{Q}}|x_k,\zeta_k\rangle=\langle x_{k+1}|e^{i\hat{H}\epsilon}|x_k\rangle\langle\zeta_{k+1}|e^{\theta\hat{Q}}|\zeta_k\rangle\,.
\tag{3.24}
$$

The supercharge of a free particle is

$$
Q=p_a\xi^a=p_A\zeta^A+p^A\beta_A\,,
\tag{3.25}
$$

where we have introduced

$$
p_{\mathrm{I}}=2^{-1/2}(p_0+p_3),\qquad p_{\mathrm{II}}=2^{-1/2}(p_1-ip_2)\,,
\tag{3.26}
$$

$$
p^{\mathrm{I}}=2^{-1/2}(p_0-p_3),\qquad p^{\mathrm{II}}=2^{-1/2}(-p_1-ip_2)\,,
\tag{3.27}
$$

and a summation over $A=$ I, II is understood. We have

$$
\langle\zeta_{k+1}|e^{\theta\hat{Q}}|\zeta_k\rangle=\langle\zeta_{k+1}|\zeta_k\rangle+\theta(p_{\mathrm{I}}\zeta_k^{\mathrm{I}}\zeta_{k+1}^{\mathrm{I}}+p^{\mathrm{I}})(1-\zeta_{k+1}^{\mathrm{II}}\zeta_k^{\mathrm{II}*})-\theta(\zeta_{k+1}^{\mathrm{I}}-\zeta_k^{\mathrm{I}})(p_{\mathrm{II}}\zeta_{k+1}^{\mathrm{II}}-p^{\mathrm{II}}\zeta_k^{\mathrm{II}*})
\tag{3.28}
$$

$$
\begin{aligned}
&=-\int d\overline{G}(\beta)\exp[-\beta_{\mathrm{I}}(\zeta_{k+1}^{\mathrm{I}}-\zeta_k^{\mathrm{I}})+\zeta_{k+1}^{\mathrm{II}}\beta_{\mathrm{II}}+\beta_{\mathrm{II}}^{*}\zeta_k^{\mathrm{II}*}]\\
&\quad\times\exp[\theta(p_{\mathrm{I}}\zeta_k^{\mathrm{I}}-p^{\mathrm{I}}\beta_{\mathrm{I}}+p^{\mathrm{II}}\beta_{\mathrm{II}}-p_{\mathrm{II}}\beta_{\mathrm{II}}^{*})]\,.
\end{aligned}
\tag{3.29}
$$

Now

$$
p_{\mathrm{I}}\zeta^{\mathrm{I}}-p^{\mathrm{I}}\beta_{\mathrm{I}}+p^{\mathrm{II}}\beta_{\mathrm{II}}-p_{\mathrm{II}}\beta_{\mathrm{II}}^{*}=p_0\xi^3+p_3\xi^0+p_1\xi^1+p_2\xi^2=:\widetilde{Q}(p,\zeta,\beta)\,.
\tag{3.30}
$$

Therefore

$$
\begin{aligned}
\langle x_{k+1},\zeta_{k+1}|e^{i\hat{H}\epsilon+\theta\hat{Q}}|x_k,\zeta_k\rangle=&\int\frac{d^4p}{(2\pi)^4}\exp\left[ip(x_k-x_{k+1})+i\frac{p^2}{2}\epsilon\right]\langle\zeta_{k+1}|e^{\theta\hat{Q}}|\zeta_k\rangle=\frac{1}{2\pi^2\epsilon^2}\int d\overline{G}(\beta_{k+1/2})\\
&\times\exp[-\beta_{\mathrm{I},k+1}(\zeta_{k+1}^{\mathrm{I}}-\zeta_k^{\mathrm{I}})+\zeta_{k+1}^{\mathrm{II}}\beta_{\mathrm{II},k+1/2}+\beta_{\mathrm{II},k+1/2}^{*}\zeta_k^{*\mathrm{II}}]\\
&\times\exp\left[-\frac{i}{2\epsilon}(x_{k+1}-x_k)^2-\theta\widetilde{Q}\left(\frac{x_{k+1}-x_k}{\epsilon},\zeta_k,\beta_{k+1/2}\right)\right],
\end{aligned}
\tag{3.31}
$$

and

$$\langle x'', \zeta''; \lambda, \theta | x', \zeta'; 0, 0\rangle \sim \int \mathscr{D}[x]\, \mathscr{D}[\xi] \exp\!\left(-\theta \tilde{Q}[\dot{x}(\lambda_0), \xi(\lambda_0)] - i \int_0^\lambda L\, d\lambda'\right). \tag{3.32}$$

The last expression is formal; λ_0 is an arbitrary value of the evolution parameter with $0 < \lambda_0 < \lambda_1$. Apparently the expression (3.32) is noncovariant. The rigorous result (3.20), however, with *one* summand of $\Sigma^n_{k=0}$ replaced according to (3.31), is covariant because the noncovariance of the factor-ordering of $\theta\tilde{Q} = \theta Q_+ + Q_-\theta$, $Q_+ + Q_- = Q$, is compensated by the noncovariance of the measure $dG(\zeta_k)d\overline{G}(\beta_{k+1/2})dG(\zeta_{k+1})$. Integrating the matrix element (3.23) over λ and θ according to (3.2) yields the desired path integral representation of the free Dirac propagator.

The results obtained so far can readily be generalized to the case of a Dirac particle minimally coupled to an external electromagnetic field. In this case the Lagrangian and Hamiltonian read

$$L = (\dot{x}^2/2) + eA\dot{x} - (i/2)(\xi\dot{\xi} - eF_{ab}\xi^a\xi^b), \tag{3.33}$$

and

$$H = \tfrac{1}{2}(p - eA)^2 - (ie/2)F_{ab}\xi^a\xi^b, \tag{3.34}$$

respectively. The path integral expression (3.20) generalizes to

$$\begin{aligned}\langle x'', \zeta''; \lambda | x', \zeta'; 0\rangle = -\lim_{n\to\infty} \int \left(\prod_{k=1}^n \frac{d^4x_k}{2\pi^2\epsilon^2}\right) d\overline{G}(\beta_{1/2}) \left[\prod_{k=1}^n dG(\zeta_k)d\overline{G}(\beta_{k+1/2})\right] \\ \times \exp\left\{\sum_{k=0}^n \left[-\frac{i}{2\epsilon}(x_{k+1} - x_k)^2 - \frac{ie}{2}(A(x_k) + A(x_{k+1}))(x_{k+1} - x_k) - \beta_{I,k+1/2}(\zeta^{\mathrm{I}}_{k+1} - \zeta^{\mathrm{I}}_k) \right.\right. \\ \left.\left. + \zeta^{\mathrm{II}}_{k+1}\beta_{\mathrm{II},k+1/2} + \beta^*_{\mathrm{II},k+1/2}\zeta^{\mathrm{II}*}_k - i\frac{\epsilon}{2}eF_{ab}(x_k)\cdot S^{ab}(\zeta_k, \zeta_{k+1}, \beta_{k+1/2})\right]\right\},\end{aligned} \tag{3.35}$$

where

$$S^{03}(\zeta, \zeta', \beta) = (i/2)(\beta_I\zeta'^{\mathrm{I}} - \zeta^I\beta_I), \tag{3.36}$$

$$S^{12}(\zeta, \zeta', \beta) = -(1/2)(\beta_{\mathrm{II}}\zeta'^{\mathrm{II}} + \zeta^{\mathrm{II}*}\beta^*_{\mathrm{II}}), \tag{3.37}$$

$$S^{01}(\zeta, \zeta', \beta) = (i/2)[\beta_{\mathrm{I}}(\zeta'^{\mathrm{II}} + \zeta^{\mathrm{II}*}) - \zeta'^{\mathrm{I}}\beta^*_{\mathrm{II}} - \zeta^{\mathrm{I}}\beta_{\mathrm{II}}], \tag{3.38}$$

$$S^{02}(\zeta, \zeta', \beta) = \tfrac{1}{2}[\beta_{\mathrm{I}}(\zeta'^{\mathrm{II}} - \zeta^{\mathrm{II}*}) - \zeta'^{\mathrm{I}}\beta^*_{\mathrm{II}} + \zeta^{\mathrm{I}}\beta_{\mathrm{II}}], \tag{3.39}$$

$$S^{13}(\zeta, \zeta', \beta) = (i/2)[\beta_{\mathrm{I}}(\zeta'^{\mathrm{II}} + \zeta^{\mathrm{II}*}) + \zeta'^{\mathrm{I}}\beta^*_{\mathrm{II}} + \zeta^{\mathrm{I}}\beta_{\mathrm{II}}], \tag{3.40}$$

$$S^{23}(\zeta, \zeta', \beta) = (i/2)[\beta_{\mathrm{I}}(\zeta'^{\mathrm{II}} - \zeta^{\mathrm{II}*}) + \zeta'^{\mathrm{I}}\beta^*_{\mathrm{II}} - \zeta^{\mathrm{I}}\beta_{\mathrm{II}}]. \tag{3.41}$$

If we now assume again that $\xi_k = \xi(k\epsilon)$, with $\xi(\lambda)$ differentiable, then

$$S^{ab}(\zeta_k, \zeta_{k+1}, \beta_{k+1/2}) = i\xi^a\xi^b + O(\epsilon), \tag{3.42}$$

and therefore the formal correspondence (3.19) remains intact.

The Berezin integration in (3.35) can be carried out explicitly using the completeness relation (3.7). With the help of (3.8) we obtain an ordinary path integral expression for the spinorial matrix elements $K_{\alpha\beta}(x'', x'; \lambda)$ of the evolution operator $e^{i\hat{H}\lambda}$:

$$\begin{aligned}K_{\alpha\beta}(x'', x'; \lambda) = -\lim_{n\to\infty}\int \prod_{k=1}^n \frac{d^4x_k}{2\pi^2\epsilon^2}\left[\prod_{k=1}^n \exp\!\left(-i\epsilon\frac{e}{2}F_{ab}(x_k)\sigma^{ab}\right)_{\alpha\beta}\right] \\ \times \exp\left\{\sum_{k=0}^{\infty}\left[-\frac{i}{2\epsilon}(x_{k+1} - x_k)^2 - i\frac{e}{2}(A(x_k) + A(x_{k+1}))(x_{k+1} - x_k)\right]\right\}\end{aligned} \tag{3.43}$$

$$\sim \int \mathscr{D}[x]\left\{T\exp\left[-i\frac{e}{2}\int_0^\lambda d\lambda'\, F_{ab}(x(\lambda'))\sigma^{ab}\right]\right\}_{\alpha\beta} \exp\left[-\frac{i}{2}\int_0^\lambda d\lambda'(\dot{x}^2 + 2e\dot{x}A)\right]. \tag{3.44}$$

Here $\sigma^{ab} = i\gamma^{[a}\gamma^{b]}$, and T denotes chronological ordering (with respect to λ') of a product of matrices.

The modification of (3.35) that is necessary to yield the kernel $\langle x'', \zeta''; \lambda, \theta | x', \zeta'; 0, 0\rangle$ required for the minimally coupled Dirac propagator is very similar to the modification embodied in Eq. (3.31) for the free particle. Since the supercharge in the present case is

$$Q = (p - eA)\xi, \tag{3.45}$$

we conclude that one simply has to add

$$-\theta\tilde{Q}\left(\frac{x_{k+1} - x_k}{\epsilon}, \zeta_k, \beta_{k+1/2}\right)$$

(for *one* value of k) to the exponent $\{\cdots\}$ appearing on the rhs of (3.35). Note that Q as a function of $\dot{x}$ and ξ is the same here as in the free-particle case.

IV. ON THE EXISTENCE OF A NICOLAI MAPPING

The possibility of integrating the anticommuting variables out in the path integral (3.35) raises the question of whether a Nicolai mapping[16] exists for the type of supersymmetry present in the Dirac particle model under consideration. The question would be answered in the affirmative if some combination of the matrix elements (3.43) could be represented by a *Gaussian* path integral. Naively one would expect this combination to be

$$\lim_{\lambda\to\infty}\int d^4x'\,d^4x''\,\mathrm{Tr}\,K(x'',x';\lambda)\,, \tag{4.1}$$

but it is apparent from (3.43) that neither this nor any other combination of matrix elements has the desired property in the presence of an external electromagnetic field, nor does (4.1) have a neat expression in terms of the ζ and β variables. It seems rather that one has to inspect every matrix element $K_{\alpha\beta}$ separately with respect to the existence of a Nicolai mapping. A similar observation was made recently[17] in a rigorous investigation of Witten's supersymmetric model. In our case this observation may be explained by the existence of a spontaneous breaking of Lorentz symmetry: There is no quantum state that corresponds to the invariant classical solution $\xi^a(\lambda)=0$ for the spin vector. Hence all the matrix elements

$$\langle\mathrm{out}|\mathrm{in}\rangle_{\alpha\delta} := \langle p''=0,\alpha;\,\infty|p'=0,\delta;-\infty\rangle \tag{4.2}$$

$$= \int d^4x'\,d^4x''\,[\gamma^0K(x'',x';\,\infty)]_{\alpha\delta} \tag{4.3}$$

may be considered as "vacuum persistence amplitudes." In (4.2) we have introduced in an obvious notation the four spin states $|\alpha\rangle$ corresponding to the pseudo-wave functions $f(\zeta)=1,\zeta^{\mathrm{I}},\zeta^{\mathrm{II}}$ and $\zeta^{\mathrm{I}}\zeta^{\mathrm{II}}$, for $\alpha=1, 2, 3$, and 4, respectively [cf. the remarks following Eq. (2.12)]. Equation (4.3) follows from (4.2) by virtue of (2.16) and (2.17). The amplitudes (4.2) will actually be divergent unless they are zero, on the other hand

$$K(x'',x';\lambda)\overset{\lambda\to\infty}{\to}0\,. \tag{4.4}$$

It is instructive to represent the amplitudes (4.2) by Grassmannian path integrals. These representations are implied by (3.35) and the following relations:

$$|1\rangle = \int|\zeta\rangle dG(\zeta)\,, \tag{4.5}$$

$$|2\rangle = |\zeta=0\rangle\,, \tag{4.6}$$

$$|3\rangle = |\beta=0\rangle\,, \tag{4.7}$$

$$|4\rangle = \int|\beta\rangle d\overline{G}(\beta)\,, \tag{4.8}$$

where the state $|\beta\rangle = |\beta_{\mathrm{I}},\beta^*_{\mathrm{II}}\rangle$ is represented by

$$f_\beta(\zeta) = e^{\zeta^{\mathrm{I}}\beta_{\mathrm{I}}}(\zeta^{\mathrm{II}}+\beta^*_{\mathrm{II}})\,. \tag{4.9}$$

Thus, e.g.,

$$[\gamma^0K(x'',x';\lambda)]_{14} = \int\int dG(\zeta'')\langle x'',\zeta'';\lambda\,|x',\beta';0\rangle d\overline{G}(\beta')\,. \tag{4.10}$$

In the free particle case, the Berezin integrations are trivial, yielding just the factor $\gamma^0_{\alpha\delta}$ appearing on the rhs of

$$[\gamma^0K(x'',x';\lambda)]_{\alpha\delta} = \gamma^0_{\alpha\delta}K_0(x'',x';\lambda)\,, \tag{4.11}$$

and we are left with the standard path integral representation for the scalar propagator K_0. From this, it may be concluded that a Nicolai mapping exists and that it is simply the identity, but apparently no such conclusion is possible in the case of the electromagnetic coupling due to the more complicated structure of the rhs of Eq. (3.43).

It is amusing to observe that the more formal manipulations usually adopted in field theory give a different result. In the case of a *constant* external electromagnetic field, it may be written formally

$$\langle\mathrm{out}|\mathrm{in}\rangle\sim\int\mathscr{D}\,[x]\,\mathscr{D}\,[\xi\,]\times\exp\biggl[-\frac{i}{2}\int_{-\infty}^{\infty}(\dot{x}^2+eA\dot{x})d\lambda\biggr]\times\exp\biggl[-\frac{1}{2}\int_{-\infty}^{\infty}\xi\Bigl(\frac{d}{d\lambda}-eF\Bigr)\xi\,d\lambda\biggr]. \tag{4.12}$$

The integration over the Grassmann variables yields formally

$$\biggl[\det\Bigl(\frac{d}{d\lambda}-eF\Bigr)\biggr]^{1/2} = \det\Bigl(\frac{d}{d\lambda}-eF\Bigr)^{1/2}, \tag{4.13}$$

which is the Jacobi determinant of the linear nonlocal transformation

$$x\to y[x] = \Bigl(\frac{d}{d\lambda}-eF\Bigr)^{1/2}x\,. \tag{4.14}$$

This is the Nicolai mapping obtained by formal considerations, resulting in the formal expression

$$\langle\mathrm{out}|\mathrm{in}\rangle\sim\int\mathscr{D}\,[y]\exp\biggl[\int_{-\infty}^{\infty}y^*\frac{d}{d\lambda}y\,d\lambda\biggr]. \tag{4.15}$$

Apparently it is not possible to construct, even perturbatively, a Nicolai mapping in the case of a generic (i.e., nonconstant) external field, because in general the electromagnetic field tensor $F^a_{\ b}(x)$ is not a Jacobian.

[1] F. A. Berezin and M. S. Marinov, Ann. Phys. (NY) **104**, 336 (1977).

[2] H. Rumpf, Gen. Relativ. Gravit. **14**, 773 (1982).

[3] P. Di Vecchia and F. Ravndal, Phys. Lett. A **73**, 371 (1979); F. Ravndal, Phys. Rev. D **10**, 2823 (1980).

[4] A. O. Barut and N. Zanghi, Phys. Rev. Lett. **52**, 2009 (1984).

[5] A. O. Barut and I. H. Duru, Phys. Rev. Lett. **53**, 2355 (1984).

[6] G. Gaveau, T. Jacobson, M. Kac, and L. S. Schulman, Phys. Rev. Lett. **53**, 419 (1984).

[7] T. Jacobson, "Feynman's checkerboard and other games," University of California at Santa Barbara, preprint, 1984.

[8] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw–Hill, New York, 1965).

[9] J. R. Klauder, Ann. Phys.(NY) **11**, 123 (1960).

[10] E. Witten, Nucl. Phys. B **188**, 513 (1981).

[11] H. Rumpf, "Supersymmetric vector particles," University of Vienna preprint, in preparation.

[12] P. Casalbuono, Nuovo Cimento A **33**, 133 (1976); **34**, 384 (1976).

[13] E. Witten, Nucl. Phys. B **202**, 253 (1982).

[14] L. Alvarez-Gaumé, Commun. Math. Phys. **90**, 161 (1983).

[15] A. T. Ogielski and J. Sobczyk, J. Math. Phys. **22**, 2060 (1981).

[16] H. Nicolai, Phys. Lett. **89** B, 341 (1980); Nucl. Phys. B **176**, 419 (1980).

[17] H. Ezawa and J. R. Klauder, Progr. Theor. Phys. **74**, 904 (1985).

A notion of duality for twistor functions

Michael Eastwood
Mathematical Institute, St Giles', Oxford OX1 3LB, England
Richard Jozsa
School of Mathematics, University of New South Wales, P.O. Box 1, Kensington, NSW 2033, Australia

(Received 15 October 1985; accepted for publication 30 January 1986)

A class of functions on the primed spin bundle of complexified Minkowski space, defined by $\nabla_{AA'}\partial f/\partial\pi_{A'} = 0$, is introduced and it is shown that these functions bear a close relationship to twistor functions, especially through their use in describing massless fields.

I. INTRODUCTION

The basic constructions in twistor theory[1-3] treat right-handed and left-handed fields separately. If a construction for right-handed fields is carried out in twistor space, then the same procedure in dual twistor space will provide the corresponding result for left-handed fields. In order to describe the mixed case of fields that are neither right-handed nor left-handed it is desirable to be able to treat both cases in terms of constructions on the same underlying space. These considerations lead naturally to a dual notion of twistor functions and provide the motivation for this paper.

II. THEORY

Let F be the projective primed spin bundle over complexified Minkowski space M and introduce coordinates $(x^{AA'},\pi_{A'})$ on F. (We use the two-component spinor notation[4] freely and, generally, follow standard twistor terminology[2].) The variable $\pi_{A'}$ is homogeneous (i.e., $\pi_{A'}$ and $\lambda\pi_{A'}$ refer to the same point for any nonzero scalar λ) and a function on F is said to be homogeneous of degree k if and only if

$$\pi_{A'}\frac{\partial f}{\partial\pi_{A'}} = kf. \tag{1}$$

Twistor functions may be defined as local holomorphic functions on F satisfying

$$\pi^{A'}\nabla_{AA'}f = 0. \tag{2}$$

Here F may be factored by the integral surfaces of $\pi^{A'}\nabla_{AA'}$ to obtain twistor space and then functions on F satisfying (2) are equivalent to unrestricted functions on this twistor space. The integral surfaces are given explicitly by $x^{AA'}\pi_{A'} =$ const so that a twistor function $f(x^{AA'},\pi_{A'})$ depends on $x^{AA'}$ only through $\omega^A \equiv x^{AA'}\pi_{A'}$, whence $(\omega^A,\pi_{A'})$ provide coordinates in twistor space.

Definition[5]: An antitwistor function f homogeneous of degree k is a holomorphic function on F homogeneous of degree k satisfying

$$\nabla_{AA'}\frac{\partial f}{\partial\pi_{A'}} = 0. \tag{3}$$

If $k=0$ then we also require $\Box f = 0$ for $\Box \equiv \nabla^{AA'}\nabla_{AA'}$ [which already follows from (1) and (3) for $k \neq 0$].

Note that antitwistor functions cannot be represented as unrestricted functions on any factor space of F, i.e., "antitwistor space" does not exist. (This follows since the product of two antitwistor functions is not necessarily antitwistor.) Though $\nabla_{AA'}\partial/\partial\pi_{A'}$ is a second-order operator, it is only first order on twistor functions as follows.

Proposition: (a) If t is a twistor function homogeneous of degree k then $\nabla_{AA'}\partial t/\partial\pi_{A'}$ is a twistor function homogeneous of degree $k-1$ given by $(k+1)\partial t/\partial\omega^A$.

(b) If a is an antitwistor function homogeneous of degree k then $\pi^{A'}\nabla_{AA'}a$ is an antitwistor function of degree $k+1$.

Proof: By direct computation. For (a) note that for any twistor function t, $\nabla_{AA'}t = \pi_{A'}\partial t/\partial\omega^A$. □

We next introduce some sheaves and an operator S required in the proof of the main results. In all cases, the local sections of the sheaves are holomorphic functions with further restrictions given below.

Sheaves on M: $\mathscr{M}^A_{(B'\cdots D')}$ is the sheaf of functions $\phi^A_{(B'\cdots D')}(x)$ taking values in the indicated spin space (and similarly for other possible spinor indices); $\mathscr{Z}_{A'B'\cdots D'}$ is the sheaf of right-handed massless fields $(\nabla^{AA'}\phi_{A'B'\cdots D'} = 0)$; $\mathscr{Z}_{AB\cdots D}$ is the sheaf of left-handed massless fields $(\nabla^{AA'}\phi_{AB\cdots D} = 0)$; $\mathscr{L}^{A'B'\cdots D'}$ is the sheaf of solutions of the multitwistor equation

$$\nabla^{E(E'}\phi^{A'B'\cdots D')} = 0 \quad \text{for } \phi^{A'B'\cdots D'} = \phi^{(A'B'\cdots D')}.$$

Sheaves on F: $\mathscr{O}(k)$ is the sheaf of functions homogeneous of degree k; $\mathscr{O}^A(k)$ is the sheaf of functions $f^A(x,\pi)$ taking values in the indicated spin space and homogeneous of degree k; $\mathscr{T}(k)$ is the sheaf of twistor functions homogeneous of degree k; $\mathscr{A}(k)$ is the sheaf of antitwistor functions homogeneous of degree k. The operator S: $\mathscr{O}(k)\rightarrow\mathscr{O}(-k-2)$ for $k \geqslant -1$ is defined as follows. If $f\in\mathscr{O}(k)$ then the k-fold derivative $\partial^k f/\partial\pi_{A'}\cdots\partial\pi_{D'}$ is homogeneous of degree 0, whence

$$\pi^{E'}\frac{\partial^k f}{\partial\pi_{A'}\cdots\partial\pi_{D'}\,\partial\pi_{E'}} = 0, \tag{4}$$

and, since $\partial^k f/\partial\pi_{A'}\cdots\partial\pi_{E'}$ is symmetric in its spinor indices, (4) gives

$$\frac{\partial^k f}{\partial\pi_{A'}\cdots\partial\pi_{D'}\partial\pi_{E'}} = \pi^{A'}\cdots\pi^{D'}\pi^{E'}g(x^{AA'},\pi_{A'}), \tag{5}$$

with g homogeneous of degree $-k-2$. We set $Sf \equiv g$.

Lemma:

(a) $\pi^{A'}(Sf) = S(\partial f/\partial\pi_{A'})$.

(b) $\partial(Sf)/\partial\pi_{A'} = -S(\pi^{A'}f)$.

0022-2488/86/061655-03$02.50

(c) There is an exact sequence

$$0\to\mathscr{M}^{A'B'\cdots D')}\xrightarrow{\pi_{A'}\pi_{B'}\cdots\pi_{D'}}\mathcal{O}(k)\xrightarrow{S}\mathcal{O}(-k-2)\to 0.$$

Proof: (a) and (b) may be derived directly from the relation (5). (a) follows easily but (b) requires a messy calculation. An easier way to obtain (b) is to note that, using Cauchy's integral formula, one may write

$$Sf(x^{AA'},\xi_{A'})=\frac{(k+1)!}{2\pi i}\oint\frac{f(x^{AA'},\pi_{A'})}{(\xi^{A'}\pi_{A'})^{k+2}}\,\pi^{B'}\,d\pi_{B'},$$

where the contour surrounds the pole at $\pi_{A'}=\xi_{A'}$. (b) follows immediately by differentiating the integrand. The kernel and image of S are easily identified using (5). □

Note that S acts only on the $\pi_{A'}$ variable (and is essentially the operator edh[6]). By introducing $\nabla_{AA'}$ on each side of (a) and (b) in the lemma, the following squares commute:

$$\begin{tikzcd}
\mathcal{O}(k) \arrow[r, "\nabla_{AA'}\,\partial/\partial\pi_{A'}"] \arrow[d, "S"'] & \mathcal{O}_A(k-1) \arrow[d, "S"'] \\
\mathcal{O}(-k-2) \arrow[r, "\pi^{A'}\nabla_{AA'}"] & \mathcal{O}_A(-k-1)
\end{tikzcd}, \quad (6)$$

$$\begin{tikzcd}
\mathcal{O}(k) \arrow[r, "\pi^{A'}\nabla_{AA'}"] \arrow[d, "S"'] & \mathcal{O}_A(k+1) \arrow[d, "S"'] \\
\mathcal{O}(-k-2) \arrow[r, "-\nabla_{AA'}\,\partial/\partial\pi_{A'}"] & \mathcal{O}_A(-k-3)
\end{tikzcd}. \quad (7)$$

Theorem 1: Let $k\geqslant 1$.

(a) The sequence (with k indices $A'\cdots D'$)

$$0\to\mathscr{Z}_{A'\cdots D'}\xrightarrow{\pi_{A'}\cdots\pi_{D'}}\mathscr{A}(k)\xrightarrow{S}\mathscr{T}(-k-2)\to 0$$

is exact.

(b) $\mathscr{A}(k)$ admits a resolution

$$0\to\mathscr{A}(k)\to\mathcal{O}(k)$$
$$\xrightarrow{\nabla_{AA'}\,\partial/\partial\pi_{A'}}\mathcal{O}_A(k-1)\xrightarrow{\nabla^A_{B'}\,\partial/\partial\pi_{B'}}\mathcal{O}(k-2)\to 0.$$

In other words, this sequence is exact.

Proof: Consider the following diagram:

$$\begin{tikzcd}
 & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & \\
0 \arrow[r] & \mathscr{Z}_{A'B'\cdots D'} \arrow[r] \arrow[d, "\pi^{A'}\cdots\pi^{D'}"] & \mathscr{M}_{(A'B'\cdots D')} \arrow[r, "k\nabla^{A'}_{A}"] \arrow[d, "\pi^{A'}\cdots\pi^{D'}"] & \mathscr{M}_{A(B'C'\cdots D')} \arrow[r, "(k-1)\nabla^{AB'}"] \arrow[d, "\pi^{B'}\cdots\pi^{D'}"] & \mathscr{M}_{(C'\cdots D')} \arrow[r] \arrow[d, "\pi^{C'}\cdots\pi^{D'}"] & 0 \\
0 \arrow[r] & \mathscr{A}(k) \arrow[r] \arrow[d, "S"] & \mathcal{O}(k) \arrow[r, "\nabla_{AA'}\,\partial/\partial\pi_{A'}"] \arrow[d, "S"] & \mathcal{O}_A(k-1) \arrow[r, "\nabla^A_{B'}\,\partial/\partial\pi_{B'}"] \arrow[d, "S"] & \mathcal{O}(k-2) \arrow[r] \arrow[d, "S"] & 0. \\
0 \arrow[r] & \mathscr{T}(-k-2) \arrow[r] \arrow[d] & \mathcal{O}(-k-2) \arrow[r, "\pi^{A'}\nabla_{AA'}"] \arrow[d] & \mathcal{O}_A(-k-1) \arrow[r, "\pi^{B'}\nabla^A_{B'}"] \arrow[d] & \mathcal{O}(-k) \arrow[r] \arrow[d] & 0 \\
 & 0 & 0 & 0 & 0 &
\end{tikzcd}$$

Using (6), the diagram is easily seen to commute. By the lemma (c) all columns save for the first are exact and this conceivably fails only at $\mathscr{A}(k)$ and $\mathscr{T}(-k-2)$. Similarly[3] only the middle row may fail to be exact at $\mathcal{O}_A(k-1)$ and $\mathcal{O}(k-2)$. By standard diagram chasing arguments[7] it follows that the diagram is exact throughout. □

Theorem 2: Let $k\geqslant -1$.

(a) The sequence (with k indices $C'\cdots D'$)

$$0\to\mathscr{L}^{C'\cdots D'}\xrightarrow{\pi_{C'}\cdots\pi_{D'}}\mathscr{T}(k)\xrightarrow{S}\mathscr{A}(-k-2)\xrightarrow{\pi^{A'}\nabla_{AA'}\cdots\pi^{D'}\nabla_{DD'}}\mathscr{Z}_{A\cdots D}\to 0$$

is exact. (In case $k=-1$, the term $\mathscr{L}$ is taken as 0.)

(b) $\mathscr{A}(-k-2)$ admits a resolution

$$0\to\mathscr{A}(-k-2)\to\mathcal{O}(-k-2)\xrightarrow{\nabla_{AA'}\,\partial/\partial\pi_{A'}}\mathcal{O}_A(-k-3)\xrightarrow{\nabla^A_{A'}\,\partial/\partial\pi_{A'}}\mathcal{O}(-k-4)\to 0.$$

Proof: Consider the following diagram:

$$\begin{tikzcd}
 & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & \\
0 \arrow[r] & \mathscr{L}^{C'\cdots D'} \arrow[r] \arrow[d, "\pi_{C'}\cdots\pi_{D'}"] & \mathscr{M}^{(C'\cdots D')} \arrow[r, "\nabla_A^{(B'}"] \arrow[d, "\pi_{C'}\cdots\pi_{D'}"] & \mathscr{M}_A^{(B'C'\cdots D')} \arrow[r, "\nabla^{AA'}"] \arrow[d, "\pi_{B'}\cdots\pi_{D'}"] & \mathscr{M}^{(A'B'\cdots D')} \arrow[r] \arrow[d, "\pi_{A'}\cdots\pi_{D'}"] & 0 \\
0 \arrow[r] & \mathscr{T}(k) \arrow[r] \arrow[d, "S"] & \mathcal{O}(k) \arrow[r, "-\pi^{B'}\nabla_{AB'}"] \arrow[d, "S"] & \mathcal{O}_A(k+1) \arrow[r, "-\pi^{A'}\nabla^A_{A'}"] \arrow[d, "S"] & \mathcal{O}(k+2) \arrow[r] \arrow[d, "S"] & 0 \\
0 \arrow[r] & \mathscr{A}(-k-2) \arrow[r] \arrow[d] & \mathcal{O}(-k-2) \arrow[r, "\nabla_{AB'}\,\partial/\partial\pi_{B'}"] \arrow[d] & \mathcal{O}_A(-k-3) \arrow[r, "\nabla^A_{A'}\,\partial/\partial\pi_{A'}"] \arrow[d] & \mathcal{O}(-k-4) \arrow[r] \arrow[d] & 0. \\
 & 0 & 0 & 0 & 0 &
\end{tikzcd}$$

Using (7), the diagram is easily seen to commute. By the Lemma (c) all columns save for the first are exact and this conceivably fails only at $\mathscr{T}(k)$ and $\mathscr{A}(-k-2)$. The bottom row may fail to be exact at $\mathcal{O}_A(-k-3)$ and $\mathcal{O}(-k-4)$ and the only other horizontal cohomology occurs[3] at $\mathscr{M}_A^{(B'\cdots D')}$, where one obtains the potential modulo gauge description[3] of the sheaf of left-handed fields $\mathscr{Z}_{AB\cdots D}$. A diagram chase[7] now yields the statement of the theorem. □

Remark: The resolutions given in the above theorems and, indeed, the crucial parts of the diagrams used in their proofs appear, in dual form, in certain Berstein–Gelfand–Gelfand resolutions.[8]

Corollary: Let $\nu: F \to M$ be the canonical projection and let $U \subseteq M$ be open. The cohomology of $\mathscr{A}(n)$ over $\nu^{-1}(U) \subseteq F$ is as follows.

(a) For $k \geqslant 1$, $H^0(\mathscr{A}(k)) = \Gamma(U, \mathscr{Z}_{A'\cdots D'})$ are the massless fields of helicity $k/2$, and $H^j(\mathscr{A}(k)) = 0$ for $j \neq 0$.

(b) For $k \geqslant -1$, $H^1(\mathscr{A}(-k-2)) = \Gamma(U, \mathscr{L}^{C'\cdots D'})$ are the multitwistors of order k, $H^2(\mathscr{A}(-k-2)) = \Gamma(U, \mathscr{Z}_{AB\cdots D})$ are the massless fields of helicity $-(k+2)/2$, and $H^j(\mathscr{A}(-k-2)) = 0$ for $j \neq 1,2$.

Proof: Follows from the resolutions of Theorems 1(a) and 2(b) and standard machinery.[9] □

To compare with the usual twistor description,[2,3,9] all cohomology on $\nu^{-1}(U)$:
{massless fields on U of helicity $n/2 > 0$}
$= H^0(\mathscr{A}(n)) = H^1(\mathscr{T}(-n-2))$,
{massless fields on U of helicity $-n/2 < 0$}
$= H^1(\mathscr{T}(n-2)) = H^2(\mathscr{A}(-n))$.

[1] R. Penrose, Gen. Relativ. Gravit. 7, 31 (1976).

[2] *Advances in Twistor Theory*, edited by L. P. Hughston and R. S. Ward (Pitman, New York, 1979).

[3] M. G. Eastwood, R. Penrose, and R. O. Wells, Jr., Commun. Math. Phys. 78, 305 (1981).

[4] R. Penrose and W. Rindler, *Spinors and Space-Time* (Cambridge U.P., Cambridge, 1984).

[5] First introduced in R. O. Jozsa, Ph.D. thesis, Oxford, 1981.

[6] M. G. Eastwood and K. P. Tod, Math. Proc. Camb. Philos. Soc. 92, 317 (1982).

[7] For example, S. MacLane, *Categories for the Working Mathematician* (Springer, Berlin, 1971).

[8] J. Lepowsky, J. Alg. 49, 496 (1977).

[9] For example, M. G. Eastwood, in Ref. 2, p. 72.

Partition functions of the quark–gluon black body and density of states

G. Auberson
Laboratoire de Physique Mathématique,[a] *Université des Sciences et Techniques du Languedoc, 34060 Montpellier Cedex, France*

L. Epele,[b] G. Mahoux, and F. R. A. Simão[c]
Service de Physique Théorique, C.E.N. Saclay, B.P.2, 91190 Gif-sur-Yvette, France

(Received 17 November 1985; accepted for publication 30 January 1986)

The problem of the determination of the asymptotic density of states of the quark–gluon gas is reconsidered. A general method emerges, which unifies and simplifies previous derivations that can be found in the literature. It takes due account of various constraints on the configurations of the system: colorlessness, conservation of electric and baryonic charges, zero total momentum, and possibly some residual flavor symmetry. It is general enough to accommodate any constraint associated to a compact Lie group. This is shown in full detail in the case of a direct product of SU(N) groups. Explicit examples are completely worked out.

I. INTRODUCTION

The thermodynamics of hadronic matter at high temperature has received great attention in the last years. Particularly exciting is the possible occurrence of a deconfining phase transition at finite temperature. A theoretical study of this phase transition should proceed in the framework of quantum chromodynamics. However, this is a difficult task, as long as the mechanism of confinement has not been understood. A more phenomenological but also more manageable approach, recently proposed,[1,2] utilizes the description of hadrons given by the bag model. There, the confinement is enforced by the model. In this second approach, it is essential to investigate first the high energy asymptotic density of states of a bag. The practical way of doing it is to compute first the partition function of a bag at high temperature, in which limit it is legitimate to view a bag as a black body of gluons, quarks, and antiquarks. The object of this paper is precisely to reconsider this problem. Indeed, what we want to do here is to simplify and unify previous derivations,[3–5] show how they can be extended to more complex situations, and bring out a general method. The emphasis is therefore put more on the method than on the results.

The physical system studied here is thus a plasma of gluons, quarks, and antiquarks, contained in a volume V, and maintained at a temperature $T = 1/\beta$. It resembles the standard black body of photons, in that the particles involved are free, massless, and have two spin states. It differs from it by the appearance of new quantum numbers, color, flavor, and electric and baryonic charges, associated with new conservation laws. The first of these laws, color symmetry, is exact, and furthermore, color confinement implies that the only allowed configurations of the physical system are those that are colorless (singlet states). As for the flavor symmetry, it is not exact, and various models may be contemplated. Concerning the conservation of the baryonic charge, it is taken, like the conservation of the electric one, as an exact law, and among the quantum numbers that characterize the plasma appear the electric and baryonic charges. All these constraints on the configurations of the system have to be (and can be exactly) implemented in the calculations.

Besides, one can be led to impose further internal constraints on the system. For example, if flavors are assigned to the quarks and antiquarks, some flavor group may be considered as a symmetry of the system, in which case its configurations are also characterized by conserved flavor quantum numbers.

A last and important constraint comes from the fact that a bag is viewed as a cavity with immaterial walls. This is a situation quite different from that of the ordinary black body, where the photon gas exchanges momentum with the walls. Here, this exchange is not possible, and it implies that the total momentum of the plasma is conserved, and vanishes in the center of mass system of the bag.

Let $\sigma(W,V,a)$ be the desired density of states of one bag, at energy W, where a stands for the collection of all conserved quantum numbers. It will be derived from the grand canonical partition function $Q(\beta,V,a)$ through the inverse Laplace transform

$$\sigma(W,V,a) = \frac{1}{2\pi i}\int_{\beta_0 - i\infty}^{\beta_0 + i\infty} d\beta\, e^{\beta W} Q(\beta,V,a). \qquad (1)$$

The function $Q(\beta,V,a)$ itself is defined by

$$Q(\beta,V,a) = \mathrm{Tr}\, \hat{\mathscr{P}}_a e^{-\beta \hat{H}}, \qquad (2)$$

where $\hat{H}$ is the Hamiltonian of the physical system, and where the projector $\hat{\mathscr{P}}_a$ selects those configurations that are allowed by all the above constraints.

We now turn to the construction of the projector.

First, let $\mathscr{H}$ be a Hilbert space and $\hat{U}(g)$ be a unitary representation in $\mathscr{H}$ of a compact lie group $\mathscr{G}$. Let j label the irreducible representations of $\mathscr{G}$, and let $\hat{\mathscr{P}}_j$ be the projector on the subspace of all the states that transform under the representation j. Then

$$\hat{\mathscr{P}}_j = d_j \int_{\mathscr{G}} d\mu(g) \chi_j(g)^* \hat{U}(g), \qquad (3)$$

[a] Equipe de Recherche associée au Centre National de la Recherche Scientifique.

[b] Member of CONICET, on leave of absence from the Laboratorio de Física Teórica, Universidad Nacional de la Plata, C.C.67-1900, La Plata, Argentina.

[c] Centro Brasileiro de Pesquisas Físicas, Consello Nacional de Pesquisas/CBPF, Rua Dr. Xavier Sigaud, 150, 22290 Rio de Janeiro, Brazil.

0022-2488/86/061658-09$02.50

where $d\mu(g)$ is the normalized Haar measure on $\mathcal{G}$, and d_j and $\chi_j(g)$ are, respectively, the dimension and the character of the representation j. For example, projecting out on the colorless configurations of the plasma will be performed by using this formula for the trivial representation, in which case $d_j = 1$ and $\chi_j(g) = 1$.

In the case of the group $U(1)$, associated to a charge $\hat{Q}$, which takes on integer values q, this formula becomes

$$\hat{\mathcal{P}}_q = \int_{-\pi}^{\pi} \frac{d\theta}{2\pi} e^{i\theta(\hat{Q}-q)}. \tag{4}$$

Similarly, the projector on the states of zero momentum, for a system enclosed in a volume V, is

$$\hat{\mathcal{P}} = \int_V \frac{d^3R}{V} e^{i\hat{\mathbf{P}}\cdot\mathbf{R}}, \tag{5}$$

where $\hat{\mathbf{P}}$ is the total momentum operator. (Strictly speaking, this is true only for a parallelepipedic box, and periodic boundary conditions.)

Formulas (3)–(5) provide us with the building blocks of the complete projector $\hat{\mathcal{P}}_a$.

An alternative way of dealing with a conserved charge $\hat{Q}$ is to introduce the chemical potential $i\theta$, and to define the new partition function

$$\tilde{Q}(\beta,V,b,\theta) = \mathrm{Tr}\, \hat{\mathcal{P}}_b e^{-\beta\hat{H} + i\theta\hat{Q}}, \tag{6}$$

where b stands for the collection of all conserved quantum numbers but $\hat{Q}$. Let $\tilde{\sigma}(W,V,b,\theta)$ be the inverse Laplace transform of $\tilde{Q}(\beta,V,b,\theta)$ through Eq. (1). Then obviously $\sigma(W,V,a)$ may be recovered from $\tilde{\sigma}(W,V,b,\theta)$ by

$$\sigma(W,V,a) = \int_{-\pi}^{\pi} \frac{d\theta}{2\pi} e^{-iq\theta}\tilde{\sigma}(W,V,b,\theta). \tag{7}$$

To show how things work, we have chosen two particular models. In Sec. II, SU(N) color symmetry and flavorless quarks are considered. In Sec. III, we introduce six flavors, with an SU(3) color group, and an SU(6) flavor symmetry, partially broken or not. In Appendix A, we give a very simple proof of a useful formula, previously established[4,5] with the help of the Bargmann space formalism. In Appendix B, we solve, in the case of SU(N) groups, the extremum problem that arises in the saddle point method repeatedly used when performing the group integration in the asymptotic limit. Appendix C deals with Gaussian integration on SU(N).

II. QUARK AND GLUON GAS, WITH SU(N) COLOR, WITHOUT FLAVORS

We consider a gas of free massless quarks and gluons, which transform under an SU(N) color group by, respectively, the fundamental and adjoint representations. We assume that the gas is an SU(N) singlet state, and has a "baryonic number" B (note that a quark has the baryonic number $1/N$). As previously explained, we also impose the vanishing of the total momentum. Consequently, the complete projector is

$$\hat{\mathcal{P}}_a = \int_{\mathrm{SU}(N)} d\mu(g)\hat{U}(g) \int_V \frac{d^3R}{V} e^{i\hat{\mathbf{P}}\cdot\mathbf{R}} \times \int_{-\pi}^{\pi} \frac{d\nu}{2\pi} \exp i\nu[(\hat{N}_q - \hat{N}_{\bar{q}} - NB)], \tag{8}$$

where $\hat{N}_q$ and $\hat{N}_{\bar{q}}$ are the number of quark and antiquark operators. Indeed, it will be more convenient to deal with the conservation of the baryonic number by introducing the chemical potential $i\nu$, as in Eq. (6). The relevant projector $\hat{\mathcal{P}}_b$ is thus built with only the first two factors in the right-hand side of Eq. (8).

The Hilbert space of the gas has the structure of a tensor product $\mathcal{H}_G \otimes \mathcal{H}_q \otimes \mathcal{H}_{\bar{q}}$ of three Fock spaces of gluons, quarks, and antiquarks. The trace involved in the definition (6) of the partition function decomposes into the product of three traces in the spaces $\mathcal{H}_G$, $\mathcal{H}_q$, and $\mathcal{H}_{\bar{q}}$. Thus

$$\tilde{Q}(\beta,V,\nu) = \int_{\mathrm{SU}(N)} d\mu(g) \int_V \frac{d^3R}{V} \{\mathrm{Tr}_G\, \hat{U}_G(g)\exp(-\beta\hat{H}_G + i\hat{\mathbf{P}}_G\cdot\mathbf{R})\}$$
$$\times\{\mathrm{Tr}_q\, \hat{U}_q(g)\exp(-\beta\hat{H}_q + i\hat{\mathbf{P}}_q\cdot\mathbf{R} + i\nu\hat{N}_q)\}\{\mathrm{Tr}_{\bar{q}}\, \hat{U}_{\bar{q}}(g)\exp(-\beta\hat{H}_{\bar{q}} + i\hat{\mathbf{P}}_{\bar{q}}\cdot\mathbf{R} - i\nu\hat{N}_{\bar{q}})\}. \tag{9}$$

Here, the indices G, q, and $\bar{q}$ of the various operators refer to their restrictions in $\mathcal{H}_G$, $\mathcal{H}_q$, and $\mathcal{H}_{\bar{q}}$.

The traces can be performed by using the following formula (see Appendix A):

$$\mathrm{Tr}_{\binom{B}{F}}\hat{U}(g)e^{\hat{A}} = \exp\left\{\mp\sum_\alpha \mathrm{tr}\ln[1 \mp \mathbb{R}(g)e^{A_\alpha}]\right\}. \tag{10}$$

The upper and lower indices distinguish the boson and fermion cases. Here $\hat{A}$ is a sum of one-particle operators, and the index α labels the eigenstates of $\hat{A}$, with eigenvalues A_α, in the one-particle subspace. The operators $\hat{A}$ and $\hat{U}(g)$ commute. $\mathbb{R}(g)$ is the (irreducible) representation under which transform the one-particle states, and the trace "tr" operates on this representation.

Now $\tilde{Q}(\beta,V,\nu)$ takes the form

$$\tilde{Q}(\beta,V,\nu) = \int_{\mathrm{SU}(N)} d\mu(g) \int_{V/\beta^3} \frac{d^3r}{V/\beta^3} e^{\Theta}, \tag{11}$$

where we have set $\mathbf{R} = \beta\mathbf{r}$, and

$$\Theta = \frac{2V}{(2\pi)^3}\int d^3p\{-\mathrm{tr}\ln[1 - \mathbb{R}_{\mathrm{adj}}(g)\exp[-\beta(|\mathbf{p}| - i\mathbf{p}\cdot\mathbf{r})]]$$
$$+\,\mathrm{tr}\ln[1 + \mathbb{R}_{\mathrm{fund}}(g)\exp[-\beta(|\mathbf{p}| - i\mathbf{p}\cdot\mathbf{r}) + i\nu]] + \mathrm{tr}\ln[1 + \mathbb{R}^*_{\mathrm{fund}}(g)\exp[-\beta(|\mathbf{p}| - i\mathbf{p}\cdot\mathbf{r}) - i\nu]]\}. \tag{12}$$

In this formula, the Σ_α of Eq. (10) has been replaced by the standard approximation for large volume $[2V/(2\pi)^3]\int d^3p$ (the factor 2 comes from the spin degeneracy); we recognize the energy $|\mathbf{p}|$ of massless particles and the fundamental and adjoint representations $\mathbf{R}_{\text{fund}}$ and $\mathbf{R}_{\text{adj}}$ of SU(N).

To proceed further, we expand the logarithms, perform the traces, and integrate over $\mathbf{p}$. The gluon contribution to Θ, for example, becomes

$$\Theta_G = \frac{2V}{(2\pi)^3} \sum_{k=1}^{\infty} \frac{1}{k} \chi_{\text{adj}}(g^k) \int d^3p\, e^{-k\beta(|\mathbf{p}| - i\mathbf{p}\cdot\mathbf{r})}$$

$$= \frac{2V}{\pi^2\beta^3} \frac{1}{(1+r^2)^2} \sum_{k=1}^{\infty} \frac{1}{k^4} \chi_{\text{adj}}(g^k),$$

where $\chi_{\text{adj}}(g)$ is the character of SU(N) in the adjoint representation.

Similar calculations for the quark and antiquark contributions lead to the following expression of Θ:

$$\Theta = \frac{4V}{\pi^2\beta^3} \frac{1}{(1+r^2)^2} [\mathscr{U}(g) + \mathscr{V}(g,\nu)], \qquad (13)$$

where we have defined

$$\mathscr{U}(g) = \frac{1}{2} \sum_{k=1}^{\infty} \frac{1}{k^4} \chi_{\text{adj}}(g^k), \qquad (14)$$

$$\mathscr{V}(g,\nu) = \text{Re} \sum_{k=1}^{\infty} \frac{(-)^{k-1}}{k^4} e^{ik\nu} \chi_{\text{fund}}(g^k), \qquad (15)$$

with $\chi_{\text{fund}}(g)$ the character in the fundamental representation.

As a function of g, Θ is a class function, and thus depends only on the eigenvalues $e^{i\theta_j}$ ($j = 1,\dots,N$) of g. The θ's, which vary[6] in $(-\pi,\pi)$, are restricted by $\Sigma_{j=1}^N \theta_j = 0$, mod 2π. Then

$$\chi_{\text{fund}}(g) = \sum_{i=1}^{N} e^{i\theta_i}, \qquad (16)$$

$$\chi_{\text{adj}}(g) = |\chi_{\text{fund}}(g)|^2 - 1 = N - 1 + 2\sum_{i<j}^{N} \cos(\theta_i - \theta_j). \qquad (17)$$

Now $\mathscr{U}$ and $\mathscr{V}$ take simple forms if we introduce the following functions[7]:

$$u(\theta) = \sum_{k=1}^{\infty} \frac{1}{k^4} \cos k\theta$$

$$= \frac{\pi^4}{90} - \frac{\pi^2}{12}\theta^2\left(1 - \frac{|\theta|}{2\pi}\right)^2 \quad (|\theta| < 2\pi), \qquad (18)$$

$$v(\theta) = \sum_{k=1}^{\infty} \frac{(-)^{k-1}}{k^4} \cos k\theta$$

$$= \frac{7\pi^4}{720} - \frac{\pi^2}{24}\theta^2\left(1 - \frac{\theta^2}{2\pi^2}\right) \quad (|\theta| < \pi). \qquad (19)$$

We have

$$\mathscr{U}(g) = (N-1)\frac{\pi^4}{90} + \sum_{i<j}^{N} u(\theta_i - \theta_j), \qquad (20a)$$

$$\mathscr{V}(g,\nu) = \sum_{i=1}^{N} v(\theta_i + \nu). \qquad (20b)$$

We note that the measure $d\mu(g)$ is invariant by changing g into $e^{2i\pi/N}g$, that is to say θ_i into $\theta_i + 2\pi/N$. Consequently, in view of Eq. (20), $\widetilde{Q}(\beta,V,\nu)$ is periodic in ν, with the period $2\pi/N$.

Since we are interested in the partition function at high temperature ($\beta \to 0$), we evaluate the integrals of Eq. (11) in the saddle point approximation. Thus we are faced with the problem of finding the maximum of Θ, when the θ_i's and r vary.

First, as a function of r, Θ has its maximum at $r = 0$.

Next, we note that the function $u(\theta)$ reaches its maximum at $\theta = 0$, mod 2π. This trivially implies that $\mathscr{U}(g)$ is maximal when g belongs to the center of SU(N), namely when

$$\theta_1 = \theta_2 = \cdots = \theta_N = n2\pi/N \quad (n \text{ integer}). \qquad (21)$$

As for the function $\mathscr{V}(g,\nu)$, it turns out that it is also maximal when g belongs to the center of SU(N). This result requires a proof. It is, in fact, a particular case of a general lemma proved in Appendix B, which in the present case states that, at the maximum of $\mathscr{V}(g,\nu)$, Eq. (21) holds with the integer n fixed by

$$-\pi/N < \nu + 2\pi n/N \leqslant \pi/N. \qquad (22)$$

Because of the periodicity of $\widetilde{Q}(\beta,V,\nu)$, we can restrict ν to lie in the interval $(-\pi/N, \pi/N)$, which implies $n = 0$. Finally, for these values of ν, Θ reaches its maximum once, for $\theta_1 = \theta_2 = \cdots = \theta_N = r = 0$, with the value

$$\Theta_{\max} = \frac{\pi^2 V}{6\beta^3}\left[\frac{4N^2 + 7N - 4}{30} - N\left(\frac{\nu}{\pi}\right)^2\left(1 - \frac{1}{2}\left(\frac{\nu}{\pi}\right)^2\right)\right]. \qquad (23)$$

The quadratic part of Θ around this maximum is

$$\Delta_2\Theta = -2\Theta_{\max} r^2 - \frac{V}{6\beta^3}\left[2N + 1 - 3\left(\frac{\nu}{\pi}\right)^2\right]\sum_{i=1}^{N} \theta_i^2, \qquad (24)$$

and in Eq. (11) we approximate Θ by $\Theta_{\max} + \Delta_2\Theta$.

The integral over r is elementary. As for the integral over SU(N), it takes a simple form[9] for a class function $f(\theta_1,\dots,\theta_N)$. For a Gaussian function, we prove in Appendix C the following formula:

$$\int_{\text{SU}(N)} d\mu(g) \exp\left(-\frac{C}{2}\sum_{i=1}^{N}\theta_i^2\right) \underset{C\to\infty}{\sim} \frac{C^{(1-N^2)/2}(\Pi_{j=1}^{N-1} j!)}{(2\pi)^{(N-1)/2}\sqrt{N}}. \qquad (25)$$

After straightforward calculations, we obtain the asymptotic expression of the partition function at high temperature:

$$\widetilde{Q}(\beta,V,\nu) \underset{(\beta^3/V)\to 0}{\sim} A\left[\frac{2N+1}{3} - \left(\frac{\nu}{\pi}\right)^2\right]^{(1-N^2)/2}$$

$$\times \varphi(\nu)^{-3/2}(\beta^3/V)^{(N^2+4)/2}$$

$$\times \exp\left[\frac{\pi^2}{3}\frac{V}{\beta^3}\varphi(\nu)\right], \qquad (26)$$

where

$$\varphi(\nu) = \frac{\pi^2}{60}(4N^2 + 7N - 4) - \frac{N}{2}\nu^2\left[1 - \frac{1}{2}\left(\frac{\nu}{\pi}\right)^2\right],$$

$$A = \frac{3\sqrt{3}(\Pi_{j=1}^{N-1} j!)}{\sqrt{N}(2\pi)^{1+N/2}}. \qquad (27)$$

We are now in a position to evaluate the asymptotic level density $\tilde{\sigma}(W,V,\nu)$ at large W, by the inverse Laplace transform of $\tilde{Q}(\beta,V,\nu)$. This is most easily done by using the following elementary formula:

$$\frac{1}{2i\pi}\int_{\beta_0 - i\infty}^{\beta_0 + i\infty} d\beta\, \beta^{c} e^{\beta W + d/3\beta^3} \underset{W\to\infty}{\sim} \frac{1}{2\sqrt{2\pi}} d^{(1+2c)/8} W^{-(5+2c)/8} e^{(4/3)(dW^3)^{1/4}}. \tag{28}$$

We obtain

$$\tilde{\sigma}(W,V,\nu) \underset{W\to\infty}{\sim} C\left[\frac{2N+1}{3} - \left(\frac{\nu}{\pi}\right)^2\right]^{(1-N^2)/2} \times \varphi(\nu)^{(3N^2+1)/8}(1/W)(VW^3)^{-(N^2+3)/8} \times \exp[(4/3)(\varphi(\nu)VW^3)^{1/4}], \tag{29}$$

with $C = (2\pi)^{5/2}A/16$.

Like the function $\tilde{Q}$, $\tilde{\sigma}(W,V,\nu)$ is $2\pi/N$ periodic in ν, and the asymptotic formula (29) is valid for $\nu \in (-\pi/N, \pi/N)$.

This is indeed the result relevant for the thermodynamic study of a gas of bags with nonvanishing baryonic numbers. Equivalently, one could use the true density of states $\sigma(W,V,B)$ with fixed total baryonic number B, given by

$$\sigma(W,V,B) = \int_{-\pi}^{\pi} \frac{d\nu}{2\pi} e^{-iNB\nu} \tilde{\sigma}(W,V,\nu). \tag{30}$$

Here NB plays the role of the integer parameter q in Eq. (7). Note that due to the $2\pi/N$ periodicity of $\tilde{\sigma}$ in ν, the above integral vanishes except when B itself is an integer, in which case it can be rewritten as

$$\sigma(W,V,B) = N\int_{-\pi/N}^{\pi/N} \frac{d\nu}{2\pi} e^{-iNB\nu} \tilde{\sigma}(W,V,\nu). \tag{31}$$

This was to be expected, since the only color singlet configurations have a quark number $N_q - N_{\bar{q}}$ that is a multiple of N.

In the asymptotic limit $W \to \infty$, if B remains finite, the dominant behavior of $\sigma(W,V,B)$ does not depend on B. Indeed, the only sensible asymptotic regime to look for is W, V, and B going to infinity, at fixed baryonic density $b = B/V$. The corresponding behavior of $\sigma(W,V,B)$ cannot be given in a completely explicit form, and after all, we could satisfy ourselves with Eq. (29). However, for the sake of completeness, we give the asymptotic expression of σ, which is obtained by applying again the saddle point method to the integral (31):

$$\sigma(W,V,bV) \underset{W\to\infty}{\sim} (D/W)(VW^3)^{-(N^2+4)/8} e^{\alpha(VW^3)^{1/4}}, \tag{32}$$

where

$$\alpha = \frac{4\sqrt{\pi}}{3^{3/4}} \frac{a + cz_0}{(a + 2cz_0 + cz_0^2)^{3/4}},$$

$$a = \frac{4N^2 + 7N - 4}{180}, \quad c = \frac{N}{12},$$

$$D = \frac{3\sqrt{3}(3\pi^2)^{1+3N^2/8}}{8(2\pi)^{N/2}} \left(\prod_{j=1}^{N-1} j!\right) \times \left(\frac{3}{2N+1+3z_0}\right)^{(N^2-1)/2} \times \frac{(a + 2cz_0 + cz_0^2)^{1+3N^2/8}}{[a + (3a-c)z_0 + cz_0^2]^{1/2}}, \tag{33}$$

and where z_0, a function of $|b|^{4/3} V/W$, is the unique positive solution of the equation

$$\frac{z_0^{2/3}(1+z_0)^{4/3}}{a + 2cz_0 + cz_0^2} = 3\frac{V}{W}(3\sqrt{\pi}|b|)^{4/3}. \tag{34}$$

Formulas (33) and (34) are valid as long as $|b|$ does not exceed some critical value that we did not calculate, but that is smaller than $4(W/NV)^{3/4}/3\sqrt{2\pi}$.

III. QUARK AND GLUON GAS, WITH SU(3) COLOR AND SIX FLAVORS

In this section, we repeat the calculations of the preceding section with the following changes. First, we restrict the number of colors to $N = 3$. Second, we introduce six flavors. We assume that the SU(6) flavor symmetry is broken only by the electromagnetic interactions, so that the remaining symmetry is the product of two SU(3) subgroups of SU(6). The first, $\mathrm{SU}(3)_1$, is the set of transformations that mix the quarks u, c, and t of charge $\frac{2}{3}$, the second, $\mathrm{SU}(3)_2$, mixes the quarks d, s, and b of charge $-\frac{1}{3}$.

We decide to focus our attention exclusively on the configurations that are singlet states not only of the SU(3) color group, but also of the $\mathrm{SU}(3)_1 \times \mathrm{SU}(3)_2$ flavor group. In so doing, we are aware of the fact that this is not sufficient for a thermodynamic study of a gas of bags, which certainly contains bags with any flavor variance. However, the following derivation intends to provide an example of the techniques to be used in the last stage of the above mentioned thermodynamic study.

Let us call N_1, $\bar{N}_1$, N_2, and $\bar{N}_2$ the numbers of quarks and antiquarks of charge $\frac{2}{3}$ and $-\frac{1}{3}$. If the total baryonic number B and the total electric charge q are fixed, it follows trivially that

$$B_1 \equiv N_1 - \bar{N}_1 = B + q,$$
$$B_2 \equiv N_2 - \bar{N}_2 = 2B - q.$$

Therefore, the relevant projector is

$$\hat{\mathscr{P}} = \int_{\mathrm{SU}(3)_c} d\mu(g)\hat{U}(g) \int_{\mathrm{SU}(3)_1} d\mu(g')\hat{U}(g') \times \int_{\mathrm{SU}(3)_2} d\mu(g'')\hat{U}(g'') \int_V \frac{d^3R}{V} e^{i\hat{\mathbf{P}}\cdot\mathbf{R}}. \tag{35}$$

Now, the Hilbert space has the structure of a tensor product $\mathscr{H}_G \otimes \mathscr{H}_{q_1} \otimes \mathscr{H}_{\bar{q}_1} \otimes \mathscr{H}_{q_2} \otimes \mathscr{H}_{\bar{q}_2}$. As in Sec. II, the trace in the definition (6) of the partition function decomposes into a product of traces

$$\tilde{Q}(\beta,V,\xi,\eta)=\int_{\mathrm{SU}(3)_c} d\mu(g)\int_{\mathrm{SU}(3)_1} d\mu(g')\int_{\mathrm{SU}(3)_2} d\mu(g'')\int_V \frac{d^3R}{V}$$
$$\times \mathrm{Tr}_G\{\hat{U}_G(g)\exp(-\beta\hat{H}_G+i\hat{\mathbf{P}}_G\mathbf{R})\}\mathrm{Tr}_{q_1}\{\hat{U}_{q_1}(g)\hat{U}_{q_1}(g')\exp(-\beta\hat{H}_{q_1}+i\hat{\mathbf{P}}_{q_1}\mathbf{R}+i\xi\hat{N}_1)\}$$
$$\times \mathrm{Tr}_{\bar{q}_1}\{\hat{U}_{\bar{q}_1}(g)\hat{U}_{\bar{q}_1}(g')\exp(-\beta\hat{H}_{\bar{q}_1}+i\hat{\mathbf{P}}_{\bar{q}_1}\mathbf{R}-i\xi\hat{\hat{N}}_1)\}\mathrm{Tr}_{q_2}\{\hat{U}_{q_2}(g)\hat{U}_{q_2}(g'')\exp(-\beta\hat{H}_{q_2}+i\hat{\mathbf{P}}_{q_2}\mathbf{R}+i\eta\hat{N}_2)\}$$
$$\times \mathrm{Tr}_{\bar{q}_2}\{\hat{U}_{\bar{q}_2}(g)\hat{U}_{\bar{q}_2}(g'')\exp(-\beta\hat{H}_{\bar{q}_2}+i\hat{\mathbf{P}}_{\bar{q}_2}\mathbf{R}-i\eta\hat{\hat{N}}_2)\}, \tag{36}$$

where $i\xi$ and $i\eta$ are chemical potentials associated with the conservation of B_1 and B_2. Equation (10) allows us to compute the various traces. The one in $\mathscr{H}_{q_1}$, for example, reads

$$\mathrm{Tr}_{q_1}\{\cdots\}=\exp\left\{\frac{2V}{(2\pi)^3}\int d^3p\,\mathrm{tr}\ln[1+\mathbf{R}_{\mathrm{fund}}(g)\otimes\mathbf{R}_{\mathrm{fund}}(g')\exp(-\beta|\mathbf{p}|+i\mathbf{p}\mathbf{R}+i\xi)]\right\}$$
$$=\exp\left\{\frac{2V}{\pi^2\beta^3}\frac{1}{(1+r^2)^2}\sum_{k=1}^{\infty}\frac{(-)^{k-1}}{k^4}e^{ik\xi}\chi_{\mathrm{fund}}(g^k)\chi_{\mathrm{fund}}(g'^k)\right\}. \tag{37}$$

Notice that the quark q_1 transforms under the direct product $\mathbf{R}_{\mathrm{fund}}(g)\otimes\mathbf{R}_{\mathrm{fund}}(g')$.

The other traces are calculated in the same way. Using the functions $\mathscr{U}$ and $\mathscr{V}$ defined by Eqs. (14) and (15), we finally obtain the following expression of the quantity Θ:

$$\Theta=\frac{4V}{\pi^2\beta^3}\frac{1}{(1+r^2)^2}[\mathscr{U}(g)+\mathscr{V}(g\times g',\xi)+\mathscr{V}(g\times g'',\eta)]. \tag{38}$$

Equation (20b) is replaced by

$$\mathscr{V}(g\times g',\xi)=\sum_{i,j=1}^{3}v(\theta_i+\theta'_j,\xi),\qquad \mathscr{V}(g\times g'',\eta)=\sum_{i,j=1}^{3}v(\theta_i+\theta''_j,\eta), \tag{39}$$

with obvious notations. It follows that $\tilde{Q}(\beta,V,\xi,\eta)$ is $(2\pi/3)$-periodic in both ξ and η.

Once more, we evaluate the various integrals of Eq. (36) in the saddle point approximation. A straightforward application of the lemma of Appendix B shows that, for $|\xi|$ and $|\eta|$ less than $\pi/3$, Θ reaches three times its maximum, namely when

$$\begin{aligned}&\theta_1=\theta_2=\theta_3=n(2\pi/3),\\&\theta'_1=\theta'_2=\theta'_3=-n(2\pi/3),\\&\theta''_1=\theta''_2=\theta''_3=-n(2\pi/3),\\&r=0,\end{aligned} \tag{40}$$

with $n=0$, ± 1. The maximum value of Θ is

$$\Theta_{\max}=\frac{\pi^2V}{\beta^3}\left[\frac{79}{90}-\frac{3}{2}\left(\frac{\xi}{\pi}\right)^2\left(1-\frac{1}{2}\left(\frac{\xi}{\pi}\right)^2\right)-\frac{3}{2}\left(\frac{\eta}{\pi}\right)^2\left(1-\frac{1}{2}\left(\frac{\eta}{\pi}\right)^2\right)\right], \tag{41}$$

and its quadratic part around the maximum corresponding to $n=0$ (say) is

$$\Delta_2\Theta=-2\Theta_{\max}r^2-\frac{V}{2\beta^3}\left[\left(4-3\left(\frac{\xi}{\pi}\right)^2-3\left(\frac{\eta}{\pi}\right)^2\right)\sum_{i=1}^{3}\theta_i^2+\left(1-3\left(\frac{\xi}{\pi}\right)^2\right)\sum_{i=1}^{3}\theta_i'^2+\left(1-3\left(\frac{\eta}{\pi}\right)^2\right)\sum_{i=1}^{3}\theta_i''^2\right]. \tag{42}$$

Then, proceeding as in Sec. II, we obtain the following asymptotic expression for the function $\tilde{\sigma}(W,V,\xi,\eta)$ [$(2\pi/3)$-periodic in ξ and η], valid when $|\xi|$ and $|\eta|<\pi/3$:

$$\tilde{\sigma}(W,V,\xi,\eta)\underset{W\to\infty}{\simeq}\left(\frac{3}{8\pi^2}\right)\Phi(\xi,\eta)^{19/2}\left[\left(1-3\left(\frac{\xi}{\pi}\right)^2\right)\times\left(1-3\left(\frac{\eta}{\pi}\right)^2\right)\left(4-3\left(\frac{\xi}{\pi}\right)^2-3\left(\frac{\eta}{\pi}\right)^2\right)\right]^{-4}\times(1/W)(VW^3)^{-7/2}\exp[\tfrac{4}{3}(\Phi(\xi,\eta)VW^3)^{1/4}], \tag{43}$$

where

$$\Phi(\xi,\eta)=\frac{79\pi^2}{30}-\frac{9}{2}\xi^2\left[1-\frac{1}{2}\left(\frac{\xi}{\pi}\right)^2\right]-\frac{9}{2}\eta^2\left[1-\frac{1}{2}\left(\frac{\eta}{\pi}\right)^2\right]. \tag{44}$$

The density of states $\sigma(W,V,B_1,B_2)$ now vanishes except when both integers B_1 and B_2 are multiples of 3 (implying that B and q are integers), in which case it can be written as

$$\sigma(W,V,B_1,B_2)=9\int_{-\pi/3}^{\pi/3}\frac{d\xi}{2\pi}\int_{-\pi/3}^{\pi/3}\frac{d\eta}{2\pi}e^{-i(\xi B_1+\eta B_2)}\tilde{\sigma}(W,V,\xi,\eta). \tag{45}$$

Inserting Eq. (43), we derive its asymptotic form, expressed in terms of $b_1\equiv B_1/V=$ baryonic charge density b + electric charge density ρ, and $b_2\equiv B_2/V=2b-\rho$:

$$\sigma(W,V,b_1V,b_2V)\underset{W\to\infty}{\simeq}(E/W)(VW^3)^{-15/4}\exp[\gamma(VW^3)^{1/4}], \tag{46}$$

where

$$\gamma=\frac{4}{3}\left(\frac{79\pi^2}{30}\right)^{1/4}\frac{1+\frac{135}{158}(z_1+z_2)}{\{1+\frac{135}{158}[z_1(2+z_1)+z_2(2+z_2)]\}^{3/4}},$$

$$E=\frac{9}{2^{12}\pi^3}\left(\frac{79\pi^2}{30}\right)^{41/4}\times\frac{\{1+\frac{135}{158}[z_1(2+z_1)+z_2(2+z_2)]\}^{43/4}}{[1+\frac{3}{4}(z_1+z_2)]^4[(1+3z_1)(1+3z_2)]^{9/2}}\times\left[1-\frac{135}{158}\left(\frac{z_1(1-z_1)}{1+3z_1}+\frac{z_2(1-z_2)}{1+3z_2}\right)\right]^{-1/2} \tag{47}$$

and where (z_1, z_2), function of $|b_1|^{4/3}V/W$ and $|b_2|^{4/3}V/W$, is the unique solution (with $z_{1,2} > 0$) of the coupled equations

$$\frac{z_i^{2/3}(1+z_i)^{4/3}}{\frac{158}{135} + z_1(2+z_1) + z_2(2+z_2)} = \frac{9}{4}\frac{V}{W}\left(\frac{\sqrt{\pi}}{3}|b_i|\right)^{4/3}, \quad i = 1,2. \tag{48}$$

These formulas are valid provided that $|b_1|$, $|b_2| < b_{\rm cr}$. Again, the value of $b_{\rm cr}$ has not been worked out, but is certainly less than $2\sqrt{2}(W/V)^{3/4}/\sqrt{3\pi}$.

Similar calculations in the case of a nonbroken SU(6) flavor symmetry yield the following results. The function $\tilde{\sigma}(W,V,\nu)$ involving the chemical potential $i\nu$ associated with the baryonic number B is $(\pi/3)$-periodic in ν, and for $|\nu| < \pi/6$:

$$\tilde{\sigma}(W,V,\nu) \underset{W\to\infty}{\simeq} \frac{405}{4\pi^2}\sqrt{\frac{3}{\pi}}\frac{\psi(\nu)^{133/8}}{[1-3(\nu/\pi)^2]^{35/2}[2-3(\nu/\pi)^2]^4} \times\frac{1}{W}(VW^3)^{-47/8}\exp\left[\frac{4}{3}(\psi(\nu)VW^3)^{1/4}\right], \tag{49}$$

where

$$\psi(\nu) = \tfrac{79}{30}\pi^2 - 9\nu^2[1-\tfrac{1}{2}(\nu/\pi)^2]. \tag{50}$$

The asymptotic behavior of the density of states $\sigma(W,V,B)$ at fixed baryonic charge density b is [for $|b| < b_{\rm cr} < 2^{7/4} \times (W/V)^{3/4}/3\sqrt{3\pi}$]

$$\sigma(W,V,bV) \underset{W\to\infty}{\simeq} (F/W)(VW^3)^{-6}\exp[\delta(VW^3)^{1/4}], \tag{51}$$

where

$$\delta = \frac{4}{3}\left(\frac{79\pi^2}{30}\right)^{1/4}\frac{1+\frac{135}{79}z_3}{[1+\frac{135}{79}z_3(2+z_3)]^{3/4}},$$
$$F = \frac{1215}{64\pi^3}\left(\frac{79\pi^2}{30}\right)^{17}\frac{[1+\frac{135}{79}z_3(2+z_3)]^{35/2}}{(1+\frac{3}{2}z_3)^4(1+3z_3)^{18}\sqrt{1-\frac{135}{79}[z_3(1-z_3)/(1+3z_3)]}}, \tag{52}$$

and where z_3 is the (unique) positive solution of

$$\frac{z_3^{2/3}(1+z_3)^{4/3}}{\frac{79}{135}+z_3(2+z_3)} = \frac{9}{2}\frac{V}{W}\left(\frac{\sqrt{\pi}}{2}|b|\right)^{4/3}. \tag{53}$$

By comparing Eqs. (48) and (53), we see that in the particular case $b_1 = b_2$ (which implies $b_i = 3b/2$), we have $z_1 = z_2 = z_3$, so that $\delta = \gamma$. Thus, when the quark densities of each type ("1" and "2") are equal, the argument of the exponential in the density of states is insensitive to the fact that the flavor symmetry is broken or not: it depends only on the number of colors and flavors. However, the power of the dimensionless variable (VW^3) in the preexponential factors decreases when the complete flavor symmetry is restored. This is in accordance with the general rule pointed out at the end of Sec. II.

As a last comment, let us insist on the following point. We do not pretend that the model of this section [partially broken SU(6) flavor symmetry] is particularly realistic. Rather, we have chosen it in order to show how it is possible to accommodate different types of constraints on the configurations of the physical system. Obviously, the methods presented here can be applied to a lot of various situations.

ACKNOWLEDGMENTS

We have benefited from useful discussions with A. Morel and M. L. Mehta. Two of us (L.E. and F.R.A.S.) wish to thank the Service de Physique Théorique de Saclay, and one of us (L.E.) the Laboratoire de Physique Mathématique de Montpellier, for the hospitality extended to them during the year 1984.

APPENDIX A: PROOF OF EQ. (10)

Let $\mathscr{H}_B$ (resp. $\mathscr{H}_F$) be the Fock space describing the physical states of an assembly of bosons (resp. fermions). Let $\mathscr{G}$, a (compact) Lie group, be a symmetry group of this physical system, and $\hat{U}(g)$, $g\in\mathscr{G}$ be the unitary representation of $\mathscr{G}$ in $\mathscr{H}_{\binom{B}{F}}$. Let $\hat{A}$ be an additive operator in $\mathscr{H}_{\binom{B}{F}}$ (for example, energy, number of particles, various kinds of charges,..., or any linear combination of these operators), which commutes with $\hat{U}(g)$, $\forall g\in\mathscr{G}$. We want to evaluate $\mathrm{Tr}_{\binom{B}{F}}\hat{U}(g)\,e^{\hat{A}}$.

Since $\hat{A}$ and $\hat{U}(g)$ commute, there exists in $\mathscr{H}_{\binom{B}{F}}$ a (g-dependent) basis that diagonalizes both operators. Let $|\alpha,\sigma\rangle$ be the one-particle states of such a basis, where α labels the eigenvalues of $\hat{A}$ (including a possible degeneracy besides the one associated to $\mathscr{G}$), and σ labels those $\mathbb{R}_{\sigma\sigma}(g)$ of $\hat{U}(g)$:

$$\langle\alpha',\sigma'|\hat{A}|\alpha,\sigma\rangle = \delta_{\alpha'\alpha}\delta_{\sigma'\sigma}A_\alpha, \tag{A1}$$

$$\langle\alpha',\sigma'|\hat{U}(g)|\alpha,\sigma\rangle = \delta_{\alpha'\alpha}\delta_{\sigma'\sigma}\mathbb{R}_{\sigma\sigma}(g). \tag{A2}$$

Any configuration of the system is defined by the set of occupation numbers $\{n_{\alpha,\sigma}\}$. The additivity of $\hat{A}$ simply means that

$$\langle\{n_{\alpha,\sigma}\}|\hat{A}|\{n_{\alpha,\sigma}\}\rangle = \sum_{\alpha,\sigma} n_{\alpha,\sigma}A_\alpha. \tag{A3}$$

Then, using the basis $|\{n_{\alpha,\sigma}\}\rangle$ to evaluate the trace, we readily obtain

$$\mathrm{Tr}_{\binom{B}{F}}\hat{U}(g)e^{\hat{A}} = \sum_{\{n_{\alpha,\sigma}\}}\prod_{\alpha,\sigma}[\mathbb{R}_{\sigma\sigma}(g)]^{n_{\alpha,\sigma}}e^{n_{\alpha,\sigma}A_\alpha}$$
$$= \prod_{\alpha,\sigma}\sum_{n_{\alpha,\sigma}}[\mathbb{R}_{\sigma\sigma}(g)e^{A_\alpha}]^{n_{\alpha,\sigma}}. \tag{A4}$$

In the boson case, this becomes

$$\prod_{\alpha,\sigma}\frac{1}{1-\mathbb{R}_{\sigma\sigma}(g)e^{A_\alpha}} = \prod_\alpha \det\frac{1}{1-\mathbb{R}(g)e^{A_\alpha}}, \tag{A5}$$

where the (finite-dimensional) matrix $\mathbb{R}(g)$, diagonal in the basis $|\sigma\rangle$, is nothing but the image of g in the (irreducible)

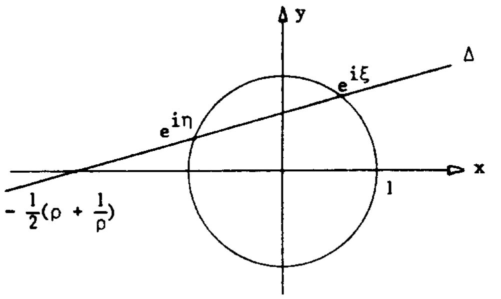

FIG. 1. Solutions of Eq. (B3).

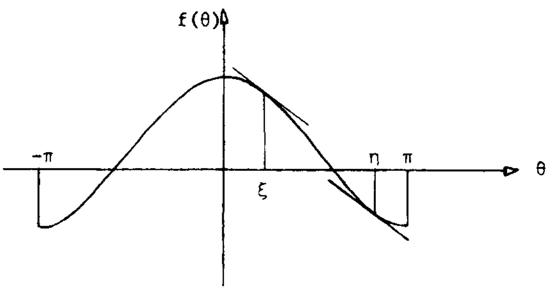

FIG. 2. Plot of the function $f(\theta)$.

representation of $\mathcal{G}$, under which transform the one-particle states.

In the fermion case, one obtains, similarly,

$$\mathrm{Tr}_F\, \hat{U}(g)e^{\hat{A}} = \prod_\alpha \det[1+\mathbb{R}(g)e^{A_\alpha}]\,. \tag{A6}$$

Finally, expressions (A5) and (A6) can be rewritten as

$$\mathrm{Tr}_{\binom{B}{F}}\hat{U}(g)e^{\hat{A}} = \exp\left\{\mp\sum_\alpha \mathrm{tr}\ln[1\mp\mathbb{R}(g)e^{A_\alpha}]\right\}, \tag{A7}$$

where the trace "tr" operates on the matrices of the one particle representations of $\mathcal{G}$. This is the announced result, Eq. (10). Notice that these formulas no longer refer to states $|\sigma\rangle$ that diagonalize $\hat{U}(g)$.

APPENDIX B: SOLVING AN EXTREMUM PROBLEM

Lemma: Let the group $\mathcal{G}$ be the direct product $\mathrm{SU}(N_1)\times\cdots\times\mathrm{SU}(N_p)$, and $\mathbb{R}(g)$ ($g\in\mathcal{G}$) be its fundamental representation. Let M be the lowest common multiple of $N_1,\ldots,N_p$. Let ρ be a positive number, and ν a real angle. Then the maximum of $|\det[1+\rho e^{i\nu}\mathbb{R}(g)]|$, when g ranges over $\mathcal{G}$, is a periodic function of ν, with period $\omega = 2\pi/M$, and it is reached on those elements $\bar{g}$ of the center of $\mathcal{G}$, defined by (i) $\bar{g} = \bar{g}_1\times\cdots\times\bar{g}_p$, $\bar{g}_i\in$center of $\mathrm{SU}(N_i)$, $i = 1,\ldots,p$, that is to say $\bar{g}_i = e^{2i\pi n_i/N_i}$, n_i integer, and

(ii) $-\dfrac{\pi}{M}<\nu+2\pi\left(\dfrac{n_1}{N_1}+\cdots+\dfrac{n_p}{N_p}\right)\leqslant\dfrac{\pi}{M}$.

Proof: (I) We first consider the case where $\mathcal{G} = \mathrm{SU}(N)$. Since the determinant of $[1+\rho e^{i\nu}\mathbb{R}(g)]$ is a class function, we can restrict ourselves to diagonal matrices $\mathbb{R}(g)$. Let $e^{i\theta_\alpha}$, $\alpha = 1,\ldots,N$, be the diagonal elements of $\mathbb{R}(g)$. The θ's are constrained by

$$\sum_{\alpha=1}^{N}\theta_\alpha = 0 \quad (\mathrm{mod}\, 2\pi). \tag{B1}$$

We thus have to look for the maximum of the function

$$|\det[1+\rho e^{i\nu}\mathbb{R}(g)]|^2 = \prod_{\alpha=1}^{N}[1+2\rho\cos(\theta_\alpha+\nu)+\rho^2], \tag{B2}$$

when the θ's vary under the constraint (B1).

We first notice that changing θ_α into $\theta_\alpha + 2\pi/N$, for all α's, preserves Eq. (B1). This is equivalent to changing ν into $\nu - 2\pi/N$ in the right-hand side of Eq. (B2). Consequently the maximum of $|\det[\cdots]|$ is a function periodic in ν, with period $\omega = 2\pi/N$.

Introducing a Lagrange parameter λ for the constraint (B1), we get the following extremum equations:

$$\cos(\theta_\alpha+\nu)+(1/\lambda)\sin(\theta_\alpha+\nu)+\tfrac{1}{2}(\rho+1/\rho)=0$$
$$(\alpha = 1,\ldots,N). \tag{B3}$$

Let Δ be the straight line of equation $x+y/\lambda+\frac{1}{2}(\rho+1/\rho)=0$ in the x, y plane (Fig. 1). It intersects the x axis at the point $x = -(\rho+1/\rho)/2\leqslant -1$. It also intersects the unit circle centered at the origin in (at most) two points $x+iy = e^{i\xi}$ and $e^{i\eta}$. We distinguish these two points by the condition $\cos\xi\geqslant\cos\eta$.

Any solution of Eqs. (B3) is such that some of the $(\theta_\alpha+\nu)$'s are equal to ξ, and the remaining ones are equal to η. For example,

$$\theta_\alpha+\nu=\begin{cases}\xi, & \alpha = 1,\ldots,P\\ \eta, & \alpha = P+1,\ldots,N\end{cases} \quad (\mathrm{mod}\, 2\pi). \tag{B4}$$

At this extremum, the constraint (B1) reads

$$P\xi+(N-P)\eta = N\nu+2\pi n \quad (n \text{ integer}). \tag{B5}$$

Here ξ and η are completely determined as functions of P and n, by Eqs. (B5) and (B6):

$$\frac{\sin\xi}{\cos\xi+\frac{1}{2}(\rho+1/\rho)} = \frac{\sin\eta}{\cos\eta+\frac{1}{2}(\rho+1/\rho)}\,. \tag{B6}$$

For convenience we will use in the following the variable $\mu = \nu+2\pi n/N$ in place of the variable n. Now, at the extremum (B4), $|\det[\cdots]|^2$ takes the value

$$F(P,\mu) = (1+2\rho\cos\xi+\rho^2)^P(1+2\rho\cos\eta+\rho^2)^{N-P}. \tag{B7}$$

We are left with the problem of finding the maximum of $F(P,\mu)$ when P and μ vary. Let us, for a moment, allow these two variables to vary continuously. Then, differentiating Eqs. (B5) and (B7) with respect to P, μ, ξ, and η, we compute the partial derivatives of $F(P,\mu)$:

$$\frac{\partial}{\partial P}\ln F(P,\mu) = \ln\frac{1+2\rho\cos\xi+\rho^2}{1+2\rho\cos\eta+\rho^2} - (\eta-\xi)\frac{\sin\xi}{\cos\xi+\frac{1}{2}(\rho+1/\rho)}, \tag{B8}$$

$$\frac{\partial}{\partial\mu}\ln F(P,\mu) = -N\frac{\sin\xi}{\cos\xi+\frac{1}{2}(\rho+1/\rho)}\,. \tag{B9}$$

Let us call $f(\theta)$ the function $\ln(1+2\rho\cos\theta+\rho^2)$. Figure (2) exhibits first its variations when θ ranges from

$-\pi$ to π, and second a couple (ξ,η). Note that according to Eq. (B6), $f'(\xi) = f'(\eta)$. Equation (B8) can now be rewritten as

$$\frac{\partial}{\partial P}\ln F(P,\mu) = f(\xi) - f(\eta) - (\xi-\eta)f'(\eta), \tag{B10}$$

and elementary geometrical considerations show that its right-hand side is positive:

$$\frac{\partial}{\partial P}\ln F(P,\mu) > 0. \tag{B11}$$

Furthermore $\partial \ln F/\partial\mu$ has the opposite sign of ξ, or equivalently the opposite sign of μ

$$\mu\frac{\partial}{\partial\mu}\ln F(P,\mu) < 0. \tag{B12}$$

Next, we remark from Eq. (B5), the right-hand side of which is equal to $N\mu$, that the variables P and μ are linearly related: When P varies from 0 to N, μ varies from η to ξ. This means that the set of points of coordinates P and μ, on which is defined the function $F(P,\mu)$, is generated by a family of segments, each segment corresponding to a possible value of the couple ξ,η. Furthermore, on such segments:

$$\frac{\partial}{\partial P}\ln F(P,\mu)\Big|_{\xi,\eta} = \ln\frac{1+2\rho\cos\xi+\rho^2}{1+2\rho\cos\eta+\rho^2} > 0. \tag{B13}$$

The reader will convince himself that the information conveyed by inequalities (B11), (B12), and (B13) is sufficient to conclude that the maximum of $|\det[\cdots]|$ is reached when the following two conditions are met: (i) $P = N$, that is to say all the $(\theta_\alpha + \nu)$'s are equal to ξ,

$$\theta_1 = \cdots = \theta_N = 2\pi n/N \quad (n \text{ integer}); \tag{B14}$$

(ii) μ is minimum in modulus, that is to say

$$-\frac{\pi}{N} < \nu + 2\pi\frac{n}{N} < \frac{\pi}{N}. \tag{B15}$$

This ends the proof of the lemma when $\mathscr{G} = \mathrm{SU}(N)$.

(II) We now consider the case where $\mathscr{G} = \mathrm{SU}(N_1)\times\cdots\times\mathrm{SU}(N_p)$. Once again we can restrict ourselves to elements $g = g_1\times\cdots\times g_p$ of $\mathscr{G}$ such that $\mathbb{R}(g) = \mathbb{R}_1(g_1)\times\cdots\times\mathbb{R}_p(g_p)$ is diagonal. Let $e^{i\theta_{\alpha j}}$ $(\alpha = 1,\ldots,N_j)$ be the diagonal elements of $\mathbb{R}_j(g_j)$, $j = 1,\ldots,p$. The constraints on the θ's read

$$\sum_{\alpha=1}^{N_j}\theta_{\alpha j} = 0 \quad (\mathrm{mod}\ 2\pi) \quad (j = 1,\ldots,p), \tag{B16}$$

and we have to look for the maximum $\mathscr{M}(\rho,\nu)$ of the function

$$\begin{aligned}&|\det[1+\rho e^{i\nu}\mathbb{R}(g)]|^2\\ &\quad = \prod_{\alpha_1=1}^{N_1}\cdots\prod_{\alpha_p=1}^{N_p}[1+2\rho\\ &\qquad \times\cos(\theta_{\alpha_1 1}+\cdots+\theta_{\alpha_p p}+\nu)+\rho^2],\end{aligned} \tag{B17}$$

when the θ's vary under the constraints (B16).

We notice that these constraints are preserved when changing $\theta_{\alpha j}$'s into $\theta_{\alpha j} + 2\pi\, n_j/N_j$ for all α and j's, where the n_j are any integers. This is equivalent to changing

$$\nu \to \nu - 2\pi\left(\frac{n_1}{N_1}+\cdots+\frac{n_p}{N_p}\right), \tag{B18}$$

in the expression (B17) of the determinant. Let M be the lowest common multiple of $N_1,\ldots,N_p$. Then $N_j = M/m_j$ $(j = 1,\ldots,p)$, where $m_1,\ldots,m_p$ are integers, the highest common factor of which is 1:

$$\mathrm{hcf}(m_1,\ldots,m_P) = 1. \tag{B19}$$

The transformation (B18) can thus be written as $\nu\to\nu - (2\pi/M)\,(m_1n_1+\cdots+m_pn_p)$. Now, it is known that when p given integers $m_1,\ldots,m_p$ $(p\geqslant 2)$ have no common factor other than 1, their linear combinations $m_1n_1+\cdots+m_pn_p$ with arbitrary integer coefficients $n_1,\ldots,n_p$ can take any given integer value n (see Ref. 10). Consequently, the transformation (B18) can finally be written as $\nu\to\nu - 2\pi n/M$, where n is any integer. We conclude from that that the maximum $\mathscr{M}(\rho,\nu)$ of $|\det[\cdots]|^2$ is a periodic function of ν, with period $\omega = 2\pi/M$.

From the constraints (B16), we can derive the following relation:

$$\begin{aligned}&\sum_{\alpha_1=1}^{N_1}\cdots\sum_{\alpha_p=1}^{N_p}(\theta_{\alpha_1 1}+\cdots+\theta_{\alpha_p p})\\ &\quad = 2\pi N_1\cdots N_p\left(\frac{n_1}{N_1}+\cdots+\frac{n_p}{N_p}\right)\\ &\quad = 2\pi N_1\cdots N_p n/M,\end{aligned} \tag{B20}$$

where $n_1,\ldots,n_p$, and also n, can take any integer values. Defining the multiple index $\{\alpha\}$ and the variables $\theta_{\{\alpha\}}$ by

$$\{\alpha\} = \{\alpha_1,\ldots,\alpha_p\}, \tag{B21}$$

$$\theta_{\{\alpha\}} = \theta_{\alpha_1 1}+\cdots+\theta_{\alpha_p p}, \tag{B22}$$

Eq. (B20) can be rewritten as

$$\sum_{\{\alpha\}}\theta_{\{\alpha\}} = 0 \quad (\mathrm{mod}\ 2\pi N_1\cdots N_p/M), \tag{B23}$$

and the right-hand side of (B17) becomes the product $\Pi_{\{\alpha\}}[1+2\rho\cos(\theta_{\{\alpha\}}+\nu)+\rho^2]$.

We now introduce the following auxiliary maximum problem: find the maximum $\mathscr{N}(\rho,\nu)$ of this product, when the $\theta_{\{\alpha\}}$'s are constrained by (B23).

Since the constraint (B23) is weaker than the constraints (B16):

$$\mathscr{M}(\rho,\nu) \leqslant \mathscr{N}(\rho,\nu). \tag{B24}$$

As a matter of fact, this auxiliary problem is nothing but the problem solved in part (I) of this proof, with N equal to $\Sigma_{\{\alpha\}}1 = N_1\cdots N_p$, but with the difference that the constraint on the θ's, which previously was written mod 2π, is now written mod $2\pi N_1\cdots N_p/M$. Very little has to be modified in the proof to conclude (i) that the maximum $\mathscr{N}(\rho,\nu)$ is periodic in ν, with period $\omega = 2\pi/M$, and (ii) that it is reached when all $\theta_{\{\alpha\}}$'s are equal,

$$\theta_{\alpha_1 1}+\cdots+\theta_{\alpha_p p} = 2\pi n/M, \tag{B25}$$

with the integer n fixed by

$$-\pi/M < \nu + 2\pi n/M \leqslant \pi/M. \tag{B26}$$

An immediate consequence of Eqs. (B25) is that all $\theta_{\alpha j}$'s for j fixed and $\alpha = 1,\ldots,N_j$ are equal. Moreover they can be given a value $2\pi n_j/N_j$, with n_j integers, in such a way as to satisfy the constraints (B16). It suffices to find integers n_j's such that

$$\frac{n_1}{N_1}+\cdots+\frac{n_p}{N_p}=\frac{n}{M}, \tag{B27}$$

or equivalently $n_1m_1+\cdots+n_pm_p=n$, and we know that this is always possible, due to condition (B19). This implies that $\mathscr{M}(\rho,\nu)$ and $\mathscr{N}(\rho,\nu)$ are equal, and that the maximum $\mathscr{M}(\rho,\nu)$ of $|\det[1+\rho e^{i\nu}\mathbb{R}(g)]|^2$ is reached on those elements of $\mathscr{G}$ defined by

$$\theta_{\alpha j}=2\pi n_j/N_j, \tag{B28}$$

where the integers n_j's satisfy the condition

$$-\frac{\pi}{M}<\nu+2\pi\left(\frac{n_1}{N_1}+\cdots+\frac{n_p}{N_p}\right)\leqslant\frac{\pi}{M}. \tag{B29}$$

This ends the proof of the lemma.

Extensions of the lemma: Since the elements of $\mathscr{G}$, where the maximum of $|\det[1+\rho e^{i\nu}\mathbb{R}(g)]|$ is reached, do not depend on ρ, the lemma is still valid if we replace $|\det[\cdots]|$ by $\int d\mu(\rho)\phi(|\det[\cdots]|)$, provided ϕ be a nondecreasing function, and $d\mu(\rho)$ be a positive measure on the positive ρ axis.

Choose the function $\phi=\ln$:

$$\ln|\det[1+\rho e^{i\nu}\mathbb{R}(g)]|=\operatorname{Re}\sum_{k=1}^{\infty}\frac{(-)^{k-1}}{k}\rho^k e^{ik\nu}\chi(g^k). \tag{B30}$$

From the elementary integral

$$\frac{1}{q!}\int_0^1 d\rho\,\rho^{k-1}\left(\ln\frac{1}{\rho}\right)^q=\frac{1}{k^{q+1}}, \tag{B31}$$

valid for any q when $k>0$, we conclude that the lemma holds for the function

$$\begin{aligned}\mathscr{V}_q(g,\nu)&=\frac{1}{(q-2)!}\int_0^1\frac{d\rho}{\rho}\left(\ln\frac{1}{\rho}\right)^{q-2}\\&\quad\times\ln|\det[1+\rho e^{i\nu}\mathbb{R}(g)]|\\&=\operatorname{Re}\sum_{k=1}^{\infty}\frac{(-)^{k-1}}{k^q}e^{ik\nu}\chi(g^k).\end{aligned} \tag{B32}$$

APPENDIX C: PROOF OF EQ. (25)

Let g be any element of the SU(N) group, with eigenvalues $e^{i\theta_k}$, $k=1,\ldots,N$. The angles $\theta_1,\ldots,\theta_N$ are restricted by $\Sigma_{k=1}^N\theta_k=0$, mod 2π. A class function $f(g)$ on SU(N) depends only on the θ's, and its mean value over the group is given by[9]

$$\begin{aligned}\int_{\mathrm{SU}(N)}&d\mu(g)f(g)\\&=\frac{1}{N!}\left(\prod_{i=1}^{N-1}\int_{-\pi}^{\pi}\frac{d\theta_i}{2\pi}\right)\left[\prod_{j<k}^{N}\left(2\sin\frac{\theta_j-\theta_k}{2}\right)^2\right]\\&\quad\times f(\theta_1,\ldots,\theta_N).\end{aligned} \tag{C1}$$

For the Gaussian function

$$f(\theta_1,\ldots,\theta_N)=\exp\left(-\frac{C}{2}\sum_{n=1}^N\theta_n^2\right),$$

in the limit where $C\to\infty$, the integral may be evaluated in the saddle point approximation, around the maximum at $\theta_1=\cdots=\theta_N=0$. We obtain

$$\begin{aligned}\int_{\mathrm{SU}(N)}&d\mu(g)\exp\left(-\frac{C}{2}\sum_{n=1}^N\theta_n^2\right)\\&\underset{C\to\infty}{\sim}\frac{1}{N!}\left(\prod_{i=1}^{N-1}\int_{-\infty}^{+\infty}\frac{d\theta_i}{2\pi}\right)\\&\quad\times\left[\prod_{j<k}^{N}(\theta_j-\theta_k)^2\right]\exp\left(-\frac{C}{2}\sum_{n=1}^N\theta_n^2\right).\end{aligned} \tag{C2}$$

After the change of variables $\theta_i\to t_i=\sqrt{C}\,\theta_i$, we transform this $(N-1)$-dimensional integral into a N-dimentional one, by introducing the missing t_N variable, through the identity

$$1=\frac{1}{2\pi}\int_{-\infty}^{+\infty}dt_N\int_{-\infty}^{+\infty}dx\exp[ix(t_1+t_2+\cdots+t_N)]. \tag{C3}$$

Straightforward manipulations lead to the result

$$\int_{\mathrm{SU}(N)}d\mu(g)\exp\left(-\frac{C}{2}\sum_{n=1}^N\theta_n^2\right)\underset{C\to\infty}{\sim}\frac{C^{(1-N^2)/2}}{(2\pi)^{N-1/2}N!\sqrt{N}}I, \tag{C4}$$

where I is the following integral, evaluated in Ref. 11:

$$\begin{aligned}I=&\left(\prod_{i=1}^N\int_{-\infty}^{+\infty}dt_i\right)\left(\prod_{j<k}^N(t_j-t_k)^2\right)\\&\times\exp\left(-\frac{1}{2}\sum_{n=1}^N t_n^2\right)=(2\pi)^{N/2}\prod_{j=1}^N j!.\end{aligned} \tag{C5}$$

This ends the proof of Eq. (25).

[1] J. I. Kapusta, Phys. Rev. D **23**, 2444 (1981).

[2] G. Auberson, L. Epele, and G. Mahoux, Saclay preprint No. SPhT/84/154.

[3] K. E. Eriksson, N. Mukunda, and B. B. Skagerstam, Phys. Rev. D **24**, 2615 (1981); M. I. Gorenstein, O. A. Mogilevsky, V. K. Petrov, and G. M. Zinovjev, Z. Phys. C **18**, 13 (1983); K. Redlich, Z. Phys. C **21**, 69 (1983); J. C. Anjos, J. Sá Borges, A. P. Malbouisson, and F. R. A. Simão, Z. Phys. C **23**, 243 (1984); R. Hagedorn and K. Redlich, CERN preprint No. TH.3889.

[4] M. I. Gorenstein, S. I. Lipskikh, V. K. Petrov, and G. M. Zinovjev, Phys. Lett. B **123**, 437 (1983).

[5] B. S. Skagerstam, Z. Phys. C **24**, 97 (1984).

[6] The authors of Ref. 4 made a mistake when they let their variables γ_1 and γ_2 vary in the interval $(-\pi,\pi)$. In so doing, they do not weight uniformly (invariantly) the different elements of SU(3). As a result they miss two of the three maxima of their extremum problem. The same error appears in M. I. Gorenstein, S. I. Lipskikh, and G. M. Zinovjev, Z. Phys. C **22**, 189 (1984).

[7] L. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products* (Academic, New York, 1965).

[8] R. Hagedorn, CERN preprint No. TH.3918/84, 1984.

[9] H. Weyl, *The Classical Groups* (Princeton U. P., Princeton, NJ, 1946); F. D. Murnaghan, *The Theory of Group Representations* (Johns Hopkins, Baltimore, MD, 1938).

[10] See for example the "additive coprimality criterion," in H. Hasse, *Number Theory* (Springer, Berlin, 1980), p. 19.

[11] M. L. Mehta, *Random Matrices* (Academic, New York, 1967).

Direct and inverse scattering in the time domain for a dissipative wave equation. I. Scattering operators

G. Kristensson
Division of Electromagnetic Theory, Royal Institute of Technology, S-100 44 Stockholm, Sweden

R. J. Krueger
Applied Mathematical Sciences, Ames Laboratory—United States Department of Energy, Iowa State University, Ames, Iowa 50011

(Received 8 August 1985; accepted for publication 30 January 1986)

This is the first part of a series of papers devoted to direct and inverse scattering of transient waves in lossy inhomogeneous media. The medium is assumed to be stratified, i.e., it varies only with depth. The wave propagation is modeled in an electromagnetic case with spatially varying permittivity and conductivity. The objective in this first paper is to analyze properties of the scattering operators (impulse responses) for the medium and to introduce the reader to the inverse problem, which is the subject of the second paper in this series. In particular, imbedding equations for the propagation operators are derived and the corresponding equations for the scattering operators are reviewed. The kernel representations of the propagation operators are shown to have compact support in the time variable. This property implies that transmission and reflection data can be extended from one round trip to arbitrary time intervals. The compact support of the propagator kernels also restricts the admissible set of transmission kernels consistent with the model employed in this paper. Special cases of scattering and propagation kernels that can be expressed in closed form are presented.

I. INTRODUCTION

The propagation of waves in lossy media can be modeled in a number of ways, depending on the features of the propagation that are of interest. This series of papers will deal with linear wave propagation in an inhomogeneous medium that is characterized by dissipation and phase velocity profiles that are independent of the frequency of the wave. A precise model for such propagation is given in Sec. II of this paper. This model involves one-dimensional electromagnetic wave propagation in the time domain in a medium that is characterized by spatially varying permittivity and conductivity profiles.

This series of papers presents a time domain approach to wave propagation that yields a unified theory for both direct and inverse scattering. The basis for this approach is in the splitting/invariant imbedding techniques that have been exploited in earlier work. Specifically, these techniques apply to time domain reflection and transmission operators for a given scattering medium.

For the convenience of the reader, the pertinent features of previous work in this area will be explicitly displayed when necessary. The present paper, Part I, deals with the direct scattering problem; i.e., given the dissipation and phase velocity profiles, determine the scattering operators (or impulse responses) for the medium. These are operators that can be used to map any transient normally incident field over to the resulting scattered fields. Various properties of these operators are developed, and it is shown how they can be utilized to "extend" scattering data.

A subsequent paper,[1] Part II, deals with the full inverse problem; i.e., given the scattering operators for a medium, determine both the dissipation and phase velocity profiles for the medium. Since a number of results derived in Part I are not used in Part II, the reader who is primarily interested in the inverse problem can proceed to Part II after reading this introduction and Sec. II of the present paper, in which notation is established and a precise statement of the problem is given. Results from Part I that are used in the inverse problem are summarized at the beginning of Part II. Some numerical examples showing scattering operators (as well as inversion procedures) will be given in Part II.

Section III of the present paper reviews the integrodifferential equations satisfied by the kernels of the scattering operators and relates these kernels to the propagator kernels for the medium. A reciprocity result is also derived. Integrodifferential equations for the propagator kernels are derived in Sec. IV. In Sec. V a result that can be used to characterize transmission data is developed. This result is also used to extend reflection and transmission data from a single round trip time trace to a time trace of arbitrary length. Section VI is a summary of the work in Part I. Appendix A supplies technical details used in Sec. V. Closed form expressions for scattering and propagator kernels in special cases are given in Appendix B. Finally, operator equations for the propagators are shown in Appendix C.

To put the present results in their proper context, some details regarding previous work are now given. Corones and Krueger[2] and Davison[3] developed a system of integrodifferential equations for the reflection and transmission operators. Those studies displayed the time domain behavior of these operators. However, it was also shown that those results could be interpreted in the frequency domain. In that case, a Riccati differential equation for the reflection coefficient was obtained.

In later work, Corones *et al.* used the reflection operator equation as the basis for an inversion algorithm in nondissipative[4–7] as well as dissipative media.[8–10] (In the dissipative

0022-2488/86/061667-16$02.50

case, *a priori* information about phase velocity or dissipation is required.) Bruckstein *et al.*[11] have given a partial review of Riccati equation techniques in the frequency domain and shown their relation to certain integral equation approaches to inversion.

There are several well-documented solutions to the full inverse problem (i.e., simultaneous reconstruction of both dissipation and phase velocity profiles). Such solutions, however, use completely different methods than the techniques presented here. Weston[12–14] was the first to use the full scattering matrix in the time domain to solve the one-dimensional dissipative inverse problem. He applied a Riemann function approach to develop a system a Gel'fand–Levitan-type equations whose solution yielded the desired profiles. The data for this problem consisted of the time domain reflection and transmission operators. These results were generalized by Krueger[15–17] to include more realistic material profiles. This had the effect of also reducing the data requirements in the problem, although transmission data were still required. More will be said about this in Part II. Jaulent[18,19] pursued a frequency domain approach to dissipative inverse problems in a variety of settings. The problems considered involved a complex potential with a linear dependence on frequency, and the required data consisted of reflection and transmission coefficients.

A model of dissipative wave propagation, which is more physically motivated than that used in this paper or in any of the above-referenced papers, is possible. Such a model is obtained by appealing to the underlying constitutive relation in the problem. In the frequency domain, this implies a certain dispersion relation, whereas in the time domain, this implies the existence of a memory function for the medium. The methods used in the present series of papers also have been applied to direct and inverse scattering problems in electromagnetic[20] and viscoelastic[21] media, which are characterized by such a memory function.

II. STATEMENT OF THE PROBLEM

In this section some notation is introduced and a precise statement of the inverse problem is given. The geometry of the problem is shown in Fig. 1. An inhomogeneous slab occupies the region $0<z<L$. This medium is assumed to be stratified so that the permittivity and conductivity are functions of depth z only. A homogeneous, lossless medium is situated on either side of this slab.

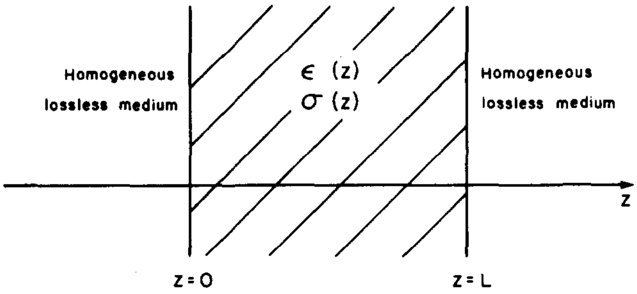

FIG. 1. The geometry of the inhomogeneous medium.

An electromagnetic plane wave is launched in the region exterior to the slab. This impinges normally on the medium, giving rise to an electric field $E(z,t)$ within the slab, with E satisfying

$$E_{zz}(z,t) - c^{-2}(z)E_{tt}(z,t) - b(z)E_t(z,t) = 0, \qquad (2.1)$$

where

$$c^{-2}(z) = \epsilon(z)\mu_0, \quad b(z) = \sigma(z)\mu_0, \qquad (2.2)$$

and μ_0 is the permeability in vacuum, $\sigma(z)$ is the conductivity, and $\epsilon(z)$ is the permittivity. The analysis becomes simpler if the phase velocity $c(z)$ is continuously differentiable within the slab. This will be assumed throughout this paper. It is further assumed that the phase velocity is continuous (although not necessarily smooth) at the boundary of the slab. Thus, in the regions exterior to the slab the phase velocity is given by

$$\begin{aligned} c(z) &= c(0^+), \quad z\leqslant 0, \\ c(z) &= c(L^-), \quad z\geqslant L \end{aligned} \qquad (2.3)$$

(where the ± superscript denotes the limit from the right and the limit from the left, respectively). These assumptions insure that E and E_z are everywhere continuous. This precludes the existence of impulsive echoes in the scattered fields.

Now if the incident plane wave is launched in the region $z<0$, then the general solution of Eq. (2.1) in the region to the left of the slab is

$$E(z,t) = E^i_+(t - z/c(0)) + E^r_+(t + z/c(0)), \quad z<0. \qquad (2.4)$$

Here, E^i_+ and E^r_+ denote the incident and reflected fields, respectively. The subscript " + " denotes the fact that the incident field is propagating in the $+z$ direction. In addition, a transmitted field is produced in the region to the right of the slab. This has the form

$$E(z,t) = E^t_+(t - l - (z-L)/c(L)), \quad z>L, \qquad (2.5)$$

where

$$l = \int_0^L c^{-1}(z)dz. \qquad (2.6)$$

The incident and scattered fields are related by the scattering operators (i.e., reflection and transmission operators) for the slab. These are integral operators represented by

$$\begin{aligned} E^r_+(t) &= \int_0^t \widetilde{R}^+(t-t')E^i_+(t')dt', \\ E^t_+(t) &= \tilde{t}^+E^i_+(t) + \int_0^t \widetilde{T}^+(t-t')E^i_+(t')dt', \end{aligned} \qquad (2.7)$$

where

$$\tilde{t}^+ = \left[\frac{c(L)}{c(0)}\right]^{1/2} \exp\left[-\frac{1}{2}\int_0^L b(z)c(z)dz\right]. \qquad (2.8)$$

In Eq. (2.7) the functions $\widetilde{R}^+$ and $\widetilde{T}^+$ are the reflection and transmission kernels, respectively, for incidence from the left. Notice that the lower limit of integration in (2.7) has been chosen to be 0, which is equivalent to assuming that the incident wave front first impinges on the slab at $t=0$. Notice also that the time variable t in Eq. (2.7) does not represent physical time, but rather a characteristic variable for Eq.

(2.1) outside the slab [cf. also Eqs. (2.4) and (2.5)].

The existence of the scattering operators in Eq. (2.7) can be verified in a number of ways, one of which is shown in Ref. 15. In particular, these operators are independent of the incident field used in the scattering experiment and depend only on the properties of the slab. Furthermore, velocity mismatch effects have been taken out of the problem by the assumption that c is continuous (although not necessarily smooth) at $z=0$ and $z=L$. Hence, in comparing the form of the operators given in Eq. (2.7) with those in Ref. 15, the constants c_0 and c_l in Ref. 15 must be set equal to 1.

A second pair of reflection and transmission operators describe scattering experiments for incident fields impinging on the medium from the right. In this case the general solution of Eq. (2.1) in the region $z>L$ is

$$E(z,t) = E^i_-\,(t+(z-L)/c(L)) + E^r_-\,(t-(z-L)/c(L)),\quad z>L, \tag{2.9}$$

where E^i_- and E^r_- are the incident and reflected fields, respectively. To the left of the slab the transmitted field is given by

$$E(z,t) = E^t_-\,((t-l+z/c(0)),\quad z<0. \tag{2.10}$$

These fields are again related by scattering operators for the slab, which are represented by

$$\begin{aligned} E^r_-\,(t) &= \int_0^t \widetilde{R}^-(t-t')E^i_-\,(t')dt', \\ E^t_-\,(t) &= \tilde{t}^-E^i_-\,(t) + \int_0^t \widetilde{T}^-(t-t')E^i_-\,(t')dt', \end{aligned} \tag{2.11}$$

where

$$\tilde{t}^- = \left[\frac{c(L)}{c(0)}\right]^{-1/2} \exp\left[-\frac{1}{2}\int_0^L b(z)c(z)dz\right]. \tag{2.12}$$

Again in Eq. (2.11) it is assumed that $t=0$ corresponds to the time the wave front first impinges on the slab at $z=L$. Notice that if $E^i_\pm\,(t)=\delta(t)$ (where δ is the Dirac delta), then from Eqs. (2.7) and (2.11), it follows that $E^r_\pm\,(t)=\widetilde{R}(t)$ and $E^t_\pm\,(t) = t^{\pm}\delta(t)+\widetilde{T}(t)$. Hence, the scattering kernels $\widetilde{R}^{\pm}$, $\widetilde{T}^{\pm}$ are the impulse responses for the medium.

The inverse problem considered in this series of papers is that of determining both $\epsilon(z)$ and $\sigma(z)$ (as well as L) for the slab through the use of scattering experiments performed on the slab. More precisely, the scattering data used in the reconstruction of ϵ and σ consist of finite time traces of both reflection kernels, $\widetilde{R}^{\pm}(t)$, and one of the transmission kernels, say, $\widetilde{T}^+(t)$ for $0<t<2l$. Here, $2l$ [with l defined by Eq. (2.6)] represents the time it takes a signal to travel one complete round trip through the medium.

The data used in this formulation of the problem are a deconvolution of Eq. (2.7) and (2.11). The effect of imperfect deconvolution can be studied (at least numerically) by means of the inversion algorithms presented in Part II.

At this point a transformation of dependent and independent variables in Eq. (2.1) is made. This transformation is not necessary for the implementation of the inversion algorithms given in Part II. However, it does result in a simpler-looking analysis (compare with Ref. 10) and numerical scheme. Begin with the change of coordinates,

$$\begin{aligned} x &= x(z) = \int_0^z \frac{c^{-1}(z')dz'}{l}, \\ s &= t/l, \\ u(x,s) &= E(z,t), \end{aligned} \tag{2.13}$$

where x is normalized travel time and s is normalized time. In these new coordinates the slab occupies the region $0\leqslant x\leqslant 1$ and a round trip time trace is described by $0<s<2$. Equation (2.1) becomes

$$u_{xx} - u_{ss} + A(x)u_x + B(x)u_s = 0, \tag{2.14}$$

where

$$A(x) = -\frac{d}{dx}\ln c(z(x)), \tag{2.15}$$

$$B(x) = -lb(z(x))c^2(z(x)), \tag{2.16}$$

and ln denotes the natural logarithm function. The coefficient functions A and B vanish outside of the interval $[0,1]$ and are continuous on the interval $(0,1)$, with possible finite discontinuities at $x=0$ and $x=1$. Typical plots of A and B are shown in Fig. 2.

It follows from the compact support of A and B that for $x<0$ and $x>1$, solutions of (2.14) reduce to right and left moving waves. These are readily related to the physical fields. In particular, scattering operators again exist for Eq.

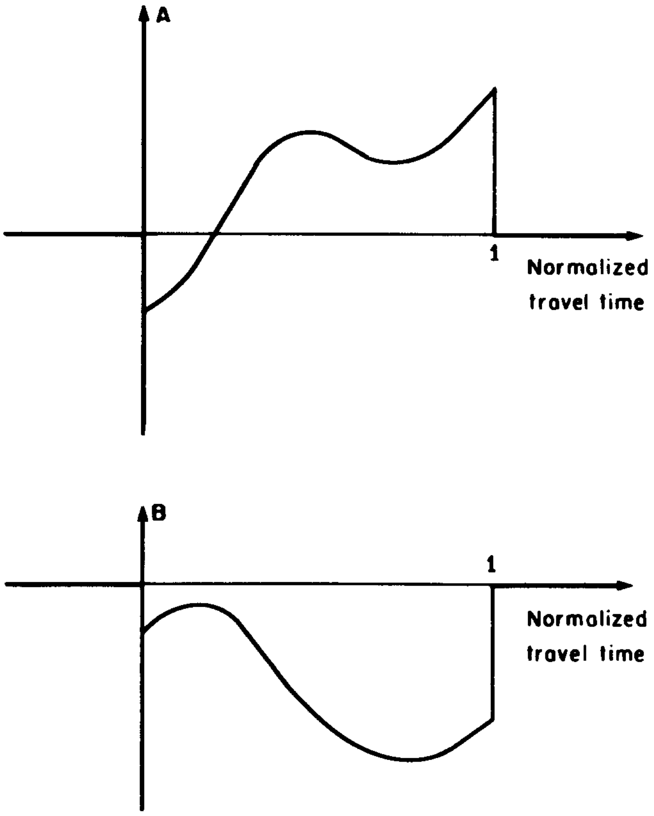

FIG. 2. Profile functions A and B.

(2.14). For a right moving incident wave $u^i_+(s-x)$, launched in the region $x<0$, it can be shown that the reflected and transmitted fields are given by

$$u^r_+(s)=\int_0^s R^+(0,1,s-s')u^i_+(s')ds', \tag{2.17}$$

$$u^t_+(s)=t^+(0,1)u^i_+(s)+\int_0^s T^+(0,1,s-s')u^i_+(s')ds', \tag{2.18}$$

while a left moving incident wave, $u^i_-(s+x-1)$, in the region $x>1$ produces reflected and transmitted fields

$$u^r_-(s)=\int_0^s R^-(0,1,s-s')u^i_-(s')ds', \tag{2.19}$$

$$u^t_-(s)=t^-(0,1)u^i_-(s)+\int_0^s T^-(0,1,s-s')u^i_-(s')ds', \tag{2.20}$$

where

$$t^\pm(0,1)=\exp\left[\mp\frac{1}{2}\int_0^1\{A(x)\mp B(x)\}dx\right]. \tag{2.21}$$

The reflection and transmission kernels in (2.17)–(2.20) are related to the physical kernels in (2.7) and (2.11) via

$$\begin{aligned}R^\pm(0,1,s)&=l\widetilde{R}^\pm(ls),\\ T^\pm(0,1,s)&=l\widetilde{T}^\pm(ls).\end{aligned} \tag{2.22}$$

Notice that these transformed kernels reference the end points of the scattering medium. This is because in later sections of this paper, scattering kernels for subsections of the original medium will be considered. Observe that the independent variable in (2.18) and (2.20) can be thought of as a characteristic variable.

Finally, it is necessary to define a second set of operators for the scattering problems relevant to (2.14). These are propagation operators[22] for the medium and are used to express the incident and reflected fields in terms of the transmitted field. They are given by (see Ref. 15)

$$u^i_\pm(s)=[t^\pm(0,1)]^{-1}u^t_\pm(s)+\int_0^s W^\pm(0,1,s-s')u^t_\pm(s')ds', \tag{2.23}$$

$$u^r_\pm(s)=[t^\pm(0,1)]^{-1}\int_0^s V^\pm(0,1,s-s')u^t_\pm(s')ds'. \tag{2.24}$$

Notice that the "W" operator is just the inverse of the corresponding "T" operator. Consequently, the kernels $W^\pm(0,1,s)$ in Eq. (2.23) are just the resolvent kernels for the functions $T^\pm(0,1,s)$. The explicit relation between these kernels is

$$[t^\pm(0,1)]^{-1}T^\pm(0,1,s)+t^\pm(0,1)W^\pm(0,1,s)+\int_0^s T^\pm(0,1,s-s')W^\pm(0,1,s')ds'=0. \tag{2.25}$$

The end points of the (transformed) slab are explicitly displayed in the arguments of $R^\pm(0,1,s)$ and $T^\pm(0,1,s)$, and as well as in $V^\pm(0,1,s)$ and $W^\pm(0,1,s)$. In the next section both of the end points of the slab are allowed to vary (see Fig. 3) and in this more general case $R^\pm(x,y,s)$ and $T^\pm(x,y,s)$ denote the reflection and transmission kernels, respectively, for the subregion of the slab with end points at x and y, with $0\leqslant x\leqslant y\leqslant 1$. A similar notation holds for $V^\pm(x,y,s)$ and $W^\pm(x,y,s)$. It should be stressed that it is only the kernels corresponding to $x=0$ and $y=1$ that are physically obtainable.

III. EQUATIONS FOR THE SCATTERING KERNELS

In the preceding section the physical reflection and transmission kernels were introduced. These are the data that are obtained from a scattering experiment. Throughout the remainder of this paper, the transformed problem given in (2.14) will be studied, and in particular the kernels on the left-hand side of (2.22) will be referred to as the physical scattering kernels (since they are easily obtained from the physical data).

The dependence of the scattering kernels on the parameters x and y (which are the end points of the subregion $[x,y]$) will be reviewed in this section. It is intuitively clear that this dependence is related to the material properties of the slab. Relations to this effect are developed in detail in Ref. 2. For the convenience of the reader and for completeness the main results of that reference are given here:

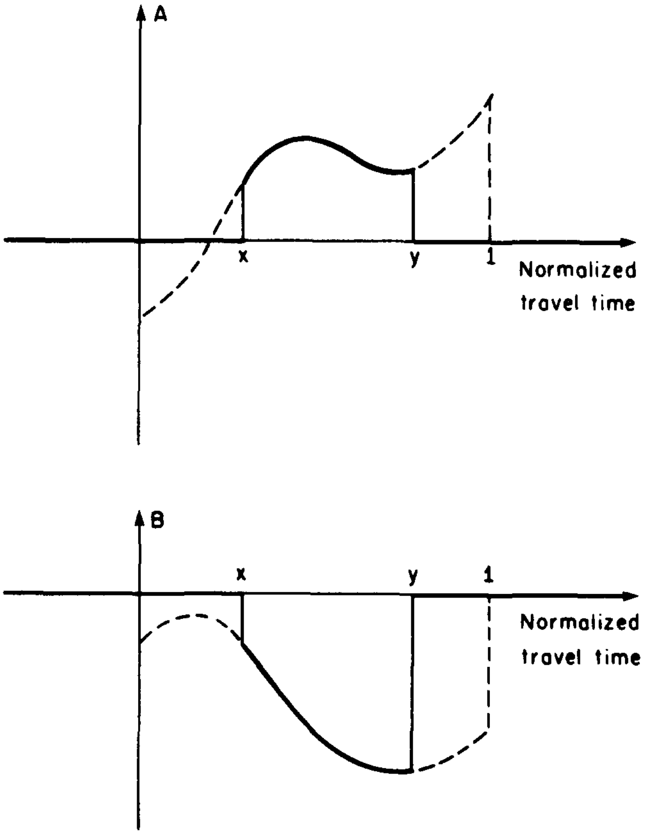

FIG. 3. Profile functions A and B for the subregion $[x,y]$. The dashed lines indicate the omitted portions of the physical region.

$$
\begin{aligned}
&R_x^+(x,y,s)\\
&\quad=2R_s^+(x,y,s)-B(x)R^+(x,y,s)\\
&\qquad-\tfrac{1}{2}[A(x)+B(x)]\\
&\qquad\times\int_0^s R^+(x,y,s')R^+(x,y,s-s')ds',\quad s>0,\\
&R^+(y,y,s)=0,\quad s>0,\\
&R^+(x,y,0^+)=-\tfrac{1}{4}[A(x)-B(x)],\quad x<y;
\end{aligned}
\tag{3.1}
$$

$$
\begin{aligned}
&T_x^+(x,y,s)\\
&\quad=\tfrac{1}{2}[A(x)-B(x)]T^+(x,y,s)\\
&\qquad-\frac{1}{2}[A(x)+B(x)]\Big\{t^+(x,y)R^+(x,y,s)\\
&\qquad+\int_0^s T^+(x,y,s')R^+(x,y,s-s')ds'\Big\},\quad s>0,
\end{aligned}
\tag{3.2}
$$

$$T^+(y,y,s)=0,\quad s>0;$$

$$
\begin{aligned}
&T_x^-(x,y,s)\\
&\quad=-\tfrac{1}{2}[A(x)+B(x)]\\
&\qquad\times\Big\{T^-(x,y,s)+t^-(x,y)R^+(x,y,s)\\
&\qquad+\int_0^s T^-(x,y,s')R^+(x,y,s-s')ds'\Big\},\quad s>0,
\end{aligned}
\tag{3.3}
$$

$$T^-(y,y,s)=0,\quad s>0;$$

$$
\begin{aligned}
&R_x^-(x,y,s)\\
&\quad=\tfrac{1}{2}[A(x)+B(x)]\Big\{t^+(x,y)T^-(x,y,s-2(y-x))\\
&\qquad+t^-(x,y)T^+(x,y,s-2(y-x))\\
&\qquad+\int_0^{s-2(y-x)}T^+(x,y,s')\\
&\qquad\times T^-(x,y,s-2(y-x)-s')ds'\Big\},\\
&\qquad s>2(y-x),
\end{aligned}
\tag{3.4}
$$

$$R^-(y,y,s)=0,\quad s>0;$$

$$
\begin{aligned}
&R_y^-(x,y,s)\\
&\quad=-2R_s^-(x,y,s)+B(y)R^-(x,y,s)\\
&\qquad-\tfrac{1}{2}[A(y)-B(y)]\\
&\qquad\times\int_0^s R^-(x,y,s')R^-(x,y,s-s')ds',\quad s>0,\\
&R^-(x,x,s)=0,\quad s>0,\\
&R^-(x,y,0^+)=\tfrac{1}{4}[A(y)+B(y)],\quad x<y;
\end{aligned}
\tag{3.5}
$$

$$
\begin{aligned}
&T_y^+(x,y,s)\\
&\quad=-\tfrac{1}{2}[A(y)-B(y)]\\
&\qquad\times\Big\{T^+(x,y,s)+t^+(x,y)R^-(x,y,s)\\
&\qquad+\int_0^s T^+(x,y,s')R^-(x,y,s-s')ds'\Big\},\quad s>0,
\end{aligned}
\tag{3.6}
$$

$$T^+(x,x,s)=0,\quad s>0;$$

$$
\begin{aligned}
&T_y^-(x,y,s)\\
&\quad=\tfrac{1}{2}[A(y)+B(y)]T^-(x,y,s)\\
&\qquad-\frac{1}{2}[A(y)-B(y)]\Big\{t^-(x,y)R^-(x,y,s)\\
&\qquad+\int_0^s T^-(x,y,s')R^-(x,y,s-s')ds'\Big\},\quad s>0,
\end{aligned}
\tag{3.7}
$$

$$T^-(x,x,s)=0,\quad s>0;$$

$$
\begin{aligned}
&R_y^+(x,y,s)\\
&\quad=-\tfrac{1}{2}[A(y)-B(y)]\Big\{t^-(x,y)T^+(x,y,s\\
&\qquad-2(y-x))+t^+(x,y)T^-(x,y,s-2(y-x))\\
&\qquad+\int_0^{s-2(y-x)}T^-(x,y,s')\\
&\qquad\times T^+(x,y,s-2(y-x)-s')ds'\Big\},\quad s>2(y-x),
\end{aligned}
\tag{3.8}
$$

$$R^+(x,x,s)=0,\quad s>0;$$

where

$$t^\pm(x,y)=\exp\Big\{\mp\frac{1}{2}\int_x^y[A(x')\mp B(x')]dx'\Big\}.\tag{3.9}$$

Equations (3.1)–(3.8) are the imbedding equations for the slab, obtained from continuously imbedding scattering kernels for subintervals of the slab into a family of scattering kernels. In particular, these equations display the change in the scattering kernels due to variations in one of the end points of the imbedded slab. As seen from above, these equations are in general nonlinear and of integrodifferential type. Note that two of the equations, Eqs. (3.1) and (3.5), are both equations for a single unknown kernel. The other six equations couple different kernels together. With each of the equations above there is also a boundary condition for the case when $x=y$. This corresponds to a slab of zero thickness. In Eqs. (3.1) and (3.5), there are also two auxiliary conditions relating the early time behavior of the reflection kernels $R^\pm(x,y,s)$ to the properties of the slab (see also Fig. 4). Equations (3.1)–(3.8) are written in a slightly different form than in Ref. 2 due to the particular representation of the scattering operators given in Eq. (2.17)–(2.20).

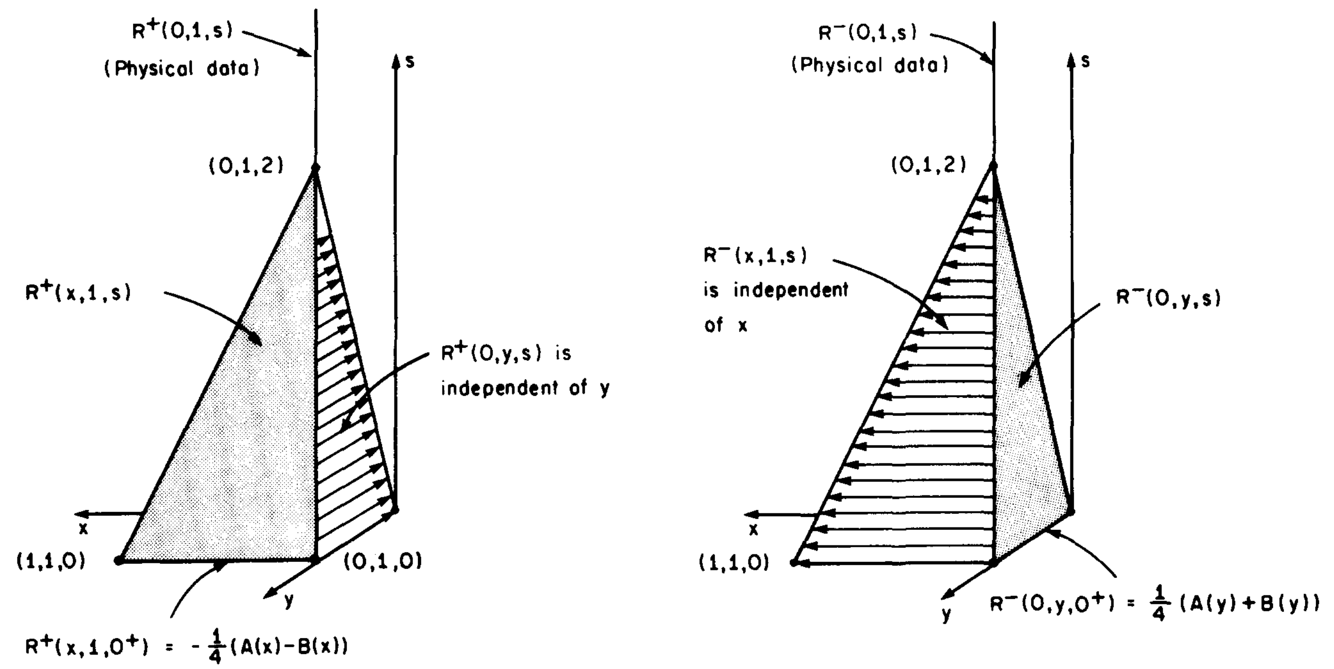
$R^+(0,1,s)$
(Physical data)
s
$(0,1,2)$
$R^+(x,1,s)$
$R^+(0,y,s)$ is independent of y
x
$(1,1,0)$
$(0,1,0)$
y
$R^+(x,1,0^+) = -\frac{1}{4}(A(x)-B(x))$
$R^-(0,1,s)$
(Physical data)
s
$(0,1,2)$
$R^-(x,1,s)$ is independent of x
$R^-(0,y,s)$
x
$(1,1,0)$
y
$R^-(0,y,0^+) = \frac{1}{4}(A(y)+B(y))$

FIG. 4. A portion of the domain $R^{\pm}(x,y,s)$. The entire domain is $0<x<y<1, s>0$. The region inside the tetrahedron is the domain of $R^{\pm}(x,y,s)$ for s limited to one round trip in the subregion $[x,y]$.

The reflection kernels $R^{\pm}(x,y,s)$ are discontinuous across the plane $s=2(y-x)$. These discontinuities are associated with the echo of the wave front from the rear interface. Again referring to Ref. 2, the jumps in the kernels along that plane are

$$\begin{aligned}&[R^+(x,y,s)]_{s=2(y-x)^-}^{s=2(y-x)^+}\\&\quad=\frac{1}{4}[A(y)-B(y)]\exp\left\{\int_x^y B(x')dx'\right\},\\&[R^-(x,y,s)]_{s=2(y-x)^-}^{s=2(y-x)^+}\qquad(3.10)\\&\quad=-\frac{1}{4}[A(x)+B(x)]\exp\left\{\int_x^y B(x')dx'\right\}.\end{aligned}$$

In Ref. 2 it is also shown that the reflection kernels $R^{\pm}(x,y,s)$ satisfy (see also Fig. 4)

$$\begin{aligned}&R^+(x,y,s)=R^+(x,x+s/2^+,s),\quad s<2(y-x),\\&\qquad\qquad(3.11)\\&R^-(x,y,s)=R^-(y-s/2^-,y,s),\quad s<2(y-x).\end{aligned}$$

These relations state that the reflected field is independent of the position of the rear interface of the slab for times less than one round trip through the subregion $[x,y]$. The properties of the reflection kernels $R^{\pm}(x,y,s)$ given by Eqs. (3.10) and (3.11) will be used in the inversion algorithm presented in Part II of this series of papers.

In the transmission kernel equations given above, there is a distinction between the T^+ and T^- kernels. Now assume that there is a relation between the kernels $T^{\pm}(x,y,s)$ of the following form:

$$T^+(x,y,s)=f(x,y)T^-(x,y,s),\qquad(3.12)$$

where $f(x,y)$ is an unknown function to be determined. A relation of this kind is suggested by the fact that a reciprocity result should exist for the transmission operators. Equations (3.2), (3.3), (3.6), (3.7), and (3.12) imply that $f(x,y)$ must satisfy

$$\begin{aligned}&f_x(x,y)=A(x)f(x,y),\\&f_y(x,y)=-A(y)f(x,y),\qquad(3.13)\\&f(x,y)=t^+(x,y)/t^-(x,y).\end{aligned}$$

These three equations are consistent and it is therefore convenient to introduce a single transmission kernel $T(x,y,s)$ defined by

$$T(x,y,s)=T^+(x,y,s)/t^+(x,y)=T^-(x,y,s)/t^-(x,y).\qquad(3.14)$$

In what follows, this new definition of the transmission kernel will be the one that is used and from now on there are only three different kinds of scattering kernels, i.e., $R^{\pm}(x,y,s)$ and $T(x,y,s)$. It is easy to see that with this new definition of the transmission kernel the Eqs. (3.2), (3.3), (3.6), and (3.7) can be replaced by two simpler ones,

$$\begin{aligned}T_x(x,y,s)&=-\frac{1}{2}[A(x)+B(x)]\left\{R^+(x,y,s)\right.\\&\quad\left.+\int_0^s T(x,y,s')R^+(x,y,s-s')ds'\right\},\\s>0,&\qquad(3.15)\\T_y(x,y,s)&=-\frac{1}{2}[A(y)-B(y)]\left\{R^-(x,y,s)\right.\\&\quad\left.+\int_0^s T(x,y,s')R^-(x,y,s-s')ds'\right\},\\s>0.&\qquad(3.16)\end{aligned}$$

The resolvent equation (2.25) for $W^{\pm}$ generalizes to the subregion $[x, y]$ in the obvious way. Now since the $T^{\pm}$ are simply related to a single kernel T, it follows that $W^{\pm}$ can be related to a single kernel W. Specifically,

$$W(x,y,s)=t^{+}(x,y)W^{+}(x,y,s)=t^{-}(x,y)W^{-}(x,y,s). \tag{3.17}$$

The resolvent equation for $W(x,y,s)$ now reads

$$T(x,y,s)+W(x,y,s)+\int_0^s T(x,y,s-s')W(x,y,s')ds'=0. \tag{3.18}$$

The equations for the $V^{\pm}(x,y,s)$ kernels are

$$R^{\pm}(x,y,s)=V^{\pm}(x,y,s)+\int_0^s T(x,y,s-s')V^{\pm}(x,y,s')ds', \tag{3.19}$$

which follows (in the "+" case) from inserting Eqs. (2.17) and (2.18) into Eq. (2.24) and using (3.14). Finally, Eq. (3.19) can be solved for $V^{\pm}$ by using the fact that W is the resolvent kernel for T. This yields

$$V^{\pm}(x,y,s)=R^{\pm}(x,y,s)+\int_0^s R^{\pm}(x,y,s-s')W(x,y,s')ds'. \tag{3.20}$$

Exact representations of the kernels $R^{\pm}$, T, W, and $V^{\pm}$ can be obtained in the special case when $A(x)$ and $B(x)$ are constants. This is done by using the Laplace transform in the variable s. Details are provided in Appendix B.

IV. THE W AND $V^{\pm}$ EQUATIONS

In this section, the dynamics of the kernels W and $V^{\pm}$ are derived. The definition of the resolvent $W(x,y,s)$ of the transmission kernel $T(x,y,s)$ is given by Eq. (3.18). Differentiation of this equation with respect to the left end point x gives

$$T_x(x,y,s)+W_x(x,y,s)+\int_0^s W_x(x,y,s')T(x,y,s-s')ds'+\int_0^s W(x,y,s')T_x(x,y,s-s')ds'=0. \tag{4.1}$$

Now use the imbedding equation for the transmission kernel T given by Eq. (3.15) and the definition of the resolvent in Eq. (3.18) to get

$$W_x(x,y,s)-\tfrac{1}{2}[A(x)+B(x)]R^{+}(x,y,s)+\int_0^s W_x(x,y,s')T(x,y,s-s')ds'=0, \tag{4.2}$$

which can be simplified to

$$W_x(x,y,s)=\frac{1}{2}[A(x)+B(x)]\left\{R^{+}(x,y,s)+\int_0^s W(x,y,s')R^{+}(x,y,s-s')ds'\right\},\quad s>0. \tag{4.3}$$

This equation gives the variation of the resolvent $W(x,y,s)$ as the left-hand side of the slab is varying. Note that this equation is very similar to the equation for the variation in the transmission kernel $T(x,y,s)$, given by (3.15).

The equation for a variation of the right-hand end point is similar to the derivation above and the result is

$$W_y(x,y,s)=\frac{1}{2}[A(y)-B(y)]\left\{R^{-}(x,y,s)+\int_0^s W(x,y,s')R^{-}(x,y,s-s')ds'\right\},\quad s>0. \tag{4.4}$$

A direct comparison between Eqs. (4.3), (4.4), and (3.20) shows that

$$W_x(x,y,s)=\tfrac{1}{2}[A(x)+B(x)]V^{+}(x,y,s), \tag{4.5}$$

$$W_y(x,y,s)=\tfrac{1}{2}[A(y)-B(y)]V^{-}(x,y,s). \tag{4.6}$$

The two pairs of equations for $V^{\pm}(x,y,s)$ now can be derived quite easily. Differentiate $V^{+}(x,y,s)$ in Eq. (3.20) once with respect to x and once with respect to s and use Eqs. (3.1) and (4.5) to obtain

$$\begin{aligned}V_x^{+}(x,y,s)&=2V_s^{+}(x,y,s)-B(x)V^{+}(x,y,s)\\&\quad+\tfrac{1}{2}[A(x)-B(x)]W(x,y,s),\quad s>0\\V^{+}(y,y,s)&=0,\quad s>0,\\V^{+}(x,y,0^{+})&=-\tfrac{1}{4}[A(x)-B(x)],\quad x<y.\end{aligned} \tag{4.7}$$

Similarly, by differentiating $V^{-}(x,y,s)$ in Eq. (3.20) with respect to y and s the following equation is obtained by the use of Eqs. (3.5) and (4.6):

$$\begin{aligned}V_y^{-}(x,y,s)&=-2V_s^{-}(x,y,s)+B(y)V^{-}(x,y,s)\\&\quad+\tfrac{1}{2}[A(y)+B(y)]W(x,y,s),\quad s>0,\\V^{-}(x,x,s)&=0,\quad s>0,\\V^{-}(x,y,0^{+})&=\tfrac{1}{4}[A(y)+B(y)],\quad x<y.\end{aligned} \tag{4.8}$$

Notice that the two Eqs. (4.7) and (4.8) do not contain any convolution integral, but couple $V^{\pm}$ with W.

The two final equations for the kernel $V^{\pm}(x,y,s)$ are derived by a differentiation with respect to the other end point in Eq. (3.20). The dynamics of V_y^{+} and V_x^{-} in the interval $0<s<2(y-x)$ are now easily obtained by the use of Eqs. (3.11), (4.5), and (4.6). This results in

$$\begin{aligned}V_y^{+}(x,y,s)&=\tfrac{1}{2}[A(y)-B(y)]\times\int_0^s R^{-}(x,y,s')V^{+}(x,y,s-s')ds'\\&=\tfrac{1}{2}[A(y)-B(y)]\times\int_0^s R^{+}(x,y,s')V^{-}(x,y,s-s')ds',\\&\quad 0<s<2(y-x),\end{aligned} \tag{4.9}$$

$$V_x^{-}(x,y,s)=\tfrac{1}{2}[A(x)+B(x)]\times\int_0^s R^{+}(x,y,s')V^{-}(x,y,s-s')ds'$$

$$=\tfrac{1}{2}[A(x)+B(x)]\times\int_0^s R^-(x,y,s')V^+(x,y,s-s')ds',$$

$$0<s<2(y-x), \qquad (4.10)$$

where the last equality in each of these equations comes from the identity

$$\int_0^s R^+(x,y,s')V^-(x,y,s-s')ds' = \int_0^s R^-(x,y,s')V^+(x,y,s-s')ds', \quad s>0, \qquad (4.11)$$

which is easily obtained from Eq. (3.20).

The equations (4.9) and (4.10) are not valid for $s>2(y-x)$. In the derivation of the corresponding equations valid for $s>2(y-x)$, the following integral is encountered:

$$\int_{2(y-x)}^s R_y^+(x,y,s')W(x,y,s-s')ds'$$
$$= -\tfrac{1}{2}[A(y)-B(y)]t^+(x,y)t^-(x,y)\Big\{W(x,y,s-2(y-x))$$
$$+2\int_0^{s-2(y-x)} T(x,y,s')W(x,y,s-2(y-x)-s')ds'$$
$$+\int_0^{s-2(y-x)}\left[\int_0^{s'} T(x,y,s'')T(x,y,s'-s'')ds''\right]W(x,y,s-2(y-x)-s')ds'\Big\}, \quad s>2(y-x). \qquad (4.12)$$

This equation can be obtained by use of Eqs. (3.8), (3.10), and (3.14).

It is now straightforward to combine Eqs. (3.20) and (4.12) and repeatedly use the resolvent equation (3.18) to get

$$V_y^+(x,y,s) = -\tfrac{1}{2}[A(y)-B(y)]\Big\{t^+(x,y)t^-(x,y)T(x,y,s-2(y-x))$$
$$-\int_0^s R^+(x,y,s')V^-(x,y,s-s')ds'\Big\}, \quad s>2(y-x). \qquad (4.13)$$

The equation for V_x^- is derived similarly. The result is

$$V_x^-(x,y,s) = -\tfrac{1}{2}[A(x)+B(x)]\Big\{t^+(x,y)t^-(x,y)T(x,y,s-2(y-x))$$
$$-\int_0^s R^-(x,y,s')V^+(x,y,s-s')ds'\Big\}, \quad s>2(y-x). \qquad (4.14)$$

Equations (4.13) and (4.14) can be simplified considerably with use of results presented in the next section. The consequences of this simplification will be discussed at the end of Sec. V. An alternative derivation of the results in this section is obtained by considering the dynamics of the propagator matrix for Eq. (2.14). This is carried out in Appendix C.

V. THE EXTENSION OF DATA

The $W(x,y,s)$ and the $V^\pm(x,y,s)$ kernels all share the important feature that they have compact support. More precisely, for times larger than one round trip in the subregion $[x,y]$, i.e., $s>2(y-x)$, these kernels are identically zero,

$$W(x,y,s)=0, \quad s>2(y-x), \qquad (5.1)$$

$$V^\pm(x,y,s)=0, \quad s>2(y-x). \qquad (5.2)$$

These relations are derived in Appendix A.

Now suppose scattering data are known for one round trip through the subregion occupying $[x,y]$. Then Eqs. (5.1) and (5.2) can be used to extend that data beyond one round trip. To be more explicit, consider the end points x and y to be fixed for the moment, and assume that $T(x,y,s)$ is known for times $0<s<2(y-x)$. Equation (3.18), which is a Volterra equation of the second kind for the kernel $W(x,y,s)$, then can be solved for $W(x,y,s)$, $0<s<2(y-x)$. The kernel $W(x,y,s)$ is thus known for all $s>0$ due to Eq. (5.1).

Now assume that $s>2(y-x)$ and rewrite the resolvent Eq. (3.18), using Eq. (5.1), in the following form:

$$T(x,y,s)+\int_{2(y-x)}^s W(x,y,s-s')T(x,y,s')ds'$$
$$=G(x,y,s)=\begin{cases}-\displaystyle\int_{s-2(y-x)}^{2(y-x)} W(x,y,s-s')T(x,y,s')ds', & 2(y-x)<s<4(y-x),\\ 0, & s>4(y-x).\end{cases} \qquad (5.3)$$

Notice that the function $G(x,y,s)$ is known as a function of s for fixed values of x and y with the assumptions made above. Equation (5.3) is a Volterra equation of second kind for $T(x,y,s)$ for $s>2(y-x)$ and this equation can be solved for the unknown $T(x,y,s)$, $s>2(y-x)$. Equation (5.3) thus provides a tool for extending the data $T(x,y,s)$, $0<s<2(y-x)$, to the time interval $s>2(y-x)$.

The extension of the reflection data now follows quite similarly, with the exception that both reflection and transmission data have to be known for one round trip. More precisely, assume that $R^{\pm}(x,y,s)$ and $T(x,y,s)$ are known for $0<s<2(y-x)$. Then from Eq. (3.19) and Eq. (5.2) above, $R^{\pm}(x,y,s)$, $s>2(y-x)$, is expressed as

$$R^{\pm}(x,y,s)=\int_0^{2(y-x)} T(x,y,s-s')V^{\pm}(x,y,s')ds', \quad s>2(y-x). \tag{5.4}$$

However, $V^{\pm}(x,y,s)$, $0<s<2(y-x)$, is related to $R^{\pm}(x,y,s)$ by Eq. (3.20),

$$V^{\pm}(x,y,s)=R^{\pm}(x,y,s)+\int_0^s W(x,y,s-s')R^{\pm}(x,y,s')ds', \quad 0<s<2(y-x). \tag{5.5}$$

Combining these last two equations gives

$$R^{\pm}(x,y,s)=\int_0^{2(y-x)} T(x,y,s-s')\left[R^{\pm}(x,y,s')+\int_0^{s'} W(x,y,s'-s'')R^{\pm}(x,y,s'')ds''\right]ds', \quad s>2(y-x). \tag{5.6}$$

Notice that in this last equation, reflection data $R^{\pm}(x,y,s)$ are used only for times less than one round trip, i.e., $0<s<2(y-x)$, while transmission data, $T(x,y,s)$, are used for all $s>0$. However, for times beyond one round trip the transmission data can be extended with the technique discussed above in Eq. (5.3). These ideas will be exploited in a special context in Part II.

An alternate approach to the extension of the reflection data is to rewrite Eq. (3.20) for $s>2(y-x)$ and use Eq. (5.2) to obtain

$$R^{\pm}(x,y,s)+\int_{2(y-x)}^{s} W(x,y,s-s')R^{\pm}(x,y,s')ds'$$
$$=\begin{cases} -\displaystyle\int_{s-2(y-x)}^{2(y-x)} W(x,y,s-s')R^{\pm}(x,y,s')ds', & 2(y-x)<s<4(y-x),\\ 0, & s>4(y-x).\end{cases}$$

Thus far, the compact support of the kernels $V^{\pm}$ has not been used in the imbedding equations derived in Sec. IV. Now using the fact that $V^{\pm}$ vanish identically for $s>2(y-x)$ reduces Eq. (4.13) and (4.14) to the following new representation of the transmission kernel T for $s>2(y-x)$:

$$t^{+}(x,y)t^{-}(x,y)T(x,y,s-2(y-x))$$
$$=\int_0^{2(y-x)} V^{+}(x,y,s')R^{-}(x,y,s-s')ds'=\int_0^{2(y-x)} V^{-}(x,y,s')R^{+}(x,y,s-s')ds', \quad s>2(y-x). \tag{5.7}$$

VI. SUMMARY AND CONCLUSIONS

In this paper some mathematical tools for transient wave propagation in lossy media have been introduced. This work primarily focuses on the direct scattering problem and the properties of the scattering operators. However, many of the equations developed in the present paper are of importance for the inverse algorithm presented in Part II (See Ref. 1).

Reciprocity is shown to imply that the two transmission kernels $T^{\pm}$ are proportional to each other, see Eq. (3.14). This property simplifies the analysis considerably and also reduces the number of independent imbedding equations for the scattering kernels.

The propagator kernels for the medium are also introduced and some of their properties are exploited. In Sec. IV the new imbedding equations for these kernels are derived in Eqs. (4.3), (4.4), and (4.7)–(4.10).

One of the main results in this paper, the compact support of the propagator kernels, has several consequences. In Sec. V this property is shown to provide a way to extend transmission data from one round trip to arbitrary time, see Eq. (5.3). This extension is also possible for the reflection data provided transmission data are available, see Eq. (5.6).

The compact support of W provides an important limitation of the functional behavior of the transmission kernel T. To be an admissible transmission kernel T for the model considered in this paper, its resolvent W also must have compact support. This observation provides an important characterization of the transmission kernel T. Furthermore, only data for times less than one round trip are needed for this characterization, due to the extension of data discussed above. This implies that all information available in the transmission kernel is contained in the time interval up to one round trip and that if it is admissible or not is based upon the functional behavior in this finite interval. Unfortunately, the compact support of $V^{\pm}$ doe not imply any simple characterizations for $R^{\pm}$. However, from the new imbedding equations for $V^{\pm}$ and the compact support of $V^{\pm}$ a new representation of the transmission kernel T is obtained, see Eq. (5.7), which relates T to $V^{\pm}$ and $R^{\pm}$.

ACKNOWLEDGMENTS

This work was performed at Ames Laboratory, U.S.D.O.E., Iowa State University, and was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under Contract No. W-7405-ENG-82, and by the Office of Naval Research Contract No. N0014-83-K-0038. One of the authors (G.K.) would like to thank Ames Laboratory for their hospitality during his visit in the United States. He is also grateful to the National Swedish Board for Technical Development (STUF) for partial support.

APPENDIX A: COMPACT SUPPORT OF W AND $V^{\pm}$

In this section it is shown that the kernels W and $V^{\pm}$ have compact support. This fact was introduced in Sec. V [Eqs. (5.1) and (5.2)]. The arguments given below suffice to show that the kernels $W(x,y,s)$ and $V^{\pm}(x,y,s)$ vanish for $s>2(y-x)$. The compact support then follows from causality, which implies that these kernels also vanish for $s<0$.

In the model problem given in Eq. (2.1), the velocity $c(z)$ is assumed to be continuous at the end points of the slab, $z=0$ and $z=L$. For the sake of proving a stronger result, which should be useful in later work, this assumption will be relaxed. Thus, $c(z)$ can have finite jump discontinuities at $z=0$ and $z=L$. This generalization alters the transformed problem given by Eq. (2.14) in that u_x is no longer continuous at $x=0$ and $x=1$. Instead, u_x satisfies the relations

$$\begin{aligned} c_0 u_x(0^-,s) &= u_x(0^+,s),\\ c_1 u_x(1^+,s) &= u_x(1^-,s), \end{aligned} \tag{A1}$$

where

$$\begin{aligned} c_0 &= c(0^+)/c(0^-),\\ c_1 &= c(L^-)/c(L^+). \end{aligned} \tag{A2}$$

It is now shown that for this more general problem, $W(0,1,s)$ vanishes for $s>2$. (The arguments given below clearly generalize to any subregion $[x,y]$ of the slab.) In order to produce the desired result, an explicit formula for W will be derived.

Being by expressing the solution u of Eq. (2.14) in terms of transmission data $u^t_+(s)$ via

$$\begin{aligned} u(x,s) = \tfrac{1}{2}[t^+(x,1)]^{-1}\Big\{&(c_1+1)u^t_+(s-x)\\ &-(c_1-1)u^t_+(s+x-2)\exp\Big[\int_x^1 B(s')ds'\Big]\\ &+\int_x^{2-x} u^t_+(s-s')N(x,s')ds'\Big\}, \end{aligned} \tag{A3}$$

for $0<x<1$. Equation (A3) is derived in Ref. 15. The function $N(x,s)$ is related to the Riemann function for Eq. (2.14) and satisfies

$$N_{xx}-N_{ss}+B(x)(N_x+N_s)+D_+(x)N=0,\quad 0<x<1, \tag{A4}$$

and boundary conditions for $0<x<1$,

$$\begin{aligned} 2N(x,x) &= c_1B(1^-)-A(1^-)-(c_1+1)\int_x^1 D_+(s')ds',\\ 2N(x,2-x)& \\ &= \Big[c_1B(1^-)-A(1^-)+(c_1-1)\int_x^1 D_-(s')ds'\Big]\\ &\quad\times\exp\Big[\int_x^1 B(s')ds'\Big], \end{aligned} \tag{A5}$$

where

$$D_\pm(x)=\tfrac{1}{4}(B^2-A^2)+\tfrac{1}{2}(-A'\pm B'). \tag{A6}$$

The prime in Eq. (A6) denotes differentiation with respect to x.

Differentiate Eq. (A3) with respect to x and set $x=0$ in the resulting equation. Rewrite the left-hand side in terms of $u^{i\prime}_+$ and $u^{r\prime}_+$ (differentiation with respect to the argument) using

$$u_x(0^+,s)=c_0[-u^{i\prime}_+(s)+u^{r\prime}_+(s)]. \tag{A7}$$

Now integrate this equation from 0 to s, using the assumption that $u^i_+(0)=u^r_+(0)=u^t_+(0)=0$, and obtain

$$\begin{aligned} 2c_0[-u^i_+(s)+u^r_+(s)]& \\ = [t^+(0,1)]^{-1}\Big\{&-(c_1+1)u^t_+(s)\\ &-(c_1-1)u^t_+(s-2)\exp\Big[\int_0^1 B(s')ds'\Big]\\ &+\int_0^s u^t_+(s')F(s-s')ds'\Big\}, \end{aligned} \tag{A8}$$

where

$$\begin{aligned} F(s)=a+bH(s-2)+\int_0^s[&N_x(0,s')\\ &-\tfrac{1}{2}(A-B)|_{0^+}N(0,s')]H(2-s')ds' \end{aligned} \tag{A9}$$

and

$$\begin{aligned} a &= -\tfrac{1}{2}(c_1+1)(A-B)|_{0^+}-N(0,0),\\ b &= \tfrac{1}{2}(c_1-1)(A+B)|_{0^+}\exp\Big[\int_0^1 B(s')ds'\Big]-N(0,2),\\ H(s) &= \text{Heaviside function} = \begin{cases} 0, & s<0,\\ 1, & s>0. \end{cases} \end{aligned}$$

Now evaluate Eq. (A3) at $x=0^+$ and rewrite the left-hand side as $u^i_+(s)+u^r_+(s)$. Use Eq. (A8) to eliminate $u^r_+(s)$ from the resulting equation and thus obtain

$$\begin{aligned} u^i_+(s)=[t^+(0,1)]^{-1}\Big\{&\frac{(c_0+1)(c_1+1)u^t_+(s)}{4c_0}\\ &-\frac{(c_0-1)(c_1-1)u^t_+(s-2)\exp[\int_0^1 B(s')ds']}{4c_0}\\ &+\int_0^s u^t_+(s')W(0,1,s-s')ds'\Big\}, \end{aligned} \tag{A10}$$

where the W kernel is given by

$$W(0,1,s)=[c_0N(0,s)H(2-s)-F(s)]/4c_0. \tag{A11}$$

Equation (A10) is the generalization of Eq. (2.23) when

$c(z)$ is discontinuous as $z=0$ and $z=L$. Notice from Eqs. (A11) and (A9) that $W(0,1,s)$ is constant for $s>2$, a fact which also follows from the domain of dependence arguments. To evaluate that constant, set

$$f(x) = \int_x^{2-x}\left[N_x(x,s') - \frac{1}{2}(A-B)\Big|_x N(x,s')\right]ds' \tag{A12}$$

so that, from Eq. (A11),

$$W(0,1,s) = [k - f(0)]/4c_0, \quad s>2, \tag{A13}$$

where

$$k = N(0,0) + N(0,2) + \tfrac{1}{2}(c_1+1)(A-B)|_{0^+} - \frac{1}{2}(c_1-1)(A+B)\Big|_{0^+}\exp\left[\int_0^1 B(s')ds'\right]. \tag{A14}$$

The constant k is known from the boundary conditions Eq. (A5), so it remains to determine $f(0)$.

Differentiate Eq. (A12) with respect to x and eliminate the N_{xx} term by using Eq. (A4). Upon performing the s' integrations, it follows that

$$f'(x) + \tfrac{1}{2}(A+B)|_x f(x) = g(x), \tag{A15}$$

where

$$g(x) = -\frac{d}{dx}N(x,x) - \frac{d}{dx}N(x,2-x) + \tfrac{1}{2}(A+B)|_x N(x,x) + \tfrac{1}{2}(A-3B)|_x N(x,2-x). \tag{A16}$$

Solving Eq. (A15) yields

$$f(0) = -\int_0^1 g(x)\exp\left[\frac{1}{2}\int_0^x [A(x') + B(x')]dx'\right]dx. \tag{A17}$$

Using Eqs. (A17) and (A5), a tedious calculation now shows that

$$f(0) = k. \tag{A18}$$

Hence, $W(0,1,s) = 0$ for $s>2$.

Fortunately, this calculation does not need to be repeated to verify the compact support of the $V^\pm$ kernels. Instead observe that if Eq. (A8) is used to eliminate $u'_+(s)$ from Eq. (A3), then

$$V^+(0,1,s) = -W(0,1,s), \quad s>2. \tag{A19}$$

In a similar manner it follows that

$$V^-(0,1,s) = -W(0,1,s), \quad s>2. \tag{A20}$$

APPENDIX B: EXACT REPRESENTATIONS OF THE SCATTERING AND PROPAGATOR KERNELS FOR CONSTANT *A*(*x*) AND *B*(*x*)

For constant $A(x)$ and $B(x)$ the scattering kernels $R^\pm$, T, W, and $V^\pm$ can be determined analytically. Throughout this appendix it is therefore assumed that the function $A(x) = A$ and $B(x) = B$, for $0<x<1$, where A and B are real constants. For the convenience of the reader the basic equations [see Eqs. (3.1), (3.5), (3.15), (4.3), (4.7), and (4.8)] that are used in this appendix are repeated here:

$$\begin{aligned} &R_x^+(x,y,s) \\ &\quad = 2R_s^+(x,y,s) - BR^+(x,y,s) \\ &\qquad - \frac{1}{2}(A+B)\int_0^s R^+(x,y,s')R^+(x,y,s-s')ds', \\ &\qquad s>0, \end{aligned} \tag{B1}$$

$$R^+(y,y,s) = 0, \quad s>0,$$
$$R^+(x,y,0^+) = -\tfrac{1}{4}(A-B), \quad x<y;$$

$$\begin{aligned} &R_y^-(x,y,s) \\ &\quad = -2R_s^-(x,y,s) + BR^-(x,y,s) \\ &\qquad - \frac{1}{2}(A-B)\int_0^s R^-(x,y,s')R^-(x,y,s-s')ds', \\ &\qquad s>0, \end{aligned} \tag{B2}$$

$$R^-(x,x,s) = 0, \quad s>0,$$
$$R^-(x,y,0^+) = \tfrac{1}{4}(A+B), \quad x<y;$$

$$\begin{aligned} &T_x(x,y,s) \\ &\quad = -\tfrac{1}{2}(A+B)\left[R^+(x,y,s) + \int_0^s T(x,y,s')R^+(x,y,s-s')ds'\right], \quad s>0, \end{aligned} \tag{B3}$$

$$T(y,y,s) = 0, \quad s>0;$$

$$\begin{aligned} &W_x(x,y,s) \\ &\quad = \tfrac{1}{2}(A+B)\left[R^+(x,y,s) + \int_0^s W(x,y,s')R^+(x,y,s-s')ds'\right], \quad s>0, \end{aligned} \tag{B4}$$

$$W(y,y,s) = 0, \quad s>0;$$

$$\begin{aligned} &V_x^+(x,y,s) \\ &\quad = 2V_s^+(x,y,s) - BV^+(x,y,s) \\ &\qquad + \tfrac{1}{2}(A-B)W(x,y,s), \\ &V^+(y,y,s) = 0, \quad s>0, \\ &V^+(x,y,0^+) = -\tfrac{1}{4}(A-B), \quad x<y; \end{aligned} \tag{B5}$$

$$\begin{aligned} &V_x^-(x,y,s) \\ &\quad = -2V_s^-(x,y,s) + BV^-(x,y,s) \\ &\qquad + \tfrac{1}{2}(A-B)W(x,y,s), \\ &V^-(x,x,s) = 0, \quad s>0, \\ &V^-(x,y,0^+) = \tfrac{1}{4}(A+B), \quad x<y. \end{aligned} \tag{B6}$$

These equations can be solved by a Laplace transformation in the "time" variable s. The Laplace transform of a function f is indicate $\hat{f}$ or $\Lambda[f]$ and the transformed time variable is denoted by p. In this notation the x and y dependence of transformed functions are suppressed for convenience. Equation (B1) transforms into the Riccati equation

$$\begin{aligned} &\hat{R}_x^+(p) - (2p-B)\hat{R}^+(p) - \tfrac{1}{2}(A-B) \\ &\qquad + \tfrac{1}{2}(A+B)\hat{R}^{+2}(p) = 0, \\ &\hat{R}^+(p) = 0, \quad x=y, \end{aligned} \tag{B7}$$

with solution

$$\hat{R}^{+}(p) = -\frac{(1-e^{-\beta})\gamma}{(A+B)(1+\delta e^{-\beta})}$$

$$= -\frac{\gamma}{(A+B)}\left\{1-(1+\delta)\sum_{n=1}^{\infty}(-\delta)^{n-1}e^{-n\beta}\right\}, \qquad \text{(B8)}$$

where

$$\begin{aligned} \alpha &= \tfrac{1}{2}(A^2-B^2)^{1/2},\\ \beta &= (y-x)((2p-B)^2+4\alpha^2)^{1/2}, \qquad \text{(B9)}\\ \gamma &= ((2p-B)^2+4\alpha^2)^{1/2}-2p+B,\\ \delta &= \gamma^2/4\alpha^2. \end{aligned}$$

It is now rather straightforward to invert each term in the bracket in Eq. (B8) with the use of the identity

$$\hat{f}((p^2+\alpha^2)^{1/2}) = \hat{f}(p) - \Lambda\left\{\alpha\int_0^s f(u)\, J_1(\alpha(s^2-u^2)^{1/2}) \times (s^2-u^2)^{-1/2}u\, du\right\}(p). \qquad \text{(B10)}$$

The final result is

$$\begin{aligned} R^{+}(x,y,s) &= -\tfrac{1}{2}(A-B)e^{Bs/2}\sum_{n=0}^{\infty}(-1)^n H(s-2n(y-x))\\ &\quad\times\Big\{S_n(s-2n(y-x))\\ &\quad - 2n(y-x)\alpha^2\int_{2n(y-x)}^{s} S_n(s'-s)\\ &\quad\times\frac{J_1(\alpha(s'^2-4n^2(y-x)^2)^{1/2})}{\alpha(s'^2-4n^2(y-x)^2)^{1/2}}ds'\Big\}, \qquad \text{(B11)} \end{aligned}$$

where

$$\begin{aligned} S_n(s) &= [(2n-1)\,J_{2n-1}(\alpha s)\\ &\quad + (2n+1)\,J_{2n+1}(\alpha s)]/\alpha s, \quad n=1,2,3,..., \qquad \text{(B12)}\\ S_0(s) &= J_1(\alpha s)/\alpha s, \end{aligned}$$

where J_n is the Bessel function of order n. Some plots of R^{+} are shown in Fig. 5.

The reflection kernel from the right-hand side $R^{-}(x,y,s)$ is easily obtained by replacing A with $-A$ in the equations above. With the same definitions of $S_n(s)$ as above, the result is

$$\begin{aligned} R^{-}(x,y,s) &= \tfrac{1}{2}(A+B)e^{Bs/2}\sum_{n=0}^{\infty}(-1)^n H(s-2n(y-x))\\ &\quad\times\Big\{S_n(s-2n(y-x))\\ &\quad - 2n(y-x)\alpha^2\int_{2n(y-x)}^{s} S_n(s'-s)\\ &\quad\times\frac{J_1(\alpha(s'^2-4n^2(y-x)^2)^{1/2})}{\alpha(s'^2-4n^2(y-x)^2)^{1/2}}ds'\Big\}. \qquad \text{(B13)} \end{aligned}$$

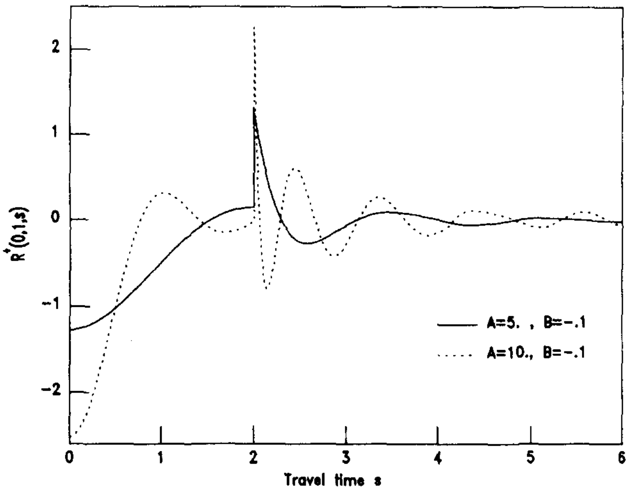

FIG. 5. The reflection kernel $R^{+}(0,1,s)$ for three round trips in the medium. Two examples with constant A and B profiles are shown.

The transmission kernel $T(x,y,s)$ is obtained from the Laplace transform of Eq. (B3). The solution in the transformed time variable p is

$$\begin{aligned} \hat{T}(p) &= -1+\frac{1+\delta}{1+\delta e^{-\beta}}e^{-\gamma(y-x)/2}\\ &= e^{-\gamma(y-x)/2}(1+\delta)-1\\ &\quad + e^{-\gamma(y-x)/2}(1+\delta)\sum_{n=1}^{\infty}(-\delta)^n e^{-n\beta}. \qquad \text{(B14)} \end{aligned}$$

The inversion of this equation leads to rather cumbersome algebra. The following identity is of great help:

$$\hat{f}(\alpha+(p^2+\alpha^2)^{1/2}-p)-\hat{f}(\alpha) = \Lambda\left\{-\alpha^2\int_0^{\infty}e^{-\alpha u}\frac{J_1(\alpha(s^2+2us)^{1/2})}{\alpha(s^2+2us)^{1/2}}uf(u)du\right\}(p). \qquad \text{(B15)}$$

The result of the inversion is

$$\begin{aligned} T(x,y,s) &= -\alpha^2 e^{Bs/2}\sum_{n=0}^{\infty}(-1)^n H(s-2n(y-x))\\ &\quad\times\Big\{P_n(s-2n(y-x))\\ &\quad - 2n(y-x)\alpha^2\int_{2n(y-x)}^{s} P_n(s-s')\\ &\quad\times\frac{J_1(\alpha(s'^2-4n^2(y-x)^2)^{1/2})}{\alpha(s'^2-4n^2(y-x)^2)^{1/2}}ds'\Big\}, \qquad \text{(B16)} \end{aligned}$$

where

$$\begin{aligned} P_n(s) &= \left[\left(\frac{\partial}{\alpha\,\partial u}\right)^{2n}+\left(\frac{\partial}{\alpha\,\partial u}\right)^{2n+2}\right]\\ &\quad\times\left[u\frac{J_1(\alpha(s^2+2us)^{1/2})}{\alpha(s^2+2us)^{1/2}}\right]_{u=2(y-x)}\\ &= 2w^{-2n-4}(\alpha s)^{2n}\{\alpha^2 s[4n(y-x)\\ &\quad\times(s+y-x)-s^2]J_{2n+2}(w) \end{aligned}$$

$$+w[(y-x)^2\alpha^2 s-2n(2n+1)$$
$$\times(s+2(y-x))]J_{2n+1}(w)\} \tag{B17}$$

and

$$w=\alpha(s^2+2s(y-x))^{1/2}.$$

Plots of T are shown in Fig. 6.

The solution to Eq. (B4) in the transformed time variable p is

$$\widehat{W}(p)=-1+\frac{1+\delta e^{-\beta}}{1+\delta}e^{\gamma(y-x)/2}. \tag{B18}$$

The inversion of this equation gives

$$W(x,y,s)=\tfrac{1}{2}H(2(y-x)-s)e^{Bs/2}\alpha^2[2(y-x)-s]$$
$$\times\frac{I_1(\alpha(2(y-x)s-s^2)^{1/2})}{\alpha(2(y-x)s-s^2)^{1/2}}. \tag{B19}$$

Proceeding to the V^+ equation, the result in the transformed variable p is

$$\widehat{V}^+(x,y,p)=-\tfrac{1}{2}(A-B)((2p-B)^2+4\alpha^2)^{-1/2}$$
$$\times e^{\gamma(y-x)/2}[1-e^{-\beta}], \tag{B20}$$

with inverse

$$V^+(x,y,s)=-\tfrac{1}{4}(A-B)H(2(y-x)-s)e^{Bs/2}$$
$$\times I_0(\alpha(2(y-x)s-s^2)^{1/2}). \tag{B21}$$

The corresponding result for $V^-(x,y,s)$ is easily obtained by replacing A with $-A$ in the equation above. The result is

$$V^-(x,y,s)=\tfrac{1}{4}(A+B)H(2(y-x)-s)e^{Bs/2}$$
$$\times I_0(\alpha(2(y-x)s-s^2)^{1/2}). \tag{B22}$$

The compact support of the kernels W and $V^\pm$, which is a general property derived in Appendix A, is clearly seen in Eqs. (B19), (B21), and (B22). Typical examples of the kernels W and V^+ are shown in Figs. 7 and 8.

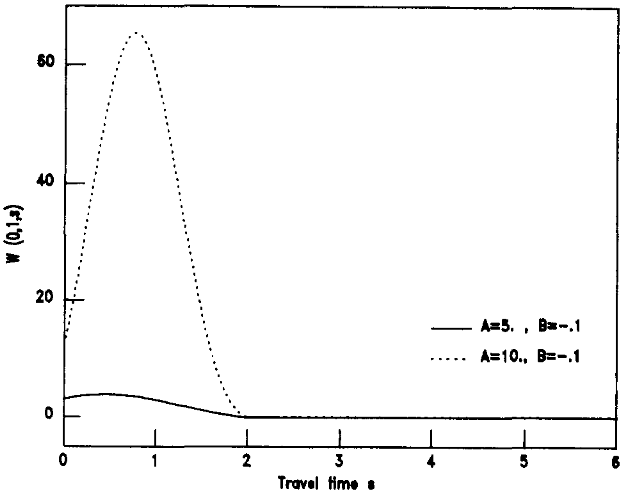

FIG. 7. The same as Fig. 5 but showing the propagation kernel $W(0,1,s)$.

Now having the explicit expressions for the kernels $R^\pm$, T, W, and $V^\pm$ for the case when A and B are constants, it is instructive to verify some of the basic equations in this paper for this special case. For example, the jump in the kernels $R^\pm$ along the plane $s=2(y-x)$ given by Eq. (3.10) is easily verified from Eqs. (B11) and (B13). It is also easy to verify Eq. (3.11).

Equations (4.9) and (4.10) can also be verified in this special case of constant A and B. These equations are equivalent to the Bessel function identity

$$\frac{d}{dx}J_0((t^2-2xt)^{1/2})$$
$$=\int_0^t J_0((t'^2-2xt')^{1/2})\frac{J_1(t-t')}{t-t'}dt'. \tag{B23}$$

This identity can be proved by showing that both sides are the same entire function in x.

APPENDIX C: PROPAGATOR DYNAMICS

In this appendix the dynamics of the kernels V^+ and W are derived in an alternate manner that does not utilize the dynamics of $R^\pm$ and T. This derivation depends on a contin-

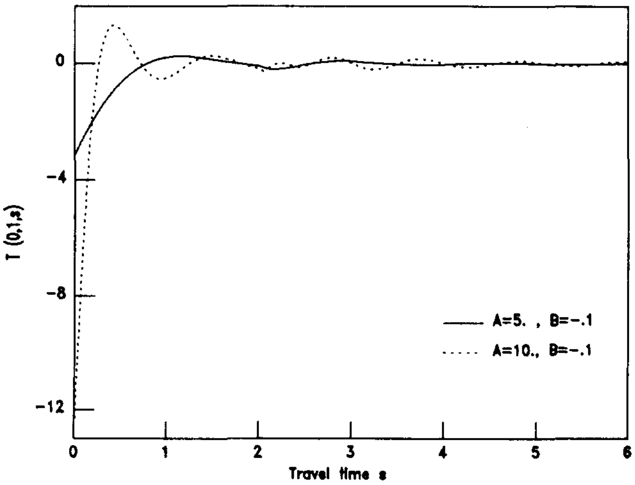

FIG. 6. The same as Fig. 5 but showing the transmission kernel $T(0,1,s)$.

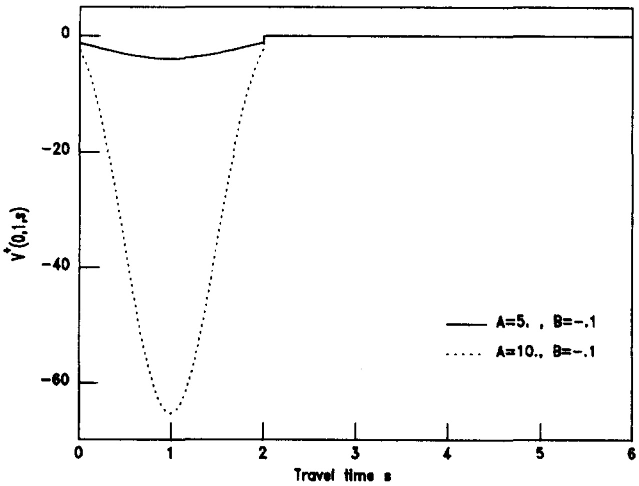

FIG. 8. The same as Fig. 5 but showing the propagation kernel $V^+(0,1,s)$.

uous representation of the propagation operator.

For the sake of convenience, the transformed problem [Eq. (2.14)] will be the starting point for this analysis, although the derivation could be carried out in terms of the physical variables that appear in Eq. (2.1). The analysis presented here is similar in spirit to that given elsewhere[3,5,10] for derivation of the scattering operator equations. The independent variable x in Eq. (2.14) will be replaced by the dummy variable z, since x is used to denote the end point of a subregion $[x,y]$. The variable z should not be confused with that appearing in Eq. (2.1). Begin by introducing a splitting[23] of the field $u(z,s)$ in Eq. (2.14), defined by

$$u^{\pm}(z,s)=\tfrac{1}{2}[u(z,s)\mp\partial_s^{-1}u_z(z,s)], \qquad \text{(C1)}$$

where

$$\partial_s^{-1}u_z(z,s)=\int_{-\infty}^{s}u_z(z,s')ds'.$$

In a homogeneous medium, this splitting merely reduces the field $u(z,s)$ into right moving (+) and left moving (−) waves. More generally, Eq. (C1) is a change of basis from $(u,u_z)^T$ to $(u^+,u^-)^T$ for Eq. (2.14). In this new basis, Eq. (2.14) becomes

$$\frac{\partial}{\partial z}\begin{bmatrix}u^+(z,s)\\u^-(z,s)\end{bmatrix}=\begin{bmatrix}\alpha(z)&\beta(z)\\\gamma(z)&\delta(z)\end{bmatrix}\begin{bmatrix}u^+(z,s)\\u^-(z,s)\end{bmatrix}$$
$$\equiv D(z)\begin{bmatrix}u^+(z,s)\\u^-(z,s)\end{bmatrix}, \qquad \text{(C2)}$$

where

$$\begin{aligned}\alpha(z)&=-\frac{1}{2}[A(z)-B(z)]-\frac{\partial}{\partial s},\\ \beta(z)&=\tfrac{1}{2}[A(z)+B(z)],\\ \gamma(z)&=\tfrac{1}{2}[A(z)-B(z)],\\ \delta(z)&=-\frac{1}{2}[A(z)+B(z)]+\frac{\partial}{\partial s}.\end{aligned} \qquad \text{(C3)}$$

Now consider a subregion $[x,y]$ of the original slab. Let $P(x,y)$ denote the propagator for the subregion $[x,y]$; i.e., P is a 2×2 matrix of operators that maps the field at y over to the field at x,

$$\begin{bmatrix}u^+(x,s)\\u^-(x,s)\end{bmatrix}=P(x,y)\begin{bmatrix}u^+(y,s)\\u^-(y,s)\end{bmatrix}. \qquad \text{(C4)}$$

Now differentiate Eq. (C4) with respect to x to obtain

$$\frac{\partial}{\partial x}\begin{bmatrix}u^+(x,s)\\u^-(x,s)\end{bmatrix}=\frac{\partial P(x,y)}{\partial x}\begin{bmatrix}u^+(y,s)\\u^-(y,s)\end{bmatrix}. \qquad \text{(C5)}$$

Use Eq. (C2) (evaluated at $z=x$) and (C4) to express the left-hand side of (C5) in terms of the fields at y:

$$D(x)P(x,y)\begin{bmatrix}u^+(y,s)\\u^-(y,s)\end{bmatrix}=\frac{\partial P(x,y)}{\partial x}\begin{bmatrix}u^+(y,s)\\u^-(y,s)\end{bmatrix}. \qquad \text{(C6)}$$

Since $(u^+(y,s),u^-(y,s))^T$ can be chosen arbitrarily, it follows that

$$\frac{\partial P(x,y)}{\partial x}=D(x)P(x,y). \qquad \text{(C7)}$$

It can similarly be shown that

$$\frac{\partial P(x,y)}{\partial y}=-P(x,y)D(y). \qquad \text{(C8)}$$

Having now obtained differential equations (C7) and (C8) for the propagator, a representation for the entries of P is required. In order to use a representation compatible with that in Sec. III, let $u^{\pm}$ be represented by

$$u^+(z,s)=\begin{cases}u^i_+\,(s-z+x), & z<x,\\ u^t_+\,(s-z+x)+u^r_-\,(s-z+y), & z>y,\end{cases} \qquad \text{(C9)}$$

$$u^-(z,s)=\begin{cases}u^r_+\,(s+z-x)+u^t_-\,(s+z-y), & z<x,\\ u^i_-\,(s+z-y), & z>y.\end{cases} \qquad \text{(C10)}$$

The fields on the right-hand sides of Eqs. (C9) and (C10) are related by [cf. Eqs. (2.17)–(2.20), (2.23), and (2.24) for the special case $x=0$ and $y=1$]

$$\begin{aligned}u^r_{\pm}\,(s)&=[\mathscr{R}^{\pm}(x,y)u^i_{\pm}\,(\cdot)](s)\\ &=\int_{-\infty}^{s}R^{\pm}(x,y,s-s')u^i_{\pm}\,(s')ds',\end{aligned} \qquad \text{(C11)}$$

$$\begin{aligned}u^t_{\pm}\,(s)&=[\mathscr{T}^{\pm}(x,y)u^i_{\pm}\,(\cdot)](s)\\ &=t^{\pm}(x,y)\Big[u^i_{\pm}\,(s)\\ &\quad+\int_{-\infty}^{s}T(x,y,s-s')u^i_{\pm}\,(s')ds'\Big],\end{aligned} \qquad \text{(C12)}$$

$$\begin{aligned}u^r_{+}\,(s)&=[\mathscr{V}^{+}(x,y)u^t_{+}\,(\cdot)](s)\\ &=[t^{+}(x,y)]^{-1}\int_{-\infty}^{s}V^{+}(x,y,s-s')u^t_{+}\,(s')ds',\end{aligned} \qquad \text{(C13)}$$

$$\begin{aligned}u^i_{+}\,(s)&=[\mathscr{W}^{+}(x,y)u^t_{+}\,(\cdot)](s)\\ &=[t^{+}(x,y)]^{-1}\Big[u^t_{+}\,(s)\\ &\quad+\int_{-\infty}^{s}W(x,y,s-s')u^t_{+}\,(s')ds'\Big].\end{aligned} \qquad \text{(C14)}$$

The relations Eqs. (3.14) and (3.17) have been used in Eqs. (C12) and (C14), respectively, and $t^{\pm}(x,y)$ is defined in Eq. (3.9). In Eqs. (C11)–(C14) it is assumed that the fields are quiescent prior to some finite time s_0, although s_0 is not necessarily zero.

It is also convenient to introduce a shift operator Q, whose action on a function of the s variable is defined by

$$Q(y,x)f(s)=f(s+x-y).$$

Repeated applications of Q have the obvious interpretation:

$$\begin{aligned}Q^2(y,x)f(s)&=Q(y,x)[Q(y,x)f(s)]\\ &=Q(y,x)f(s+x-y)\\ &=f(s+2x-2y),\\ Q(x,y)Q(y,x)f(s)&=f(s).\end{aligned}$$

Still confining attention to the subregion $[x, y]$, it now follows that

$$\begin{aligned}u^+(x,s) &= u^i_+(s)\\ &= [\mathscr{W}^+(x,y)u^i_+(\cdot)](s)\\ &= [\mathscr{W}^+(x,y)\{u^+(y,\cdot) - u^r_-(\cdot)\}](s+y-x)\\ &= Q(x,y)[\mathscr{W}^+(x,y)u^+(y,\cdot)](s)\\ &\quad - Q(x,y)[\mathscr{W}^+(x,y)\mathscr{R}^-(x,y)u^-(y,\cdot)](s),\end{aligned} \tag{C15}$$

and

$$\begin{aligned}u^-(x,s) &= u^r_+(s) + u^t_-(s+x-y)\\ &= [\mathscr{V}^+(x,y)u^i_+(\cdot)](s) + Q(y,x)\\ &\quad\times[\mathscr{T}^-(x,y)u^-(y,\cdot)](s)\\ &= Q(x,y)[\mathscr{V}^+(x,y)u^+(y,\cdot)](s)\\ &\quad - Q(x,y)[\mathscr{V}^+(x,y)\mathscr{R}^-(x,y,)u^-(y,\cdot)](s)\\ &\quad + Q(y,x)[\mathscr{T}^-(x,y)u^-(y,\cdot)](s).\end{aligned} \tag{C16}$$

Using Eqs. (C15) and (C16), the propagator can be written in the explicit form

$$P(x,y) = Q(x,y)\begin{bmatrix}\mathscr{W}^+(x,y) & -\mathscr{W}^+(x,y)\mathscr{R}^-(x,y)\\ \mathscr{V}^+(x,y) & Q^2(y,x)\mathscr{T}^-(x,y) - \mathscr{V}^+(x,y)\mathscr{R}^-(x,y)\end{bmatrix}. \tag{C17}$$

In order to pass from the operator equations (C7) and (C8) to equations involving the kernels W and V^+, it is easier to consider these equations at the level of Eq. (C6). Setting $u^-(y,s)\equiv 0$ yields

$$\begin{aligned}&Q(y,x)\frac{\partial}{\partial x}[Q(x,y)\mathscr{W}^+(x,y)u^+(y,\cdot)](s)\\ &\quad= [(\alpha(x)\mathscr{W}^+(x,y) + \beta(x)\mathscr{V}^+(x,y))u^+(y,\cdot)](s),\\ &Q(y,x)\frac{\partial}{\partial x}[Q(x,y)\mathscr{V}^+(x,y)u^+(y,\cdot)](s)\\ &\quad= [(\gamma(x)\mathscr{W}^+(x,y) + \delta(x)\mathscr{V}^+(x,y))u^+(y,\cdot)](s).\end{aligned}$$

Expressing these equations in terms of kernels produces Eqs. (4.5) and (4.7). Similarly, applying Eq. (C8) to $(u^+(y,s),0)^T$ yields Eqs. (4.4), (4.9), and (4.13) and verifies that the jump in $V^+(x,y,s)$ at $s = 2(y-x)$ is the same as that in $R^+(x,y,s)$, as given in Eq. (3.10). This last fact also follows from Eq. (3.20).

With the notation established above, it is now easy to compare the action of the propagator matrix $P(x,y)$ with that of the scattering matrix $S(x,y)$. Here S relates the ± components of u according to

$$\begin{bmatrix}u^+(y,s)\\ u^-(x,s)\end{bmatrix} = S(x,y)\begin{bmatrix}u^+(x,s)\\ u^-(y,s)\end{bmatrix},$$

and is represented by

$$S(x,y) = \begin{bmatrix}\mathscr{T}^+(x,y) & \mathscr{R}^-(x,y)\\ \mathscr{R}^+(x,y) & \mathscr{T}^-(x,y)\end{bmatrix}.$$

It can be shown (see Ref. 3, 5, or 10) that S satisfies

$$\begin{aligned}\frac{\partial S}{\partial x} &= -\begin{bmatrix}\mathscr{T}^+(x,y) & 0\\ \mathscr{R}^+(x,y) & I\end{bmatrix}\begin{bmatrix}\alpha(x) & \beta(x)\\ -\gamma(x) & -\delta(x)\end{bmatrix}\\ &\quad\times\begin{bmatrix}I & 0\\ \mathscr{R}^+(x,y) & \mathscr{T}^-(x,y)\end{bmatrix},\\ \frac{\partial S}{\partial y} &= -\begin{bmatrix}I & \mathscr{R}^-(x,y)\\ 0 & \mathscr{T}^-(x,y)\end{bmatrix}\begin{bmatrix}\alpha(y) & \beta(y)\\ -\gamma(y) & -\delta(y)\end{bmatrix}\\ &\quad\times\begin{bmatrix}\mathscr{T}^+(x,y) & \mathscr{R}^-(x,y)\\ 0 & I\end{bmatrix},\end{aligned} \tag{C18}$$

where 0 and I denote the zero and identity operators, respectively. Equations (3.1)–(3.8) can be obtained from the operator equations (C18) by rewriting the latter in terms of the representations (C11) and (C12).

[1] G. Kristensson and R. J. Krueger, "Direct and inverse scattering in the time domain for a dissipative wave equation. II. Simultaneous reconstruction of dissipation and phase velocity profiles," J. Math. Phys. **27**, 1683 (1986).

[2] J. P. Corones and R. J. Krueger, "Obtaining scattering kernels using invariant imbedding," J. Math. Anal. Appl. **95**, 393 (1983).

[3] M. Davison, "A general approach to splitting and invariant imbedding techniques for linear wave equations," Ames Laboratory preprint, to appear in J. Math. Anal. Appl.

[4] J. P. Corones, M. E. Davison, and R. J. Krueger, "Direct and inverse scattering in the time domain via invariant imbedding equations," J. Acoust. Soc. Am. **74**, 1535 (1983).

[5] J. P. Corones, M. E. Davison, and R. J. Krueger, "Wave splittings, invariant imbedding and inverse scattering," in *Inverse Optics, Proceedings of the SPIE*, Vol. 413, edited by A. J. Devaney (SPIE, Bellingham, WA, 1983), pp. 102–106.

[6] J. P. Corones, "Wave splitting and invariant imbedding in direct and inverse scattering," in *Wave Propagation in Homogeneous Media and Ultrasonic Nondestructive Evaluation*, AMD-Vol. 62, edited by G. C. Johnson (ASME, New York, 1984), pp. 31–35.

[7] J. P. Corones, R. J. Krueger, and C. R. Vogel, "The effects of noise and bandlimiting in a one dimensional time dependent inverse scattering technique," in *Review of Progress in Quantitative Nondestructive Evaluation*, Vol. 4, edited by D. O. Thompson and D. E. Chimenti (Plenum, New York, 1985), pp. 551–558.

[8] J. P. Corones, M. E. Davison, and R. J. Krueger, "The effects of dissipation in one-dimensional inverse problem," in Ref. 5, pp. 107–114.

[9] J. P. Corones, R. J. Krueger, and V. H. Weston, "Some recent results in inverse scattering theory," in *Inverse Problems of Acoustic and Elastic Waves*, edited by F. Santosa, Y. Pao, W. Symes, and C. Holland (SIAM, Philadelphia, PA, 1985), pp. 65–81.

[10] J. P. Corones, M. E. Davison, and R. J. Krueger, "Dissipative inverse problems in the time domain," in *Inverse Methods in Electromagnetic Imaging*, NATO ASI series, Series C, Vol. 143, edited by W-M. Boerner (Reidel, Dordrecht, 1985), pp. 121–130.

[11] A. M. Bruckstein, B. C. Levy, and T. Kailath, "Differential methods in inverse scattering," SIAM J. Appl. Math. **45**, 312 (1985).

[12] V. Weston, "On the inverse problem for a hyperbolic dispersive partial differential equation," J. Math. Phys. **13**, 1952 (1972).

[13] V. Weston and R. J. Krueger, "On the inverse problem for a hyperbolic

dispersive partial differential equation. II," J. Math. Phys. **14**, 406 (1973).

[14] V. Weston, "On inverse scattering," J. Math. Phys. **15**, 209 (1974).

[15] R. J. Krueger, "An inverse problem for a dissipative hyperbolic equation with discontinuous coefficients," Quart. Appl. Math. **34**, 129 (1976).

[16] R. J. Krueger, "An inverse problem for an absorbing medium with multiple discontinuities," Quart. Appl. Math. **36**, 235 (1978).

[17] R. J. Krueger, "Numerical aspects of a dissipative inverse problem," IEEE Trans. Antennas Propag. **AP-29**, 253 (1981).

[18] M. Jaulent, "Inverse scattering problems in absorbing media," J. Math. Phys. **17**, 1351 (1976).

[19] M. Jaulent, "Inverse scattering problems for LCRG transmission lines," J. Math. Phys. **23**, 2286 (1982).

[20] R. S. Beezley and R. J. Krueger, "An electromagnetic inverse problem for dispersive media," J. Math. Phys. **26**, 317 (1985).

[21] E. Ammicht, J. P. Corones, and R. J. Krueger, "Direct and inverse scattering for viscoelastic media," preprint.

[22] B. L. N. Kennet, *Seismic Wave Propagation in Stratified Media* (Cambridge U. P., Cambridge, 1983).

[23] J. P. Corones, "Bremmer series that correct parabolic approximations," J. Math. Anal. Appl. **50**, 361 (1975).

Direct and inverse scattering in the time domain for a dissipative wave equation. II. Simultaneous reconstruction of dissipation and phase velocity profiles

G. Kristensson
Division of Electromagnetic Theory, Royal Institute of Technology, S-100 44 Stockholm, Sweden
R. J. Krueger
Applied Mathematical Sciences, Ames Laboratory—United States Department of Energy, Iowa State University, Ames, Iowa 50011

(Received 8 August 1985; accepted for publication 30 January 1986)

The one-dimensional inverse scattering problem for inhomogeneous lossy media is considered. The model problem involves electromagnetic wave propagation in a medium of unknown thickness with spatially varying conductivity and permittivity. Two inversion algorithms are developed in the time domain using data obtained from normally incident plane waves. These algorithms utilize reflection data from both sides of the medium, and one of them also uses transmission data. These algorithms are implemented numerically on several examples, one of which includes the effects of noisy data. The possibility of using one-sided reflection data and no transmission data is reviewed and analyzed.

I. INTRODUCTION

Inverse scattering problems for lossy media are not well understood, even in the one-dimensional case. Such problems can be addressed on a variety of levels, depending on the underlying model of dissipation and the information sought from the inversion procedure. In this paper a one-dimensional wave propagation model is considered in which the dissipation and phase velocity are spatially varying functions; i.e., functions of depth in the medium. The analysis is carried out in the time domain. Inversion procedures are developed for simultaneously reconstructing the dissipation and phase velocity profiles using data obtained from normally incident plane waves.

In a previous paper[1] (hereafter called Part I) various aspects of the direct scattering problem were developed. The pertinent results from Part I will be summarized in Sec. II below. Hence, the reader who is primarily interested in the inverse problem will find this paper fairly self-contained, with the exception that the first two sections of Part I should be consulted for an overview of the problem at hand and also for an explanation of the notation.

A model problem for the techniques presented here involves one-dimensional electromagnetic wave propagation in a medium characterized by nonconstant permittivity and conductivity profiles. A precise statement of the model problem is given in Part I, Sec. II.

Two inversion algorithms are developed in this paper. In both of them it is assumed that the medium has finite but unknown thickness and that reflection data are available on both sides of the medium. One of the algorithms also requires transmission data. All of these data are in the form of finite time traces of impulse responses. The specific data requirements are given in Sec. III.

The inversion procedure using transmission data and both sets of reflection data is shown in Sec. III. Two numerical examples are also given, one of which shows the performance of the algorithm using noisy data. In Sec. IV the inversion algorithm using only reflection data is given. This is an iteration procedure, and a numerical example of its performance is also provided. In Sec. V the question of inverting reflection data from only one side of the medium is considered. Inversions of this nature have been addressed in previous works[2,3] under the assumption that either the conductivity or permittivity is known *a priori*. In the present paper, it is shown that if only a finite time trace of the (reflected) impulse response is known, and no information regarding the medium is given, then an infinite number of medium profiles can be found that produce such a time trace.

A number of authors[4–8] have developed inversion procedures for dissipative media that require two-sided reflection data as well as transmission data. The inversion procedure given in Sec. III seems to be more intuitive than these other procedures since it clearly shows the interplay between the early time behavior of one reflected signal with the late time behavior of the reflected signal from the other side. This is also evident in the inversion procedure in Sec. IV. An inversion procedure using transmission data and one-sided reflection data has been previously developed,[9–11] although the model problem is different from that considered in this paper.

A brief summary is presented in Sec. VI. Also, an example is provided that demonstrates that under certain conditions it is possible for two different media to produce the same two-sided reflection data for time traces corresponding to one round trip in the medium.

The paper concludes with an Appendix that provides sufficient conditions for the inversion procedure of Sec. IV to be well posed.

II. SUMMARY OF PREVIOUS RESULTS

The equations used in the inverse algorithms presented in this paper are summarized in this section. The reader interested in the details in the derivations is referred to Part I.

The reflection kernels $R^{\pm}(x,y,s)$ for the subregion

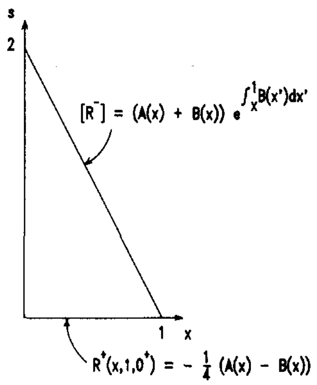
s
2
$[R^-] = (A(x) + B(x))\, e^{\int_x^1 B(x')dx'}$
1
x
$R^+(x,1,0^+) = -\frac{1}{4}(A(x) - B(x))$

FIG. 1. The domain of $R^+(x,1,s)$ for one round trip. The boundary value of R^+ at $s = 0^+$ and the discontinuity of R^- along the line $s = 2(1-x)$ are also shown.

$[x,y]$ satisfy (see also Sec. III, Part I)

$$\begin{aligned} R_x^+(x,y,s) &= 2R_s^+(x,y,s) - B(x)R^+(x,y,s) \\ &\quad - \tfrac{1}{2}[A(x)+B(x)] \\ &\quad \times \int_0^s R^+(x,y,s')R^+(x,y,s-s')ds', \\ &\quad s>0, \end{aligned} \tag{2.1}$$

$$R^+(x,y,0^+) = -\tfrac{1}{4}[A(x)-B(x)], \quad x<y,$$

$$\begin{aligned} R_y^-(x,y,s) &= -2R_s^-(x,y,s) + B(y)R^-(x,y,s) \\ &\quad - \tfrac{1}{2}[A(y)-B(y)] \\ &\quad \times \int_0^s R^-(x,y,s')R^-(x,y,s-s')ds', \quad s>0, \end{aligned} \tag{2.2}$$

$$R^-(x,y,0^+) = \tfrac{1}{4}[A(y)+B(y)], \quad x<y.$$

These kernels are discontinuous across the plane $s = 2(y-x)$. The discontinuities can be related to the internal properties of the slab (see Fig. 1):

$$\begin{aligned} &[R^+(x,y,s)]_{s=2(y-x)^-}^{s=2(y-x)^+} \\ &\quad = \tfrac{1}{4}[A(y)-B(y)] \exp\left\{\int_x^y B(x')dx'\right\}, \\ &[R^-(x,y,s)]_{s=2(y-x)^-}^{s=2(y-x)^+} \\ &\quad = -\tfrac{1}{4}[A(x)+B(x)] \exp\left\{\int_x^y B(x')dx'\right\}. \end{aligned} \tag{2.3}$$

Furthermore, the kernels satisfy

$$\begin{aligned} R^+(x,y,s) &= R^+(x,x+s/2^+,s), \quad s<2(y-x), \\ R^-(x,y,s) &= R^-(y-s/2^-,y,s), \quad s<2(y-x). \end{aligned} \tag{2.4}$$

These last relations express the property that the reflected field is independent of position of the rear interface for times less than one round trip.

In Sec. III in Part I, the effect of reciprocity on the transmission kernels $T^\pm(x,y,s)$ was analyzed. It was shown that the two transmission kernels $T^\pm(x,y,s)$ are proportional to each other as functions of s, as are the propagator kernels $W^\pm(x,y,s)$. Thus it suffices to consider just one transmission kernel T and one propagator kernel W. The relations are

$$\begin{aligned} T(x,y,s) &= T^+(x,y,s)/t^+(x,y) \\ &= T^-(x,y,s)/t^-(x,y), \end{aligned} \tag{2.5}$$

$$\begin{aligned} W(x,y,s) &= W^+(x,y,s)t^+(x,y) \\ &= W^-(x,y,s)t^-(x,y), \end{aligned} \tag{2.6}$$

where

$$t^\pm(x,y) = \exp\left\{\mp\frac{1}{2}\int_x^y [A(x') \mp B(x')]dx'\right\}. \tag{2.7}$$

The resolvent equation which relates T and W to each other is

$$\begin{aligned} &T(x,y,s) + W(x,y,s) \\ &\quad + \int_0^s T(x,y,s-s')\, W(x,y,s')ds' = 0. \end{aligned} \tag{2.8}$$

The propagator kernel W satisfies the imbedding equations

$$\begin{aligned} W_x(x,y,s) &= \frac{1}{2}[A(x)+B(x)] \Big\{R^+(x,y,s) \\ &\quad + \int_0^s W(x,y,s')R^+(x,y,s-s')ds'\Big\}, \quad s>0, \end{aligned} \tag{2.9}$$

$$\begin{aligned} W_y(x,y,s) &= \frac{1}{2}[A(y)-B(y)] \Big\{R^-(x,y,s) \\ &\quad + \int_0^s W(x,y,s')R^-(x,y,s-s')ds'\Big\}, \quad s>0. \end{aligned} \tag{2.10}$$

In Sec. V in Part I the extension of data from one round trip, $0<s<2(y-x)$, to arbitrary time s is derived. Transmission data and reflection data for $0<s<2(y-x)$ are extended to $s>2(y-x)$ by the following equations:

$$\begin{aligned} &T(x,y,s) + \int_{2(y-x)}^s W(x,y,s-s')T(x,y,s')ds' \\ &\quad = G(x,y,s) = \begin{cases} -\displaystyle\int_{s-2(y-x)}^{2(y-x)} W(x,y,s-s')T(x,y,s')ds', & 2(y-x)<s<4(y-x), \\ 0, & s>4(y-x), \end{cases} \end{aligned} \tag{2.11}$$

$$R^\pm(x,y,s) = \int_0^{2(y-x)} T(x,y,s-s')\left[R^\pm(x,y,s') + \int_0^{s'} W(x,y,s'-s'')R^\pm(x,y,s'')ds''\right]ds', \quad s>2(y-x). \tag{2.12}$$

III. THE INVERSION ALGORITHM WITH COMPLETE DATA

The new algorithm presented in this section utilizes a complete set of data, namely the two (physical) reflection kernels $R^{\pm}(0,1,s)$ and the (physical) transmission kernel $T(0,1,s)$ for a complete round trip in the slab, $0<s<2$. These data are complete in the sense that they can be extended to arbitrary time s by the extension procedure described in Sec. V in Part I. Loosely speaking, the algorithm combines an early time behavior in $R^{\pm}$ with a late time behavior in $R^{\mp}$ and the properties of the discontinuity in $R^{\mp}$. This statement and its more precise meaning will become much clearer in this section.

All the data described above and the constant $G(1)$ defined below are needed to recover the two unknown functions $A(x)$ and $B(x)$, $0<x<1$. From these two functions it is then easy to find the unknown permittivity and conductivity as a function of z as well as the total length L of the slab. However, two more constants are needed to transform from A and B to ϵ and σ. Thus the complete set of data to simultaneously recover both the permittivity and the conductivity are

$$\begin{aligned}&R^{+}(0,1,s), \quad 0<s<2,\\ &R^{-}(0,1,s), \quad 0<s<2,\\ &T(0,1,s), \quad 0<s<2,\\ &G(1),\\ &l,\\ &\epsilon(0) \quad \text{or} \quad \epsilon(L),\end{aligned} \tag{3.1}$$

where

$$G(x)=1/[t^{+}(0,x)t^{-}(0,x)] = \exp\left\{-\int_0^x B(x')\right\} dx'$$

[see Eq. (2.7) for a definition of $t^{\pm}(x,y)$], and $G(1)$ is a constant associated with the attenuation of the field within the slab. From the definition of the transmission operators [Eqs. (2.18) and (2.20) in Part I], $G(1)$ is a measurable quantity. The constant l [see Eq. (2.6) in Part I] is a constant related to the total time of measurement. The permittivity $\epsilon(0)$ at the left interface [or $\epsilon(L)$ at the right] is also assumed to be known from experimental data.

The inversion algorithm works from one side of the medium to the other. For convenience the algorithm is presented for a propagation from the left-hand side of the slab towards the right and all the details of the alogorithm will be shown for this particular choice. Thus, the subregions to be considered are of the form $[x,1]$ with y being fixed at 1. The necessary modifications to propagate from the right-hand side are rather straightforward.

In Eq. (2.3), the jump across the plane $s=2(y-x)$ was given as a function of the internal properties of the slab. This jump can, however, be expressed in an alternative way by the extension of data presented in Sec. V in Part I. Suppose the reflection data are known for $s<2(y-x)$. In terms of these data the value just above the plane $s=2(y-x)$ can be calculated from Eq. (2.12). The resulting jump is

$$\begin{aligned}&[R^{\pm}(x,y,s)]_{s=2(y-x)^-}^{s=2(y-x)^+}\\ &\quad=\int_0^{2(y-x)} T(x,y,2(y-x)-s')\left[R^{\pm}(x,y,s')\right.\\ &\qquad\left.+\int_0^{s'} W(x,y,s'-s'')R^{\pm}(x,y,s'')ds''\right]ds'\\ &\qquad-R^{\pm}(x,y,2(y-x)^{-})\\ &\quad=-\int_0^{2(y-x)} W(x,y,2(y-x)-s')R^{\pm}(x,y,s')ds'\\ &\qquad-R^{\pm}(x,y,2(y-x)^{-}).\end{aligned} \tag{3.2}$$

The resolvent equation, Eq. (2.8), has been used to simplify the expression above. In particular, the jump in the reflection kernel R^{-} for $y=1$ can, with use of Eq. (2.4), be expressed as

$$\begin{aligned}&[R^{-}(x,1,s)]_{s=2(1-x)^-}^{s=2(1-x)^+}\\ &\quad=-\int_0^{2(1-x)} W(x,1,2(1-x)-s')R^{-}(0,1,s')ds'\\ &\qquad-R^{-}(0,1,2(1-x)^{-}).\end{aligned} \tag{3.3}$$

It should be noted that only the physical kernel $R^{-}(0,1,s)$, $0<s<2$, is used in Eq. (3.3).

From the equations above it is now clear that knowing $R^{\pm}(x,1,s)$ and $W(x,1,s)$ for a fixed x gives two linearly independent relations between the two unknown functions $A(x)$ and $B(x)$ at the point x. This can be seen by combining Eqs. (2.3) and (3.3) together with the early time behavior of $R^{+}(x,1,s)$ in Eq. (2.1),

$$\begin{aligned}&\int_0^{2(1-x)} W(x,1,2(1-x)-s')R^{-}(0,1,s')ds'\\ &\qquad+R^{-}(0,1,2(1-x)^{-})\\ &\quad=\frac{1}{4}[A(x)+B(x)]\exp\left\{\int_x^1 B(x')dx'\right\},\\ &R^{+}(x,1,0^{+})=-\tfrac{1}{4}[A(x)-B(x)].\end{aligned} \tag{3.4}$$

Before describing the general inversion algorithm, the initialization of the procedure is addressed. From the data in Eq. (3.1) the resolvent $W(0,1,s)$, $0<s<2$, is obtained by solving Eq. (2.8) at $x=0$. Equations (3.4) are then easily solved for $A(x)$ and $B(x)$ at $x=0^{+}$ and the initialization of $A(x)$ and $B(x)$ is completed.

The inversion scheme can now be written down in a general setting. As in earlier works,[2,12] which used only the R^{+} equation [Eq. (2.1)], a grid of points is established in (x,s) space. The mesh is uniform in each direction, with $\Delta s=2\Delta x$, which takes advantage of the directional derivative nature of Eq. (2.1). Now Eqs. (2.1) and (2.9) are discretized on this grid. The calculation proceeds from left to right across the grid, starting at $x=0$ and marching to $x=1$, with $0<s<2(1-x)$. In its most basic form, the inversion algorithm for determining $A(x)$ and $B(x)$ is as follows.

(1) Equation (2.9) is used to explicitly step $W(x,1,s)$ forward in the x direction to the next set of x grid points.

(2) Equation (2.1) is used to implicitly step a portion of $R^{+}(x,1,s)$ forward in the x direction to the next x grid point at $s=0$.

(3) Equations (3.4) are used at these new x grid points to obtain $A(x)$ and $B(x)$.

(4) Equation (2.1) is used to implicitly step the remaining $R^{+}(x,1,s)$ data forward in the x direction to the next set of x grid points.

(5) Now repeat steps (1) through (4) to move one step deeper into the slab.

This procedure can be modified in a number of ways to improve its numerical accuracy. Details regarding the numerical implementation are not discussed here.

There are some interesting points to notice in the inversion algorithm outlined above. First, the transmission data $T(0,1,s)$ are used only in the initialization step, and thereafter it is the resolvent of T that is used to step into the medium. Second, since the calculation is being carried out in the plane $y=1$ (see Part I, Fig. 4), the $R^{-}(0,1,s)$ data are constant on lines of constant s. Therefore, it is not necessary to propagate R^{-} into the medium via an integrodifferential equation; rather, it is the physical data $R^{-}(0,1,s)$ that appears in Eq. (3.4).

The final step in the inversion scheme is to calculate the depth $z(x)$, the total length L, the permittivity $\epsilon(z)$, and the conductivity $\sigma(z)$ from the profile functions $A(x)$ and $B(x)$, $0<x<1$. From the definitions of $A(x)$ and $B(x)$, given by Eqs. (2.15) and (2.16) in Part I, it is easy to obtain the following relations:

$$z(x)=\frac{l}{\sqrt{\mu_0\epsilon(0)}}\int_0^x \exp\left[-\textstyle\int_0^{x'} A(x'')dx''\right]dx',$$

$$0<x<1, \tag{3.5}$$

$$\epsilon(z(x))=\epsilon(0)\exp\left[2\int_0^x A(x')dx'\right], \tag{3.6}$$

$$\sigma(z(x))=\frac{-\epsilon(0)B(x)\exp\left[2\int_0^x A(x')dx'\right]}{l}. \tag{3.7}$$

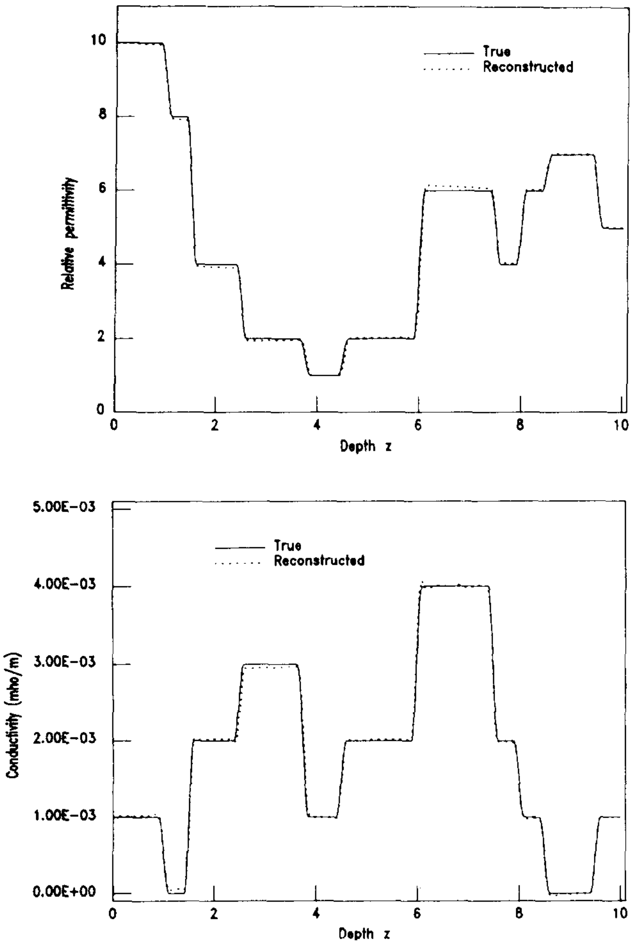

FIG. 2. The relative permittivity and conductivity profiles in example 1. The depth is given in m.

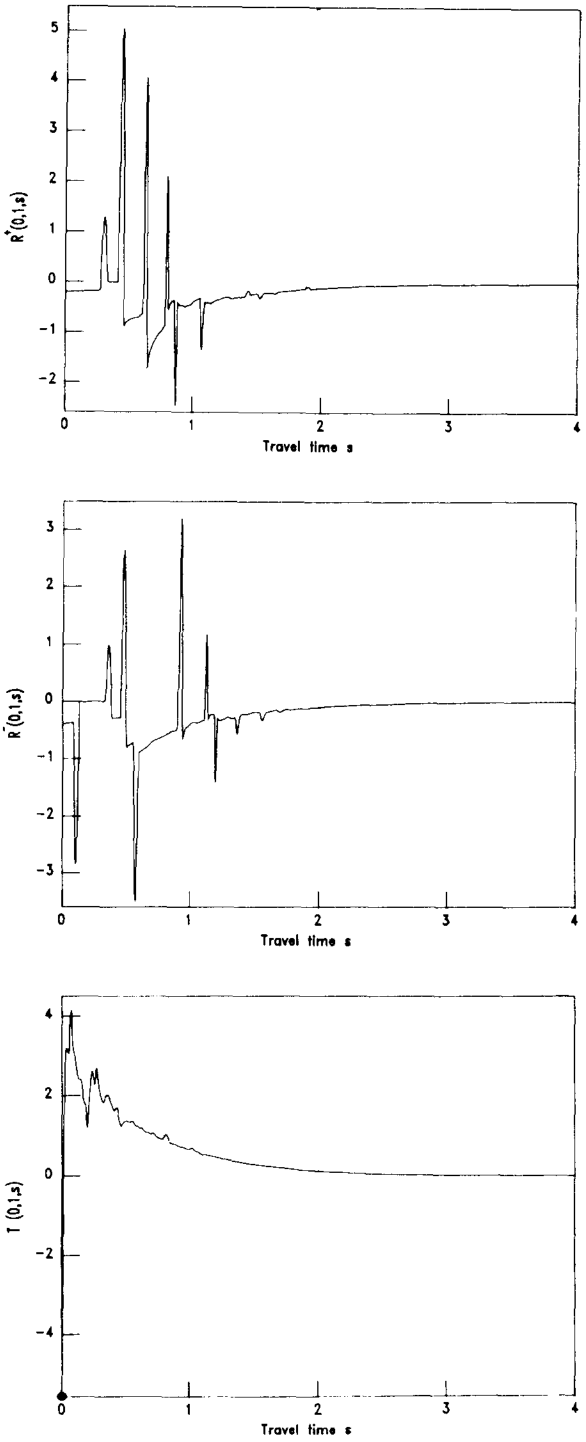

FIG. 3. The physical scattering kernels $R^{\pm}(0,1,s)$ and $T(0,1,s)$ for example 1. Two round trips are shown. The value of $T(0,1,0^{+})$ is marked with a solid dot.

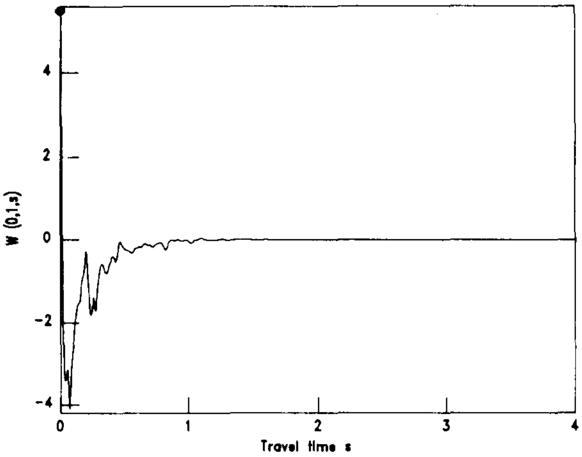

FIG. 4. The resolvent kernel $W(0,1,s)$ for example 1. Two round trips are shown. The value of $W(0,1,0^+)$ is marked with a solid dot.

In particular, the total length L of the slab is

$$L = \frac{l}{\sqrt{\mu_0 \epsilon(0)}} \int_0^1 \exp[-\textstyle\int_0^{x'} A(x'')dx'']dx'. \qquad (3.8)$$

The results of some inversions are now shown. In all of the examples in this paper, synthetic $R^{\pm}$ data were generated using Eqs. (2.1) and (2.2), and T data were generated via Eq. (3.15) in Part I. The procedure for doing this was to first choose an $(\epsilon(z),\sigma(z))$ profile, convert to an $(A(x),B(x))$ profile, generate $R^{\pm}$ and T for two different step sizes (Δx), and then extrapolate those results to obtain the data for the inverse problem. The accuracy of all numerical algorithms was verified using the exact solutions displayed in Part I, Appendix B. All calculations were performed in single precision on a VAX 11/750.

Example 1: The (ϵ,σ) profiles in this example are approximately piecewise constant, as shown in Fig. 2. (Recall that the derivations required that ϵ be smooth.) The length of the medium is 10 m and permittivity relative to that of free space is shown. Scattering data for this medium are displayed in Fig. 3 for two round trips in the medium, although it is only the data for $0<s<2$ that are used in the inversion algorithm. It is difficult to see the discontinuities in $R^{\pm}(0,1,s)$ at $s=2$. Figure 4 shows the resolvent kernel $W(0,1,s)$, which is obtained from Eq. (2.8). Notice the compact support, with W vanishing for $s>2$. The $R^{\pm}$ data are shown differently in Fig. 5, with the R^+ time scale running along the bottom axis and the R^- along the top. The spikes in the two time traces line up at corresponding regions of high reflectivity in the medium. Those traces decay toward zero quite rapidly due to the absorption of energy in the medium and the reflection of energy out of the medium.

The reconstructed profiles are shown in Fig. 2. These reconstructions used 513 data points from each of the time traces for $R^{\pm}$ and T. There is essentially no difference if 257 points are used instead.

Example 2: The performance of the inversion algorithm with noisy data is now examined. The medium profiles are shown by the solid lines in Fig. 6. The exact scattering data for these profiles are shown by the solid lines in Fig. 7. Gaussian white noise was then added to the kernels, resulting in

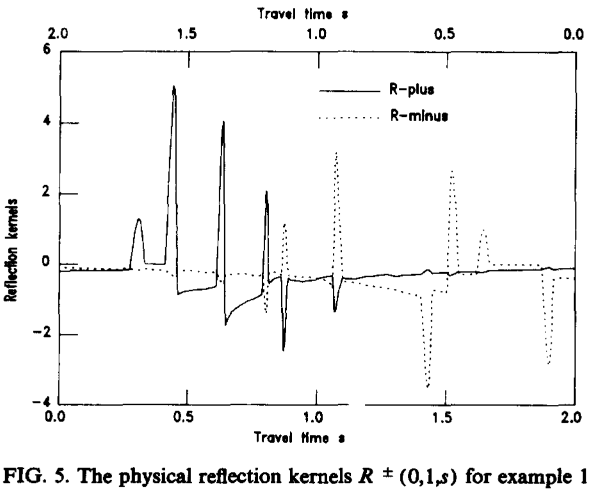

FIG. 5. The physical reflection kernels $R^{\pm}(0,1,s)$ for example 1 for one round trip. The solid line is the time trace for R^+ and should be read from left to right using the lower scale in the figure. The dotted line shows R^- and should be read from right to left using the upper scale.

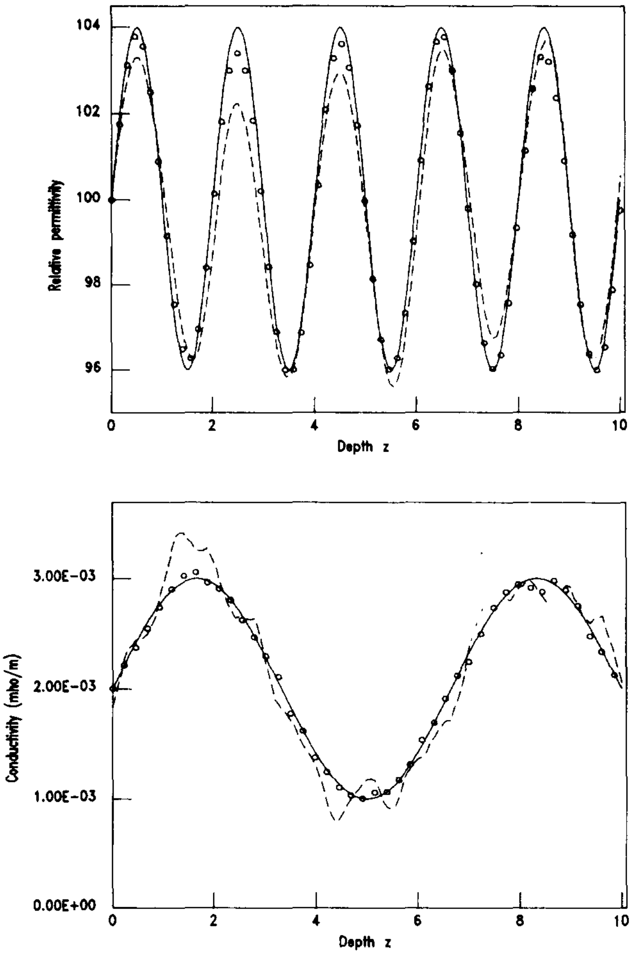

FIG. 6. The relative permittivity and conductivity profiles in example 2. The solid lines are the true profiles and the broken lines and the circles are reconstructions using noisy data. Each reconstruction uses 129 data points, but for graphical clarity not all circles are displayed. For an explanation of the noise, see the text. The depth is given in *m*.

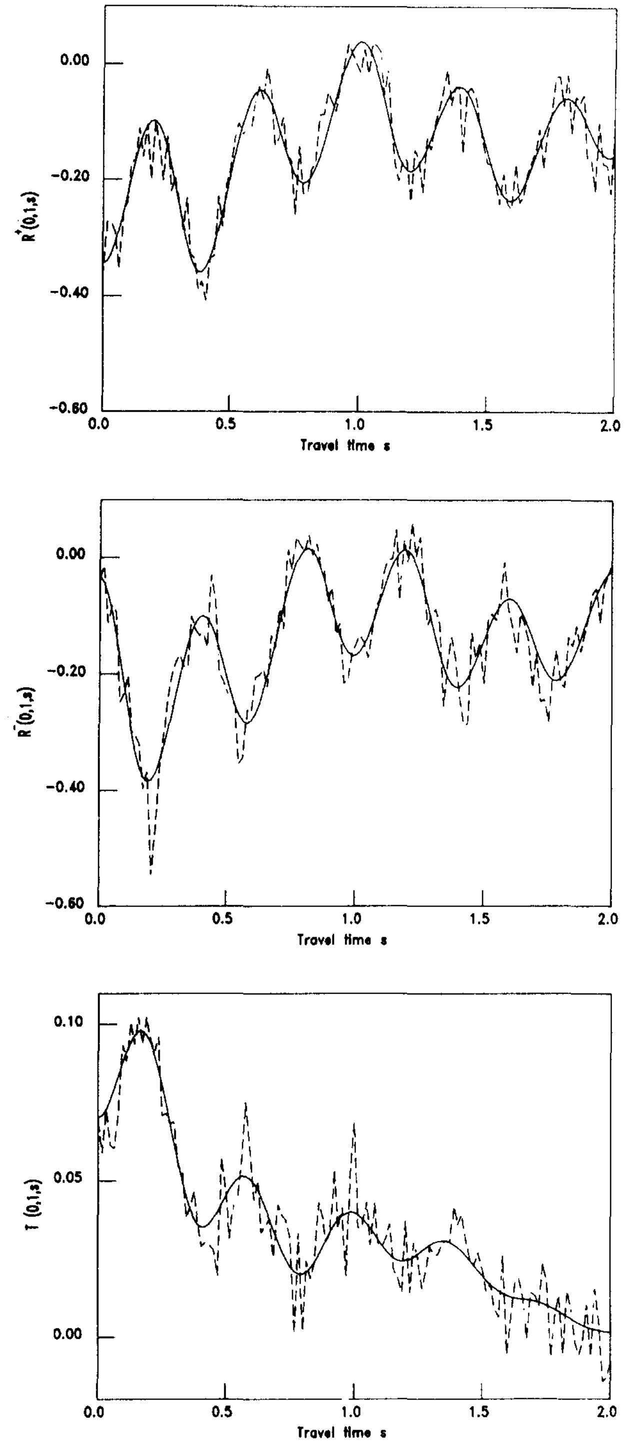

FIG. 7. The physical scattering kernels $R^{\pm}(0,1,s)$ and $T(0,1,s)$ for example 2. The solid lines are the time traces without noise and the broken lines show the noisy data with rms $S/N \doteq 1.8$. For further details, see the text.

the corrupted data shown by the broken lines in Fig. 7. The signal to noise ratio for these data is approximately 1.8. The noisy data were smoothed using two applications of a five point linear least squares smoother. (The first application left too much high-frequency noise in the data.) The broken line in Fig. 6 shows the resulting reconstructions using 129 data points. The reconstruction was carried out a second time with noisy data having signal to noise ratio of approximately 6.8. The results are much improved, and are shown with circles in Fig. 6. In the absence of noise, the reconstructed profiles are indistinguishable from the original profiles.

The definition of the root mean square signal to noise ratio (rms S/N) that was used above is

$$\text{rms}\, S/N = \left[\int_0^2 [K(0,1,s) - \overline{K}\,]^2\, ds\right]^{1/2} \Big/ (2\sigma).$$

Here, σ denotes the standard deviation of the noise, $K(0,1,s)$ denotes a noisy scattering kernel, and

$$\overline{K} = \frac{1}{2}\int_0^2 K(0,1,s)\,ds.$$

IV. INVERSION USING REFLECTION DATA FROM BOTH INTERFACES

In the previous section an inversion algorithm was presented that utilized both of the reflection kernels and the transmission kernel for one complete round trip in the slab. These data are sufficient to recover both A and B (i.e., ϵ and σ) for the medium. In this section an inversion algorithm is given that uses only the reflection data $R^{\pm}(0,1,s)$ for a complete round trip. More explicitly, the data are a subset of (3.1), namely, $R^{\pm}(0,1,s)$, $0<s<2$, and constants l and $\epsilon(0)$ [or $\epsilon(L)$].

The algorithm is an iteration procedure. It has the property that the iterates may not converge, and if they do converge, the result may not be the correct solution. However, sufficient conditions for convergence to the correct solution are supplied in the Appendix.

The basis for the inversion algorithm is Eqs. (2.1) and (2.2). Begin by setting $y = 1$ in Eq. (2.1) and using the directional derivative nature of that equation to rewrite Eq. (2.1) in integrated form as

$$\begin{aligned} R^{+}(x,1,s) = R^{+}(0,1,s+2x) - \int_0^x \{&B(x')R^{+}(x',1,s+2(x-x')) \\ &+ \tfrac{1}{2}[A(x') + B(x')](R^{+} * R^{+})(x',1,s+2(x-x'))\}dx', \end{aligned} \tag{4.1}$$

where the * operation denotes convolution in s,

$$(f * g)(x,y,s) = \int_0^s f(x,y,s')g(x,y,s-s')ds'. \tag{4.2}$$

Similarly, in integrated form, Eq. (2.2) becomes (with $x = 0$)

$$R^-(0,y,s) = R^-(0,1,s+2(1-y)) - \int_y^1 \{B(y')R^-(0,y',s+2(y'-y)) - \tfrac{1}{2}[A(y') - B(y')](R^- * R^-)(0,y',s+2(y'-y))\}dy'. \tag{4.3}$$

Notice that the first term on the right-hand side of Eqs. (4.1) and (4.3) is the given reflection data. Denote this by

$$F^{\pm}(s) = R^{\pm}(0,1,s). \tag{4.4}$$

Now Eqs. (4.1) and (4.3) form the basis for an iteration procedure given by

$$R^+_{n+1}(x,1,s) = F^+(s+2x) - \int_0^x \{B_n(x')\,R^+_n(x',1,s+2(x-x')) + \tfrac{1}{2}[A_n(x') + B_n(x')](R^+_n * R^+_n)(x',1,s+2(x-x'))\}dx', \tag{4.5}$$

with $0<x<1$, $0<s<2(1-x)$, $n = 1, 2, 3,\ldots$, and

$$R^-_{n+1}(0,y,s) = F^-(s+2(1-y)) - \int_y^1 \{B_n(y')\,R^-_n(0,y',s+2(y'-y)) - \tfrac{1}{2}[A_n(y') - B_n(y')](R^-_n * R^-_n)(0,y',s+2(y'-y))\}dy', \tag{4.6}$$

with $0<y<1$, $0<s<2y$, and $n = 1, 2, 3,\ldots$. The functions A_n and B_n are defined as

$$A_n(x) = 2[R^-_n(0,x,0^+) - R^+_n(x,1,0^+)],$$
$$B_n(x) = 2[R^-_n(0,x,0^+) + R^+_n(x,1,0^+)], \tag{4.7}$$

which is suggested by the initial conditions given in Eqs. (2.1) and (2.2). One method for starting the iteration is to choose

$$R^+_1(x,1,s) = F^+(s+2x),$$
$$R^-_1(0,y,s) = F^-(s+2(1-y)). \tag{4.8}$$

Now if the iterates converge,

$$R^{\pm}_n \to R^{\pm}, \tag{4.9}$$

then it is natural to define $A(x)$ and $B(x)$ by Eq. (4.7), with subscript n removed from all quantities. Also, notice that if the iterates converge, then the limit functions given in Eq. (4.9) agree with the given reflection data when x is set equal to 0 and y is set equal to 1.

It is interesting to note that the initialization procedure given in Eq. (4.8) is a generalization of the nondissipative Bremmer approximation given in Ref. 12. It corresponds to ignoring dissipative effects on the reflected fields as well as ignoring multiple scattering effects. Hence, in a weakly dissipative, weakly scattering medium, Eqs. (4.7) and (4.8) themselves yield a good approximation to A and B given by

$$A(x) \doteq A_1(x) = 2[R^-_1(0,1,2(1-x)) - R^+_1(0,1,2x)],$$
$$B(x) \doteq B_1(x) = 2[R^-_1(0,1,2(1-x)) + R^+_1(0,1,2x)]. \tag{4.10}$$

Continuing the iteration can be thought of as bringing higher-order effects into the calculation.

Sufficient conditions exist to guarantee that the scheme

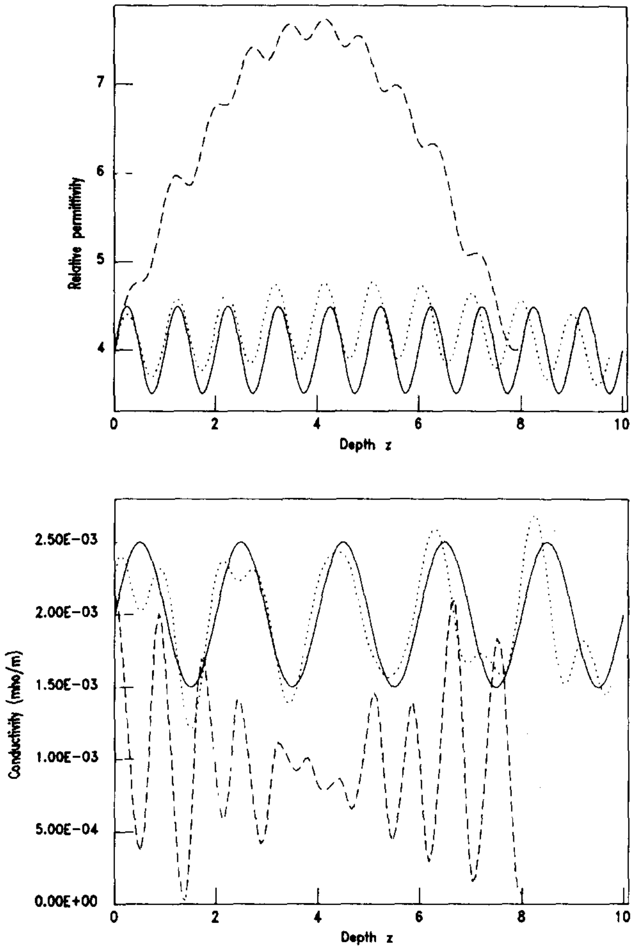

FIG. 8. The relative permittivity and conductivity profiles in example 3. The solid lines are the true profiles, and the broken lines are the initial approximations given by Eq. (4.10). The dotted lines show the profiles after 20 iterations. After 60 iterations the profiles coincide with the solid lines. The depth is given in *m*.

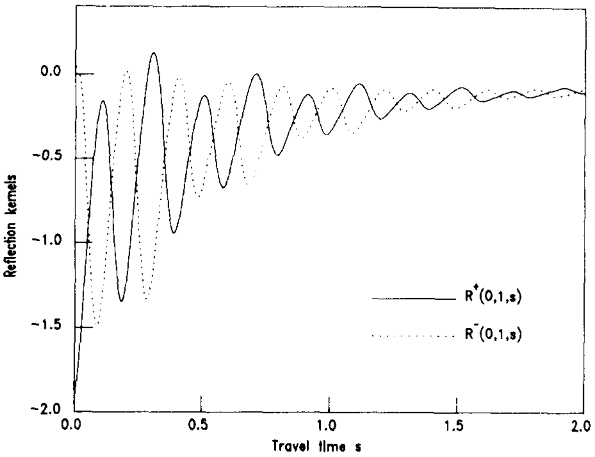

FIG. 9. The physical reflection kernels $R^{\pm}(0,1,s)$ for example 3.

does converge to one and only one solution. These are conditions on the physical reflection data and therefore have practical implication. The conditions are

$$|F^{\pm}(s)| < f, \quad 0 < s < 2, \tag{4.11}$$

where $f = (11/\sqrt{22} - 50)/27 \pm 0.059\,06$. In the Appendix, it is shown that if the condition (4.11) is satisfied, then the solution of the inverse problem, (A,B), exists, is unique, and depends continuously on the data $F^{\pm}$. Thus, in this case, reflection data alone suffice to reconstruct A and B.

These positive results do not imply that any two functions $F^{\pm}$ satisfying (4.11) correspond to scattering data for some physical medium. This is because the reconstructed $B(x)$ may be greater than 0, a result that is nonphysical for the model problem, Eq. (2.1) in Part I. Also, condition (4.11) is not a necessary condition for convergence, as will be apparent from the following example.

Example 3: The ϵ and σ profiles are shown in solid lines in Fig. 8, and the corresponding $R^{\pm}$ data are shown in Fig. 9. Notice that these data do not satisfy condition (4.11). Nevertheless, the iterates converge to the original profiles in Fig. 8. The broken lines in Fig. 8 display the estimates of ϵ and σ given by the initialization procedure in Eqs. (4.8), or equivalently in Eqs. (4.10). After 20 iterations the estimates of ϵ and σ are given by the dotted lines in Fig. 8. After 60 iterations the estimates coincide with the original profiles. Continuing the iteration procedure produces no change in the estimated profiles.

V. INVERSION USING REFLECTION DATA FROM ONE INTERFACE

It was shown in the previous section that under certain circumstances, reflection data from both interfaces can be used to uniquely reconstruct A and B. In this section some aspects of reconstructing A and B are considered for the case in which the data consist only of $R^{+}(0,1,s)$ for $0 < s < 2$. This is an important problem since it corresponds to the case in which all data measurement is carried out on one side of the slab and consequently, a semi-infinite medium can be considered. In such a case, the parameter l defined in Part I, Eq. (2.6), is given by

$$l = t_{\max}/2, \tag{5.1}$$

where data is collected for physical time t in the interval $0 < t < t_{\max}$. Thus, only a finite portion of the medium can be probed, namely, that portion for $0 < z < L$, where L is given in Part I, Eq. (2.6), with l as in Eq. (5.1) above.

In this case it seems intuitively clear that nonunique solutions (A,B) should exist, provided the data correspond to a physical reflection kernel. The intuition here is that a single function of s (for $0 < s < 2$) cannot be used to reconstruct two independent functions $A(x)$, $B(x)$ for $0 < x < 1$.

If it is known *a priori* that the medium is nondissipative so that $B = 0$, then Eq. (2.1) can be used in an inversion algorithm to recover $A(x)$. This has been shown in Refs. 2 and 12. More generally, if the conductivity $\sigma(z)$ is known, then Eq. (2.1) can be used to recover the permittivity $\epsilon(z)$ or vice versa. Such problems have been considered in Refs. 2, 3, and 13. Integral equation methods for solving problems of this latter variety have been considered by Bolomey *et al.*[14] and Tijhuis.[15] It has also been shown by Corones *et al.*[2,3,13] that if the *a priori* information about conductivity (or permittivity) is incorrect, then the resulting reconstruction can degrade somewhat dramatically.

The question now addressed is, "What profiles pairs (A,B) [or (ϵ,σ)] produce the same one-sided reflection data, $R^{+}(0,1,s)$, for $0 < s < 2$?" A partial answer to this question will be given by considering media with "small" profile functions A and B. In this case the explicit dependence of the reflection data on A,B can be given asymptotically.

To carry this out, set $y = 1$ in Eq. (2.1) and again use the directional derivative nature of that equation to obtain for $0 < s < 2(1-x)$,

$$\begin{aligned} R^{+}(x,1,s) &= \tfrac{1}{4}\,[B(x+s/2) - A(x+s/2)] \\ &\quad + \int_{x}^{x+s/2} \{B(x')R^{+}(x',1,s+2(x-x')) \\ &\quad + \tfrac{1}{2}\,[A(x') + B(x')](R^{+} * R^{+}) \\ &\quad (x',1,s+2(x-x'))\}dx'. \end{aligned} \tag{5.2}$$

This integrated form of Eq. (2.1) is well suited to the study of the direct problem, while Eq. (4.1) is better suited to the inverse problem. Now define a sequence of iterates given by

$$\begin{aligned} R_1(x,1,s) &= \tfrac{1}{4}\,[B(x+s/2) - A(x+s/2)], \\ R_{n+1}(x,1,s) &= \tfrac{1}{4}\,(B(x+s/2) - A\,(x+s/2)) \\ &\quad + \int_{x}^{x+s/2} \{B(x')R_n(x',1,s+2(x-x')) \\ &\quad + \tfrac{1}{2}\,[A(x') + B(x')](R_n * R_n) \\ &\quad (x',1,s+2(x-x'))\}dx', \end{aligned} \tag{5.3}$$

where $n = 1,2,3,\ldots$. Define λ by

$$\lambda = \sup_{0<x<1} \{|A(x)|,|B(x)|\}.$$

For small λ it follows that the reflection data, $R^{+}(0,1,s)$, are asymptotic to $R_2(0,1,s)$, with

$$R^+(0,1,s) \sim R_2(0,1,s)$$
$$= \frac{1}{4}\left(B\left(\frac{s}{2}\right) - A\left(\frac{s}{2}\right)\right)\left(1 + \int_0^{s/2} B(x')dx'\right)$$
$$+ O(\lambda^3), \lambda \to 0. \qquad (5.4)$$

Thus, two profile pairs, (A_0, B_0) and (A_1, B_1), produce the same reflection data (asymptotically) for $0<s<2$ if

$$(B_0(x) - A_0(x))\left(1 + \int_0^x B_0(x')dx'\right)$$
$$= (B_1(x) - A_1(x))\left(1 + \int_0^x B_1(x')dx'\right). \qquad (5.5)$$

Notice that it follows from Eq. (5.5) that $A_0 = A_1$ if and only if $B_0 = B_1$.

It is interesting to consider Eq. (5.5) for the special case involving a homogeneous, dissipative medium. Thus, assume that both $\epsilon(z)$ and $\sigma(z)$ are constants denoted by ϵ and σ, respectively. Then $A_0(x) \equiv 0$ and $B_0(x) = -\beta = -l\sigma/\epsilon$, where β is small. An equivalent, nondissipative scatterer is then obtained from Eq. (5.5) by setting $B_1(x) \equiv 0$ and solving for $A_1(x)$ [with corresponding permittivity $\epsilon_1(z)$]. This yields

$$A_1(x) = \beta(1-\beta x), \quad 0<x<1, \qquad (5.6)$$

and so, from Eqs. (3.5) and (3.6),

$$z(x) = \frac{l}{\sqrt{\epsilon\mu_0}} \int_0^x \exp\left[-\beta x'\left(1 - \frac{\beta x'}{2}\right)\right]dx', \qquad (5.7)$$

$$\epsilon_1(z(x)) = \epsilon \exp\left[2\beta x(1-\beta x/2)\right]. \qquad (5.8)$$

Notice from Eq. (5.8) that $\epsilon_1(z)$ is an increasing function of z, while Eq. (5.7) shows that the depth L_1 of this equivalent medium has decreased from the original depth L_0 to

$$L_1 = L_0 \int_0^1 \exp\left[-\beta x'\left(1 - \frac{\beta x'}{2}\right)\right] dx'. \qquad (5.9)$$

These conclusions are in agreement with the numerical results given in example 1 of Ref. 2, which suggest that equivalent scatterers that are obtained by decreasing σ result in an increasing permittivity profile and a more shallow medium.

VI. SUMMARY AND CONCLUSIONS

In Sec. III a new time domain inversion procedure for lossy media is developed. The algorithm uses the set of data given by (3.1). With the concept of "extension of data" developed in Part I, this set of data can be used to derive the entire time trace of the scattering kernels. However, data from only one round trip are explicitly used in the algorithm. The possibility of using longer time traces is not addressed in this paper.

At first sight it may seem a little surprising that three functions of time ($R^{\pm}$ and T) have to be given in order to obtain the two unknown functions $A(x)$ and $B(x)$ [or $\epsilon(z)$ and $\sigma(z)$]. It is, however, interesting to observe that other authors[4–8] use similar data sets to invert lossy profiles. The next example shows the importance of transmission data for reliable reconstructions when data from only one round trip are used.

Example 4: In this example it is shown that two different media can produce virtually identical $R^{\pm}$ reflection data for times less than one round trip through the slab, while at the same time producing different transmission data. The two different profiles are shown in Fig. 10 and the corresponding scattering data are given in Fig. 11. The dotted line profile was found by the iteration scheme presented in Sec. IV. It is seen that reflection data are virtually identical up to one round trip. At later times the reflection data are different as well as their discontinuities at one round trip. The transmission data, however, are different for all times. The two profiles are thus equivalent in that they are indistinguishable by just using reflection data for times less than one round trip. Consequently, transmission data are necessary (in general) for reliable reconstructions.

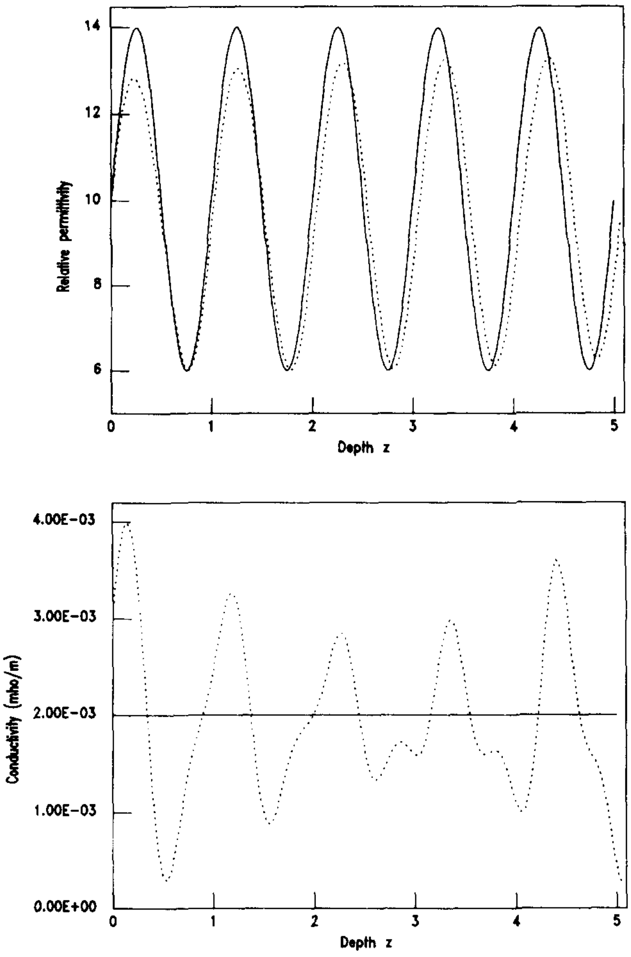

FIG. 10. The relative permittivity and conductivity in example 4. The depth is given in m.

An iterative inversion scheme using only reflection data for one round trip is presented in Sec. IV. The limitations of this inversion algorithm are illustrated by the example in this section. However, sufficient conditions for convergence of the iteration scheme are derived in the Appendix. Notice that this scheme is much more computer intensive than that of Sec. III, with one step of the iteration taking as long as the entire inversion procedure when transmission data are also available.

The effect of using reflection data from one side only is

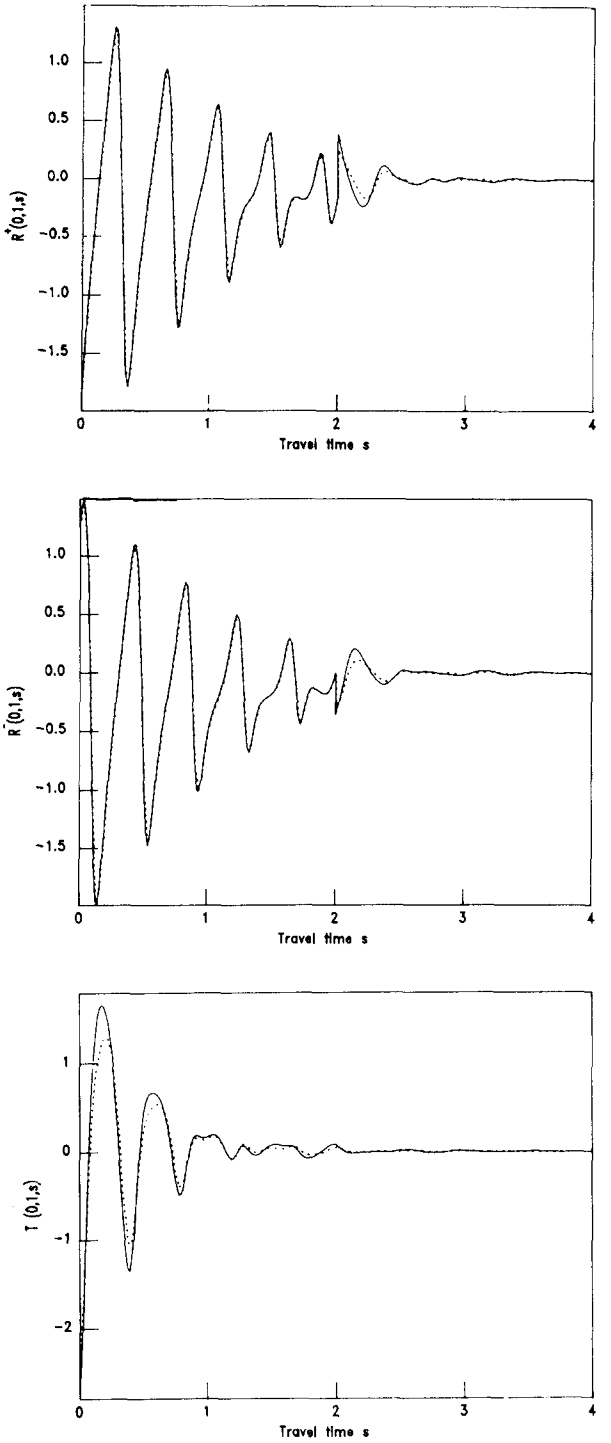

FIG. 11. The physical scattering kernels $R^{\pm}(0,1,s)$ and $T(0,1,s)$ for example 4. Two round trips are shown. The solid (dotted) lines correspond to the solid (dotted) line profiles in Fig. 10.

discussed in Sec. V. It is shown that for weakly scattering media (in which only the lowest-order multiple scattering effects are important), an entire family of media can be generated that produce the same one-sided reflection data for one round trip in travel time. In particular, this implies that for a semi-infinite medium, it is impossible to determine both $\epsilon(z)$ and $\sigma(z)$ from reflection data using normal incidence.

ACKNOWLEDGMENTS

This work was performed at Ames Laboratory, U. S. D. O. E., Iowa State University, and was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U. S. Department of Energy, under Contract No. W-7405-ENG-82, and by the Office of Naval Research Contract No. N0014-83-K-0038. One of the authors (G. K.) would like to thank Ames Laboratory for their hospitality during his visit in the United States. He is also grateful to the National Swedish Board for Technical Development (STUF) for partial support.

APPENDIX: CONVERGENCE OF THE ITERATION PROCEDURE

This appendix shows an analysis of the iteration procedure given in Sec. IV. In particular, it is shown that the condition in Eq. (4.11) guarantees the convergence of the iterates, the uniqueness of the solution and the continuous dependence of the solution on the data $F^{\pm}$.

To begin, suppose the reflection data are bounded by a constant f over one round trip in the slab, i.e.,

$$|F^{\pm}(s)| \leqslant f, \quad 0 < s < 2. \tag{A1}$$

Does it follow that all the iterates $R_n^{\pm}$, given in Eqs. (4.5) and (4.6), are uniformly bounded? Assume there is a constant b such that, for all n,

$$\begin{aligned} &|R_n^{+}(x,1,s)| \leqslant b, \quad 0 \leqslant x \leqslant 1, \quad 0 < s < 2(1-x), \\ &|R_n^{-}(0,y,s)| \leqslant b, \quad 0 \leqslant y \leqslant 1, \quad 0 < s < 2y. \end{aligned} \tag{A2}$$

In this case, it follows that

$$\begin{aligned} &|A_n(x) + B_n(x)| = 4|R_n^{-}(0,x,0^{+})| \leqslant 4b, \\ &|B_n(x)| \leqslant 4b, \end{aligned} \tag{A3}$$

from Eqs. (4.7). Consequently, from Eq. (4.5) it is seen that

$$\begin{aligned} |R_{n+1}^{+}(x,1,s)| &\leqslant f + 4b^2x + 2b^3x(s+x) \\ &\leqslant f + 4b^2 + 2b^3, \end{aligned} \tag{A4}$$

for $0 \leqslant x \leqslant 1$, $0 < s < 2(1-x)$. Similarly, it can be shown that

$$|R_{n+1}^{-}(0,y,s)| \leqslant f + 4b^2 + 2b^3, \tag{A5}$$

for $0 \leqslant y \leqslant 1$, $0 < s < 2y$. In order to satisfy the uniform bound on the iterates given in Eq. (A2), it therefore suffices to require that

$$f + 4b^2 + 2b^3 \leqslant b. \tag{A6}$$

The object is to now choose the largest value of f such that a positive b exists which satisfies Eq. (A6) and therefore Eq. (A2). This occurs when

$$8b + 6b^2 = 1 \tag{A7}$$

or

$$b = b_0 = (\sqrt{22} - 4)/6 \doteq 0.115\,07, \tag{A8}$$

and consequently

$$f = f_0 = (11\sqrt{22} - 50)/27 \doteq 0.059\,06. \tag{A9}$$

Having shown that it is possible for the iterates to remain uniformly bounded, it now must be demonstrated that the iterates actually converge. This follows from the contraction mapping principle, or equivalently from a comparison of successive iterates. To see this, assume

$$|F^{\pm}(s)| < f_1 < f_0, \quad 0 < s < 2, \tag{A10}$$

so that

$$|R_n^+(x,1,s)|, |R_n^-(0,y,s)| < b_1 < b_0$$

in the appropriate triangular regions. Define for $n = 1,2,...$ (with $R_0^{\pm} \equiv 0$)

$$c_n = \sup_{(x,y,s)} [|R_n^+(x,1,s) - R_{n-1}^+(x,1,s)|, |R_n^-(0,y,s) - R_{n-1}^-(0,y,s)|], \tag{A11}$$

with arguments x, y, s in the relevant domains. Now from Eq. (4.5) it follows that

$$\begin{aligned}|R_{n+1}^+(x,1,s) - R_n^+(x,1,s)| \leqslant \int_0^x & |B_n(R_n^+ - R_{n-1}^+) + R_{n-1}^+(B_n - B_{n-1}) \\ & + 2R_n^-(0,x',0^+)[R_n^+ * R_n^+ - R_{n-1}^+ * R_{n-1}^+] \\ & + 2[R_n^-(0,x',0^+) - R_{n-1}^-(0,x',0^+)](R_{n-1}^+ * R_{n-1}^+)|dx' \leqslant kc_n\end{aligned} \tag{A12}$$

where

$$k = 8b_1 + 6b_1^2 < 1, \tag{A13}$$

this last inequality following from the fact that $b_1 < b_0$. [In Eq. (A12) the $*$ denotes convolution in s and the suppressed arguments in B_n and the R^+ iterates are x' and $(x',1,s+2(x-x'))$, respectively.] Similarly, it can be shown that

$$|R_{n+1}^-(0,y,s) - R_n^-(0,y,s)| < kc_n,$$

and consequently

$$c_{n+1} < kc_n.$$

Hence, the iteration converges since $k < 1$. Standard arguments now show that if Eq. (A10) is satisfied, then the iteration converges to a unique limit as long as the initial iterates $R_1^{\pm}$ are bounded by $b_1 < b_0$.

Finally, to show continuous dependence on the scattering data, suppose two sets of reflection data $F^{\pm}$ and $\widetilde{F}^{\pm}$ both satisfy Eq. (A10) and

$$|F^{\pm}(s) - \widetilde{F}^{\pm}(s)| < \epsilon.$$

Denote the corresponding iterates by $R_n^{\pm}$ and $\widetilde{R}_n^{\pm}$. All of these iterates are uniformly bounded by some $b_1 < b_0$. Define

$$d_n = \sup_{(x,y,s)} [|R_n^+(x,1,s) - \widetilde{R}_n^+(x,1,s)|, |R_n^-(0,y,s) - \widetilde{R}_n^-(0,y,s)|], \tag{A14}$$

with x, y, s in the appropriate domains. Using Eq. (4.5) with each set of data then yields [in a manner similar to that in which Eq. (A12) was derived]

$$|R_{n+1}^+(x,1,s) - \widetilde{R}_{n+1}^+(x,1,s)| < \epsilon + kd_n,$$

where k is given by Eq. (A13). Consequently, it can be shown that

$$d_{n+1} < \epsilon + kd_n. \tag{A15}$$

Since $d_1 < \epsilon$ it follows from (A15) that

$$d_n < \epsilon \sum_{j=0}^{n-1} k^j < \frac{\epsilon}{1-k},$$

for $n = 1,2,...$. Hence,

$$|A(x) - \widetilde{A}(x)| < 4\epsilon/(1-k),$$

$$|B(x) - \widetilde{B}(x)| < 4\epsilon/(1-k),$$

where $k < 1$. This established the continuous dependence of A and B on the scattering data.

[1] G. Kristensson and R. J. Krueger, "Direct and inverse scattering in the time domain for a dissipative wave equation. Scattering operators," J. Math. Phys. **27**, 1667 (1986).

[2] J. P. Corones, M. E. Davison, and R. J. Krueger, "The effects of dissipation in one-dimensional inverse problems," in *Inverse Optics, Proceedings of the SPIE*, Vol. 413, edited by A. J. Devaney (SPIE, Bellingham, WA, 1983), pp. 107–114.

[3] J. P. Corones, M. E. Davison, and R. J. Krueger, "Dissipative inverse problems in the time domain," in *Inverse Methods in Electromagnetic Imaging*, NATO ASI series, Series C, Vol. 143, edited by W-M. Boerner (Reidel, Dordrecht, 1985), pp. 121–130.

[4] M. Jaulent, "Inverse scattering problems in absorbing media," J. Math. Phys. **17**, 1351 (1976).

[5] M. Jaulent, "Inverse scattering problem for LCRG transmission lines," J. Math. Phys. **23**, 2286 (1982).

[6] V. Weston, "On the inverse problem for a hyperbolic dispersive partial differential equation," J. Math. Phys. **13**, 1952 (1972).

[7] V. Weston and R. J. Krueger, "On the inverse problem for a hyperbolic dispersive partial differential equation. II," J. Math. Phys. **14**, 406 (1973).

[8] V. Weston, "On inverse scattering," J. Math. Phys. **15**, 209 (1974).

[9] R. J. Krueger, "An inverse problem for a dissipative hyperbolic equation with discontinuous coefficients," Quart. Appl. Math. **34**, 129 (1976).

[10] R. J. Krueger, "An inverse problem for an absorbing medium with multiple discontinuities," Quart. Appl. Math. **36**, 235 (1978).

[11] R. J. Krueger, "Numerical aspects of a dissipative inverse problem," IEEE Trans. Antennas Propag. **AP-29**, 253 (1981).

[12] J. P. Corones, M. E. Davison, and R. J. Krueger, "Direct and inverse scattering in the time domain via invariant imbedding equations," J. Acoust. Soc. Am. **74**, 1535 (1983).

[13] J. P. Corones, R. J. Krueger, and V. H. Weston, "Some recent results in inverse scattering theory," in *Inverse Problems of Acoustic and Elastic Waves*, edited by F. Santosa, Y. Pao, W. Symes, and C. Holland (SIAM, Philadelphia, PA, 1984), pp. 65–81.

[14] J. C. Bolomey, D. Lesselier, C. Pichot, and W. Tabbara, "Spectral and time domain approaches to some inverse scattering problems," IEEE Trans. Antennas Propag. **AP-29**, 206 (1981).

[15] A. G. Tijhuis, "Iterative determination of permittivity and conductivity profiles of a dielectric slab in the time domain," IEEE Trans. Antennas Propag. **AP-29**, 239 (1981).

Reflection and refraction of a spherical acoustic wave from a thin layer

N. C. Banik
Arco Oil and Gas Company, P. O. Box 2819, Dallas, Texas 75221

I. Lerche
Department of Geology, University of South Carolina, Columbia, South Carolina 29208

(Received 30 April 1985; accepted for publication 5 February 1986)

The behavior of a spherical acoustic wave impacting on a thin, parallel slab of material of thickness L is investigated. It is found that the reflected wave may have a so-called head-wave contribution when the slab has a higher acoustic velocity than the surrounding medium. However, the effects of the finite slab thickness are to delay the head-wave arrival time relative to that from a single interface, to diminish its amplitude, and modify the frequency response of the amplitude, and to cause multiples from the base of the slab to produce far-field interference fringes (the analog of Newton's rings). In addition, in the case where the slab has a lower acoustic velocity than the surrounding medium, no head wave results, but the far-field interference pattern persists. As the slab thickness is increased relative to the acoustic wavelength both the interference effects and the head-wave modifications increase with increasing thickness, for thicknesses small compared to the acoustic wavelength.

I. INTRODUCTION

In previous papers in this series Hill and Lerche[1] and Lerche[2] considered the behavior of the reflected component of a spherical acoustic wave incident upon an interface. The response of the wave to reflection from a rough interface was discussed,[1] as was the response to reflection from a smooth, but slightly curved, interface.[2] In particular the interest in both papers centered on the so-called head-wave component of the reflected signal,[3-6] which occurs ahead of the directly reflected specular signal,[3] at lateral reception points far distant from the source. The head wave then provides a clear diagnostic, uncluttered by overlapping detritus from the directly reflected signal, of acoustic impedence conditions at the interface. This fact is often used in seismology[5-8] to provide a measure of variations in subsurface lithologies with a depth for thick sedimentary layers that would otherwise be difficult to obtain.

On the other hand, it is certainly the case that not all lithologic units are sufficiently thick that they can be treated as single interface reflectors. In particular, a thin interbedding of shale and evaporite layers, as occurs, for example, in Ras-al-Khaimah,[9,10] where the beds are only a few feet thick, much less than the typical seismic wavelength of a 100 ft or so, would obviously produce a different head-wave response than would a single interface. This can be an important consideration in analyses attempting to unravel the causes for the behavior of acoustic wave amplitude with offset.

The question we address here is related to this problem. What is the acoustic reflection response of a layer of material, sandwiched between two identical semi-infinite media, to a spherical acoustic wave incident upon the interfaces between the layer and the surrounding media? Clearly as the layer thickness tends to 0 both the direct wave and its multiples, as well as the head wave, must become vanishingly small. Equally clearly, as the layer thickness increases the reflection response must reduce to that for a single interface, but at finite thicknesses the multiple reflections internal to the slab will modify both the direct reflection response as well as the head-wave reflection response. Our task is to determine the modifications as functions of slab thickness, wave frequency, and density and velocity of the slab and enveloping material. The organization of this paper is as follows. Section II sets up the basic acoustic equations and provides a formal expression for the total reflected field from a slab of thickness L immersed in an homogeneous medium. This is then converted to a far-field asymptotic expression and the contributions to the direct reflected wave and to the head wave are identified. The behavior of the reflected wave, when the slab's parameters forbid creation of a head wave, is taken up in Sec. III. Section IV evaluates the far-field asymptotic expressions and determines the modification to both the head-wave critical angle,[1] and to the head-wave amplitude and arrival time brought about by the internal multiple reflections in the slab.

Section V considers the contribution to the total reflected signal when the observation point is located in the vicinity of the critical angle so that the direct reflection and the head wave "merge."[1]

Section VI explores the variation of amplitude and phase behaviors for the head wave and direct wave as slab thickness, frequency, density, and acoustic velocity contrast between the slab and enveloping media vary. Finally in Sec. VII we provide a discussion of the results, suggest several areas of application, and comment on what remains to be done in order that we may use the results presented here and earlier.[1,2]

II. BASIC EQUATIONS, NOTATION, AND THE REFLECTED ACOUSTIC FIELDS

Acoustic waves of angular frequency ω are governed by the equation

$$\rho\nabla\cdot(\rho^{-1}\nabla p)+\omega^2 s^2 p=0\,, \tag{1}$$

where p is the pressure field and ρ and s are the position-

0022-2488/86/061694-07$02.50

dependent density and slowness. We place smooth boundaries at $z = 0$ and $z = L$ so that ρ and s take on the values ρ_1 and s_1 (ρ_2 and s_2) for $z<0$ and $z>L$ ($0<z<L$) (see Fig. 1).

We will use the convention of representing a three-dimensional vector by lowercase letters and its projection in the x-y plane by uppercase letters; e.g., a three-dimensional position vector is written $r = (\mathbf{R},z)$ with $\mathbf{R} = (x,y)$. Wave vectors are denoted in a similar fashion. In the medium with subscript properties 1, a plane wave component satisfying Eq. (1) has a three-dimensional wave vector $(\mathbf{K},k_1)$ with the z component

$$k_1 = (\omega^2 s_1^2 - K^2)^{1/2} . \tag{2}$$

The root in Eq. (2) is always taken such that $\operatorname{sgn}(z)\operatorname{Im}\{k_1\} < 0$. Likewise, in medium 2, the slab occupying $0<z<L$, we have the three-dimensional wave vector $(\mathbf{K},k_2)$ with the z component

$$k_2 = (\omega^2 s_2^2 - K^2)^{1/2} . \tag{3}$$

A point source is located on the positive z axis at z_0. The incident field is

$$P_{\text{inc}}(r) = \frac{i}{2\pi}\int dK \frac{\exp(i\mathbf{K}\cdot\mathbf{R} + ik_1|z-z_0|)}{k_1} \tag{4a}$$

$$= i\int_0^\infty J_0(KR)k_1^{-1}\exp(ik_1|z-z_0|)|K\, dK . \tag{4b}$$

Inspection of Eq. (1) shows that p and the normal derivative $\rho^{-1}\,\partial p/\partial n$ must be continuous across the boundaries $z = 0$ and $z = L$. Thus, the reflection coefficient for an incident plane-wave component of the source field making an angle θ with the normal to the slab is[11]

$$B(\theta) = r(\theta)[1 - \exp(i\delta(\theta))] \times [1 - \exp(i\delta(\theta))r(\theta)^2]^{-1} , \tag{5}$$

where $\delta = 2LK_*(1 - \alpha^2 \sin^2\theta)^{1/2}$ with

$$K_* = \omega s_2 ,$$

$$r(\theta) = [(1-\alpha^2\sin^2\theta)^{1/2} - \rho\cos\theta] \times [(1-\alpha^2\sin^2\theta)^{1/2} + \rho\cos\theta]^{-1} ,$$

$$\rho = \rho_2/\rho_1, \quad \alpha = s_1/s_2 ,$$

$$K = K_0 \sin\theta, \quad K_0 = s_1\omega .$$

Note that $B(\theta)$ possesses a branch cut with branch points at $\alpha\sin\theta = \pm 1$, which influence the structure of the far-field response for $\alpha > 1$. (The range of integration is only over $0<\theta<\pi/2$ for propagating disturbances in the far field.) Note also that $B(\theta)$ contains poles that contribute to the far-field radiation pattern. These poles occur at the roots of the transcendental equation

$$r(\theta) = \pm\exp(i\delta(\theta)/2) ,$$

which occur off the real θ axis. It is known[11] that these roots correspond to the points where the phase of the multiple along their paths is ($2\pi\times$integer) leading to a significant far-field contribution to the reflected radiation. As we shall see later, these multiples produce an analog of the Newton's rings effect.

Note further that for $\alpha<1$ it is not possible for the branch cut to contribute to the far-field integral; for $\alpha > 1$, however, not only does the branch cut provide a contribution, but the phase along the branch can exactly match with the phase along a multiple's path leading to a comingled contribution of head wave and multiple. We shall address this point later. (See Fig. 2.) For later use we note here that we can also write δ in the form

$$\delta = 2LK_0\alpha^{-1}[1-\alpha^2\sin\theta]^{1/2}$$
$$= 2L\omega s_2(1-\alpha^2\sin^2\theta)^{1/2} .$$

Since (4a) is an integral over plane waves, we have the reflected field

$$P_{\text{refl}} = i\int_0^\infty K\,dK \exp\{i[K_0^2 - K^2]^{1/2}|z+z_0|\} \times (K_0^2-K^2)^{-1/2}J_0(KR)B(\theta) . \tag{6}$$

Defining $z_* = r\cos\phi$, $R = r(\sin\phi)^{-1}$, we can write Eq. (6) in the form

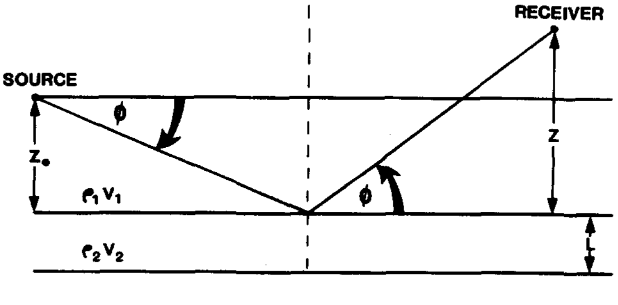

FIG. 1. Sketch of the geometric configuration.

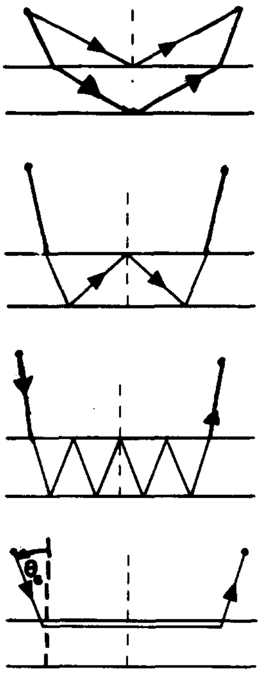

FIG. 2. Illustrative examples of ray paths that provide a far-field contribution. Note the presence of the multiples in the slab and the presence of the head-wave refraction path (for $\alpha > 1$).

$$P_{\text{refl}} = iK_0 \int_0^{i\infty} d\theta (\sin\theta)^{-1} B(\theta) \times \exp\{iK_0 r \cos\phi \cos\theta\} \times J_0(K_0 r \sin\phi \sin\theta) \quad (7)$$

along that path in complex-θ space corresponding to $0 \leqslant K \leqslant \infty$. We are interested in propagating disturbances at large distances from the source. Hence in (7) we can restrict the range of integration to $0 \leqslant K \leqslant \omega s_1$ (i.e., $0 \leqslant \theta \leqslant \pi/2$) and use the asymptotic representation for the Bessel function

$$J_0(KR) \simeq [2/(\pi KR)]^{1/2} \cos(KR - \pi/4)\,. \quad (8)$$

The problem of determining the mapping of such paths of integration from transverse wave number space K to angular space θ has been intensively investigated over the years. Kong,[12] in his Fig. 6.10 in particular, and in his Chap. 6 overall, provides an elegant treatment of the problem. For further details concerning the relevant path deformations and contours of integration in θ space we refer the interested reader to Ref. 12. We see no need to repeat that information here.[13]

When using (8) in (7), it is simple, but tedious, to show that stationary points occur only for the component of the cosine in (8) varying as $\frac{1}{2}\exp[i(KR - \pi/4)]$. Thus, retaining this term in the integral we obtain the asymptotic formula

$$P_{\text{refl}} = iK_0^{1/2} (2\pi r \sin\phi)^{-1/2} e^{-i\pi/4} \times \int_0^{i\infty} (\sin\theta)^{1/2} B(\theta) \exp[iK_0 r \cos(\theta - \phi)]\,d\theta \equiv iK_0^{1/2}\,(2\pi r)^{-1/2} e^{-i\pi/4} I\,. \quad (9)$$

Evaluation of (9) by the method of steepest descents is then in line with the asymptotic development of the far field. The factor $(\sin\theta)^{1/2}$ in (9) is caused by the asymptotic expansion of $J_0(KR)$. Thus stationary points near $\theta = 0$ caused by this factor are only apparent and we can see from (6) that actually there are no stationary phase contributions near $\theta = K = 0$.

The phase in (9) has a stationary point near $\theta = \phi$. We identify this stationary phase point in the usual manner[1-6] as providing the directly reflected wave component of the incident field. Any stationary phase point, say $\theta = \theta_*$, arising as a direct consequence of the square root $(1 - \alpha^2 \sin^2\theta)^{1/2}$ becoming pure imaginary in the domain of integration, we regard as providing the head-wave response. This separation of direct and head-wave contributions can be maintained as long as the phase points near $\theta = \phi$ and $\theta = \theta_*$ do not coalesce to within an angle of order $(K_0 r)^{-\beta}$ with $\beta = O(1)$ (see Ref. 1). At coalescence, the distinction between head wave and direct wave is moot as the two contributions merge into one.

Expression (9) provides the basis for the rest of this paper. As we shall see directly, the reflected wave behavior is crucially dependent upon whether $\alpha (\equiv s_1/s_2)$ is greater or less than unity. To anticipate: for $\alpha > 1$ we obtain a head-wave contribution, for $\alpha < 1$ we do not, i.e., head waves arise when a spherical acoustic wave is incident on a high velocity slab from a low velocity medium.

III. STATIONARY PHASE EVALUATION OF THE REFLECTED WAVE FOR $\alpha < 1$

Inspection of the structure of $B(\theta)$ given by Eq. (5) shows that the square root $(1 - \alpha^2 \sin^2\theta)$ remains real in $0 \leqslant \theta \leqslant \pi/2$.

In addition since $r(\theta)$ is then less than unity for all p and $\alpha (<1)$ it follows that $B(\theta)$ has no poles.

Stationary phase points in the reflected wave integral (9) can then arise only from the term involving $\exp[iN\cos(\theta - \phi)]$ (where $N \equiv K_0 r$) or from factors involving zeros of $1 - e^{i\delta(\theta)}$. Therefore we write the integral I in Eq. (9) in the form

$$I = \int_0^{\pi/2} \left(\frac{\sin\theta}{\sin\phi}\right)^{1/2} r(\theta)\,[1 - e^{i\delta(\theta)} r(\theta)^2]^{-1} \times \exp\{iN\cos(\theta - \phi) + \ln[1 - e^{i\delta(\theta)}]\}\,d\theta\,. \quad (10)$$

Stationary phase points of the exponential factor in Eq. (10) occur when

$$\sin(\theta - \phi)\,[1 - e^{i\delta(\theta)}] = (2LK_0)^2 e^{i\delta(\theta)}/(N\delta(\theta))\,. \quad (11)$$

Since $\delta > 0$ throughout $0 \leqslant \theta \leqslant \pi/2$, and since $N \gg 1$ because we are in the asymptotic regime, two different solutions are available for the stationary phase points.

A. The specular reflection phase point

One phase point of (11) is given approximately by

$$\theta_1 = \phi + (2LK_0)^2 e^{i\delta(\theta)} [1 - e^{i\delta(\phi)}]^{-1} \times [N\delta(\varphi)]^{-1} + O(N^{-2})\,, \quad (12)$$

and provides a contribution to the integral (10) in the amount

$$I_1 \simeq B(\phi)\,(2\pi/N)^{1/2} \exp[i(N - \pi/4)]\,. \quad (13)$$

B. The multiple reflection phase points

A second set of stationary phase points of (11) are given approximately by

$$\theta = \theta_n + i(2LK_0)^2 [N\sin(\theta_n - \phi) 2\pi n]^{-1} + O(N^{-2})\,, \quad (14)$$

where

$$2n\pi = [1 - \alpha^2 \sin^2\theta_n]^{1/2} 2K_0 L/\alpha\,, \quad n = 1, 2, \ldots, N_{\max}\,, \quad (15)$$

with $N_{\max}$ given by the nearest integer to, but less than,

$$\xi = LK_0/\pi\,. \quad (16)$$

Note that these stationary phase points, which represent the contribution to the reflected field from multiple "bounces" off the surfaces of the slab, provide contributions in the far field only if

$$\min\{1, (LK_0/\pi)^2\} \gg \alpha^2 \gg [1 + (\pi LK_0)^2]^{-1} \equiv \alpha_{\min}^2\,, \quad (17)$$

representing the fact that multiples produced in the slab have to have a total travel time in the slab such that they stay in phase to within a fraction of a wavelength with the primary spherical wave.

Multiples generated outside of the range of angles allowed by (15)–(17) provide a much smaller contribution to the reflected wave in the far field.

We shall assume in writing the far-field stationary phase expressions that we are dealing with multiple reflections satisfying (15)–(17). The stationary phase contribution to I produced by the nth multiple then provides the contribution

$$\begin{aligned} I_n &= (\sin\theta_n/\sin\phi)^{1/2} r(\theta_n)(2LK_0)^2 \\ &\quad \times (2\pi)^{-1/2}[1-r(\theta_n)^2]^{-1}[N\sin(\theta_n-\phi)]^{-2} \\ &\quad \times \exp\{iN\cos(\theta_n-\phi)+i(2LK_0)^2 \\ &\quad \times [2\pi n\sin(\theta_n-\phi)]^{-1}\}, \end{aligned} \tag{18}$$

where

$$r(\theta_n) = (n\pi\alpha - \rho K_0 L\cos\theta_n)/(n\pi\alpha+\rho K_0 L\cos\theta_n). \tag{19}$$

In computing the stationary phase point contributions we have made the implicit assumption that the specular reflection phase point is distinct in angular position from any of the multiple reflection phase points by an angle of about $O(N^{-1/2})$ in order that we can treat the stationary phase contributions from each separately.

We now work out the contribution when this is not the case.

C. Comingled phase points: $|\theta_n - \phi| < O(N^{-1/2})$

In the case both factors on the left-hand side of (11) contribute to the stationary phase point at some particular n value, say $n=m$. Then

$$\sin^2\phi = \alpha^{-2} - m^2\pi^2/(LK_0)^2, \tag{20}$$

so that a discrete set of reflection angles exist where a multiple and the primary interfere constructively or destructively. The comingled stationary phase point is then at

$$\theta_c = \phi + (\sin\phi)^{1/2}e^{i\pi/4}N^{-1/2} + O(N^{-1}) \tag{21}$$

and the comingled stationary phase contribution to the integral I is

$$\begin{aligned} I_c &= (\sin\phi)^{3/2}r(\phi)[1-r(\phi)^2]^{-1}(\cos\phi)^{1/2} \\ &\quad \times (LK_0)^2N^{-1}2^{5/4}e^{-1/2}(\alpha m\pi^{1/2})^{-1} \\ &\quad \times \exp[i(N+3\pi/8)], \end{aligned} \tag{22}$$

where

$$r(\phi) = (\alpha m\pi - \rho LK_0\cos\phi)/(\alpha m\pi + \rho LK_0\cos\phi). \tag{23}$$

We defer discussion of these contributions to the reflected wave field until after we have evaluated the similar contributions arising when the head wave is permitted to exist. Contrasting behaviors are then more easily compared and discussed.

D. Stationary phase points for $\alpha = 1$

In this case the only difference between the slab and surrounding medium is due to density contrast. In this case $r(\theta) = (1-\rho)/(1+\rho) \equiv r_0$ and $\delta = 2LK_0\cos\theta \equiv \Delta\cos\theta$, so that the integral I in Eq. (9) reduces to

$$\begin{aligned} J &\equiv (1+\rho)(1-\rho)^{-1}I \\ &= \int_0^{\pi/2} [1-e^{i\Delta\cos\theta}][1-r_0^2e^{i\Delta\cos\theta}]^{-1} \\ &\quad \times (\sin\theta/\sin\phi)^{1/2}\exp[iN\cos(\theta-\phi)]d\theta. \end{aligned} \tag{24}$$

Equation (24) has separate stationary phase points at

$$\begin{aligned} \theta_0 &= \phi + \Delta\sin\phi e^{i\Delta\cos\phi} \\ &\quad \times [N(1-e^{i\Delta\cos\phi})]^{-1} + O(N^{-2}) \end{aligned} \tag{25a}$$

and at

$$\begin{aligned} \cos\theta_n &= 2n\pi/\Delta + i\Delta\sin\theta_n^{(0)}[N\sin(\theta_n^{(0)}-\phi)]^{-1} \\ n &= 1,2,\ldots,n_{\max}; \end{aligned} \tag{25b}$$

where $\cos\theta_n^{(0)} = 2n\pi/\Delta$ and where $n_{\max}$ is the nearest integer less than $\Delta/2\pi$. If $\Delta < 2\pi$, there are no stationary phase points from the multiples. When a phase point θ_n with $n=m$ is comingled with θ_0, the comingled stationary phase point is at

$$\theta = \phi + (\sin\phi)^{1/2}e^{i\pi/4}N^{-1/2} + O(N^{-1}). \tag{26}$$

It is then a simple matter to write down the stationary phase contributions to the integral I in (24). From the direct reflection phase point $\theta = \theta_0$ we obtain the contribution

$$\begin{aligned} J_0 &= [1-e^{i\Delta\cos\phi}][1-r_0^2e^{i\Delta\cos\phi}]^{-1} \\ &\quad \times (2\pi/N)^{1/2}e^{i(N-\pi/4)}. \end{aligned} \tag{27a}$$

From the nth multiple stationary phase point at $\theta = \theta_n$, we obtain the contribution

$$\begin{aligned} J_n &= (\sin^3\theta_n^{(0)}/\sin\phi)^{1/2}\Delta(2\pi)^{1/2}(1-r_0^2)^{-1} \\ &\quad \times [N\sin(\theta_n^{(0)}-\phi)]^{-2}\exp\{iN\cos(\theta_n^{(0)}-\phi) \\ &\quad + i\Delta\sin\theta_n^{(0)} + i\pi/2\}. \end{aligned} \tag{27b}$$

In the event that the mth multiple stationary phase point comingles with the point near $\theta = \phi$, i.e., $|\theta_n^{(0)} - \phi| \leqslant O(N^{-1/2})$, we then obtain a comingled contribution

$$J_0 = -\sin\phi\Delta(\pi/e)^{1/2}[1-r_0^2e^{i\Delta\cos\phi}]^{-1}N^{-1}e^{iN} \tag{27c}$$

from the contribution in the vicinity of the stationary point

$$\theta(\text{comingled}) = \phi + e^{3i\pi/4}N^{-1/2} \text{ with } \Delta\cos\theta = 2n\pi. \tag{28}$$

IV. STATIONARY PHASE POINT EVALUATION OF THE REFLECTED WAVE FOR $\alpha > 1$

In this case a fundamentally different type of behavior can occur than for the cases $\alpha \leqslant 1$. The reason is that the square-root factor $(1-\alpha^2\sin^2\theta)^{1/2}$ not only goes through zero in the domain $0 \leqslant \theta \leqslant \pi/2$ at $\theta = \theta_c$ ($\sin\theta_c = \alpha^{-1}$), but its behavior changes from being purely real for $\theta < \theta_c$ to pure imaginary for $\theta > \theta_c$. Thus we have the possibility of producing a different "character" to the stationary phase contributions including different coherence properties—the head wave is the direct consequence of this change in character.[3–6]

We already have an expression in the literature[1–6] for the behavior of the head-wave response for the case of a single interface. We can anticipate here and note that we expect major modifications to be made to the head wave when the slab is thin (in the sense $LK_0 \ll 1$) for it is under such conditions that the phase factor $\exp[i\delta(\theta)]$ not only varies slowly but also has a small imaginary part in $\theta > \theta_c$ [a large imaginary part would reduce $B(\theta)$ to $r(\theta)$ within a small angular range $O((LK_0)^{-2})$ centered on $\theta = \theta_c$ so that the head-wave behavior would then be as though from a single interface].

Hence the dominant modifications to the head wave behavior occur for $LK_0 \ll 1$ and, accordingly, we restrict our investigation to the thin slab.

Expanding $B(\theta)$ in powers of LK_0, we obtain

$$B(\theta) = B_1 + B_2\,, \tag{29}$$

with

$$\begin{aligned} B_1 = &-iLK_0[2\alpha\rho\cos\theta]^{-1}[1-\alpha^2\sin^2\theta-\rho^2\cos^2\theta] \\ &\times\exp(iLK_0\rho\cos\theta) \end{aligned} \tag{30}$$

and

$$B_2 = -2iLK_0\alpha^{-1}(1-\alpha^2\sin^2\theta)^{1/2}B_1(\theta) + O((LK_0)^2)\,, \tag{31}$$

where B_2 is the lowest power in LK_0 that contains the square-root factor. Following a conventional format[1-3] we identify B_1 as providing the directly reflected wave behavior and the stationary phase point in B_2 near $\theta = \theta_c$ as contributing to the head-wave behavior. Other stationary phase contributions from B_2 we regard as providing modifications to direct wave brought about by the existence of the head wave.[3] The asymptotic integral I in Eq. (9) then also splits into two parts, $I = I_1 + I_2$, with

$$\begin{aligned} I_1 = &-iLK_0(2\alpha\rho)^{-1}\int_0^{\pi/2}\left(\frac{\sin\theta}{\sin\phi}\right)^{1/2} \\ &\times(\cos\theta)^{-1}(1-\alpha^2\sin^2\theta-\rho^2\cos^2\theta) \\ &\times\exp[iN\cos(\theta-\phi)+iLK_0\rho\cos\theta]\mathrm{d}\theta\,, \end{aligned} \tag{32}$$

and

$$\begin{aligned} I_2 = &-(LK_0)^2(2\rho\alpha^2)^{-1}\int_0^{\pi/2}\left(\frac{\sin\theta}{\sin\phi}\right)^{1/2} \\ &\times(\cos\theta)^{-1}(1-\alpha^2\sin^2\theta)^{1/2} \\ &\times(1-\alpha^2\sin^2\theta-\rho^2\cos^2\theta) \\ &\times\exp[iN\cos(\theta-\phi)+iLK_0\rho_0\cos\theta]d\theta\,. \end{aligned} \tag{33}$$

The integral I_1 has a stationary phase point at $\theta \simeq \phi$ yielding the contribution

$$\begin{aligned} I_1 \simeq &-iLK_0(2\alpha\rho)^{-1}(2\pi/N)^{1/2}(\sec\phi)^{1/2} \\ &\times(1-\alpha^2+(\alpha^2-\rho^2)\cos^2\phi) \\ &\times\exp[i(N+LK_0\rho\cos\phi-\pi/4)]\,. \end{aligned} \tag{34}$$

The integral I_2 has two stationary phase points, one in the vicinity of $\theta = \phi$, which we regard as contributing to the direct wave, the other in the vicinity of $\theta = \theta_c$, which we regard as giving rise to the head-wave behavior. For the moment we deal with the phase points as separate, not comingled.

A. Contribution to the direct wave

Inspection of Eq. (32) and (33) shows that the direct wave contribution from I_2 is of order (LK_0) smaller than that arising from I_1. Hence we can ignore the stationary phase point contribution in I_2 from the vicinity of $\theta = \phi$.

The contribution to I provided by the stationary phase point in I_1 is then about

$$\begin{aligned} I_{\text{direct}} \simeq &-iLK_0(2\alpha\rho)^{-1}\sec\phi(2\pi/N)^{1/2} \\ &\times(1-\alpha^2+(\alpha^2-\rho^2)\cos^2\phi) \\ &\times\exp[i(N+LK_0\rho\cos\phi)]\,. \end{aligned} \tag{35}$$

B. Contribution to the head wave

Following a procedure similar to one laid down elsewhere,[1,2] we rewrite I_2 in the form

$$\begin{aligned} I_2 = &-(LK_0)^2\sin\theta_c(2\rho)^{-1}\int_0^{\pi/2}\left(\frac{\sin\theta}{\sin\phi}\right)^{1/2} \\ &\times\sec\theta(1-\alpha^2+(\alpha^2-\rho^2)\cos^2\theta) \\ &\times\exp\{iN\cos(\theta-\phi)+\tfrac{1}{2}\ln[\sin^2\theta_c-\sin^2\theta]\}d\theta\,. \end{aligned} \tag{36}$$

The exponential in (36) has a head-wave stationary phase point at

$$\theta \simeq \theta_c + i(2N\sin(\theta_c-\phi))^{-1}\,, \tag{37}$$

which yields the head-wave contribution from I_2 of

$$\begin{aligned} I_{\text{head}} \simeq &(LK_0)^2(\sin^3\theta_c/\sin\phi)^{1/2}\rho\cos\theta_c \\ &\times(\pi/(8e))^{1/2}[N|\sin(\theta_c-\phi)|]^{-1} \\ &\times\exp\{i[N\cos(\theta_c-\phi)+LK_0\rho\cos\theta_c]\}\,, \end{aligned} \tag{38a}$$

and, following standard procedure,[1-3] we have evaluated the head-wave contribution only in the regime $\phi > \theta_c$, since it is only for large angles that the reflected head wave arrives at a receiver prior to the direct wave. In terms of the conventional notation[3] using the head-wave travel time

$$\tau_h \equiv s_1 r\cos(\theta_c-\phi)$$

and the distance l $[\equiv r\sin(\theta_c-\phi)]$ that the head wave travels through the medium with slowness s_2 (see Fig. 6.10 of Ref. 3), we can use Eqs. (9) and (38) to write the head-wave pressure field in the form

$$\begin{aligned} P_{\text{head}} = &\left(\frac{2}{e}\right)^{1/2} r^{-1/2}\frac{L^2K_*^2}{4K_0 l^{3/2}}\left(\frac{\rho_2}{\rho_1}\right)\left(1-\frac{s_1^2}{s_2^2}\right)^{3/2} \\ &\times\exp[i\omega(\tau_h+s_2L\cos\theta_c-t)]\,, \end{aligned} \tag{38b}$$

so that, as well as having its amplitude cut down by the interference between the upper and lower bedding planes, the head wave is also delayed by the extra amount $s_2L\cos\theta_c$ relative to head-wave arrivals from a single interface.

C. Comingled stationary phase points: $|\phi-\theta_c| < O(N^{-1/2})$

In this case the direct wave stationary phase point at $\theta = \phi + O(N^{-1})$ combines with the head-wave phase point to yield a single, comingled, stationary phase point at

$$\theta = \theta_c + e^{i\pi/4}(2N)^{-1/2}\,,$$

which provides a contribution to I_2 in the amount

$$I_{\text{comingled}} \simeq i\rho(LK_0)^2(\sin\theta_c\cos\theta_c)^{3/2}(\tfrac{1}{2}\pi/e)^{1/2}N^{-3/4} \times\exp[i(N+LK_0\rho\cos\theta_c)]\,. \quad (39)$$

Having explored the basic mathematical contributions to the reflected wave behavior we now turn our attention to an assessment of their relative behaviors.

V. REFLECTED WAVE BEHAVIOR FOR $\alpha \lesssim 1$

A. The case $\alpha < 1$ (low velocity slab, high velocity medium)

The relative contributions to the integral I in Eq. (9) arise from the direct wave (stationary phase point near $\theta = \phi$) as given through Eq. (13) and from the multiples generated in the slab as given through Eq. (18) [stationary phase points near $\theta = \theta_n$ provided inequality (17) is obeyed].

Apart from factors of order unity, the relative magnitudes of the separated (not comingled) multiple contributions relative to the direct wave have the large N behavior $|I_n|/|I_1| = O(N^{-3/2})$ so that, in the far field, the multiples are always a very small contribution. When comingled, the mth multiple plus direct wave has magnitude $|I_m|/|I_1| = O(N^{-1/2})$ so that the direct wave is diminished by $O(N^{-1/2})$ in amplitude by interference with the mth multiple.

The multiple contributions $[I_n = O(N^{-2})]$ to the far field are so small because of the curvature of the incident spherical wave. Unlike a single plane-wave incident at a fixed angle on the slab, the spherical wave consists of a superposition of plane waves. The usual plane-wave phase coherence of the reflected multiples is diminished by the superposition effect of the spherical wave. The individual plane wave contributions rapidly become phase incoherent, one with respect to another, thereby destructively interfering to diminish the multiples' far-field contribution.

Unless a multiple comingles in phase with the direct wave, its far-field behavior is minute compared to the direct wave.

At the specific angles $\theta_m = \phi$ the *joint* contribution of the direct wave and mth multiple is smaller than the direct wave so that dark bands, of intensity $O(N^{-1/2})$ compared to the direct wave at angles away from $\theta_m = \phi$, will be present in the far-field radiation pattern—an analog to Newton's rings. [Newton's rings are caused by interference of the reflected primary waves from the two interfaces of a wedge (varying thickness). In this sense $B(\phi)$ itself is directly analogous to Newton's rings.]

These bands narrow as distance increases, and there are a finite number $n_{\max}$ of them, with $n_{\max}$ given through Eqs. (15) and (16), provided further that inequality (17) is in force.

If α steps out of the bounds allowed by Eq. (17) or if, for a given α, the wave number K_0 or slab thickness L are such that inequality (17) is violated, then there are no far-field multiple contributions. The far-field pattern is then due to the direct contribution $B(\phi)$. Reflection of a spherical wave from the slab behaves as it would from a plane wave (apart from a divergence factor r^{-1}). The thickness would modify the amplitude as a function of offset as if the source were a plane wave.

B. The case $\alpha = 1$ (equal velocities for slab and medium)

Here the far-field behavior is essentially identical in character to that for $\alpha < 1$. Again there is a direct wave contribution [Eq. (27a)], multiple contributions [Eq. (27b)], which are of order $N^{-3/2}$ compared to the direct wave, and a comingled contribution [Eq. (27c)] of order $N^{-1/2}$ compared to the direct wave so that, once again, dark angular bands of width $O(N^{-1/2})$ centered on $\phi = \theta_m$ will be present in the reflected wave's angular intensity. In this case the dark zones occur at angular positions where $\phi = \cos^{-1}(n\pi/LK_0)$, which is the limiting case for satisfaction of inequality (17).

C. The case $\alpha > 1$ (high velocity slab, low velocity medium)

In this case we restricted our attention to the case of a thin slab $LK_0 \ll 1$ since we wished to concentrate on the head-wave behavior as modified by the slab nature of the medium.

In terms of intensity, the relative contributions to the integral I from the head wave [Eq. (38)] and the direct wave [Eq. (35)] are in the ratio $LK_0/N^{1/2}$ apart from factors of order unity (for comparison, in the case of a single interface the ratio is of order[1-3] $N^{-1/2}$). Thus the effect of a thin slab is to diminish the head wave amplitude by a factor $O(LK_0)$ relative to the direct wave, and to diminish the head-wave amplitude by a factor $O((LK_0)^2)$ relative to the head wave from a single interface.

From Eq. (38) we see that the usual propagation aspects of the head wave are maintained[1-6]: it arrives ahead of the direct wave for large offset angles θ_c and the head-wave arrival time is independent of frequency (nondispersive). An "extra" time delay, $s_2L\cos\theta_c$, now arises because of wave interference at the bedding plane boundaries of the slab. The factor $(LK_0)^2$ in the amplitude and the term $s_2L\cos\theta_c$ in the phase are consequences of the assumption that $LK_0 \ll 1$. When $LK_0 \approx 1$ the derivation fails. Of course then the various reflections separate out temporally.

In the case of a comingled direct wave and head wave, Eq. (39) shows that the joint amplitude varies as $O(LK_0)^2/N^{3/4}$ over an angular range $O(N^{-1/2})$ centered on $\theta_c = \phi$.

VI. DISCUSSION AND CONCLUSION

The purpose of this paper was to see what modifications were introduced to reflection of a spherical acoustic wave that impacted on a slab of finite thickness compared to impaction on a single interface.

We found that the dominant effect, irrespective of the relative contrast in density or velocity, was for a thin slab to cut down the total reflected wave by amounts of order $LK_0(\ll 1)$ for the direct wave and of order $(LK_0)^2$ for head waves relative to single interface results. We also found that the existence of multiples, caused by reflections of the incident spherical wave off the top and bottom surfaces of the slab, has the effect of introducing narrow, angular bands of darkness into the reflected field.

In the case where a low velocity slab is surrounded by a high velocity medium, these bands are dispersive in frequency, exist only in restricted frequency bands (for a given velocity contrast and given slab thickness), and become narrower in angle as the observation point moves further and further away from the slab. The direct wave dominates under all such conditions.

In the opposite case, where a high velocity slab is surrounded by a low velocity medium, we found that the head wave still has its single interface propagation characteristics (e.g., arrives ahead of the direct wave for observation angles greater than θ_c, is nondispersive in frequency), but that its amplitude is not only smaller than that from a single interface but also has a different frequency dependence, so that the temporal structure of the head wave is changed. The direct wave, while also weakened, is still more dominant in intensity than the head wave. There is also an extra time delay for the head wave due to the finite thickness of the slab in the amount $\Delta t = s_2 L \cos \theta_c$.

In order to improve our understanding of the behavior of head waves in the subsurface of the earth (so that we can better use them as diagnostic devices for determining subsurface structure) we still need to investigate the behavior of a spherical acoustic source impacting on one or more interfaces (flat, rough, curved) when due allowance is made for conversion of part of the incident wave to shear waves[6] at the boundaries between the different media.

These problems will be taken up in future papers in this series.

ACKNOWLEDGMENTS

That portion of I.L.'s work performed at the University of South Carolina was funded by the Basin Modelling Industrial Associates Program.

This work was performed after N.C.B. left Gulf Research and Development Company and before he joined Arco Oil and Gas Company.

[1] N. R. Hill and I. Lerche, J. Math. Phys. **26**, 1420 (1985).

[2] I. Lerche, accepted by J. Acoust. Soc. Am.

[3] K. Aki and P. G. Richards, *Quantitative Seismology: Theory and Methods* (Freeman, San Francisco, 1980).

[4] D. C. Stickler, J. Acoust. Soc. Am. **60**, 1061 (1976); **64**, 869 (1978).

[5] L. M. Brekhovskikh, *Waves in Layered Media* (Academic, New York, 1960).

[6] J. A. DeSanto, "Coherent scattering from rough surfaces," in *Mathematical Methods and Applications of Scattering Theory. Lecture Notes in Physics*, edited by J. A. DeSanto, A. W. Saenz, and W. W. Zackary (Springer, New York, 1980).

[7] W. C. Chew and J. A. Kong, Geophys. **46**, 309 (1981).

[8] V. Cervený and R. Ravindra, *Theory of Seismic Head Waves* (Univ. Toronto, Toronto, 1971).

[9] N. C. Banik, I. Lerche, and R. T. Shuey, Geophys. **50**, 2768 (1985).

[10] J. Resnick, I. Lerche, and R. T. Shuey, accepted by Geophys. J. R. Astron. Soc.

[11] M. Born and E. Wolf, *Principles of Optics* (Pergamon, New York, 1964).

[12] J. A. Kong, *Theory of Electromagnetic Waves* (Wiley, New York, 1975).

[13] We drew the incorrect contours of integration in an earlier version of this paper, as was pointed out to us by the referee, although the mathematic results were correct. The referee made us aware of Ref. 12, and it does an excellent job of describing the transformations necessary—far better than we could do. We urge the reader to consult Ref. 12 for details.

Layer-stripping solutions of multidimensional inverse scattering problems

Andrew E. Yagle and Bernard C. Levy
Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 11 June 1985; accepted for publication 21 February 1986)

A layer-stripping procedure for solving three-dimensional Schrödinger equation inverse scattering problems is developed. This method operates by recursively reconstructing the potential from the jump in the scattered field at the wave front, and then using the reconstructed potential to propagate the wave front and the scattered field further into the inhomogeneous region. It is thus a generalization of algorithms that have been developed for one-dimensional inverse scattering problems. Although the procedure has not yet been numerically tested, the corresponding one-dimensional algorithms have performed well on synthetic data. The procedure is applied to a two-dimensional inverse seismic problem. Connections between simplifications of this method and Born approximation inverse scattering methods are also noted.

I. INTRODUCTION

The inverse scattering problem for the Schrödinger equation in three dimensions with a time-independent, local, nonspherically symmetric potential has a wide variety of applications. In particular, the inverse seismic problem of reconstructing the density and wave speed of an inhomogeneous isotropic acoustic medium from surface measurements of the medium reponse to an excitation can be formulated as a Schrödinger inverse scattering problem, as was done by Coen *et al.*[1] The plasma wave equation, which describes the propagation of electromagnetic waves in the ionosphere, is also related to the Schrödinger equation by a Fourier transformation with respect to time. Some connections between inverse scattering problems for the Schrödinger equations and for the plasma wave equation have been noted by Rose *et al.*[2]

Two major approaches for solving Schrödinger inverse scattering problems in dimensions greater than 1 are available. The first approach consists of using the *first Born approximation*, in which the wave field inside the inhomogeneous region (where the potential differs from zero) is approximated by the incident wave being used to probe the region. This approach has been applied to the inhomogeneous wave equation by Cohen and Bleistein,[3] Devaney,[4] and others. The second approach is the generalized Marchenko procedure due to Newton,[5] in which the wave field is reconstructed inside the inhomogeneous region by solving a Marchenko integral equation for each direction of the incident probing. The potential is then recovered by an equation commonly referred to as the "miracle" [Eq. (2.11) below].

Both of these methods have shortcomings. The Born approximation constitutes a single scattering approximation, and thus requires the assumption of weak scattering. Newton's integral equation method avoids this problem, but requires the scattering amplitude (measured in the far field) for all incident and outgoing directions and all frequencies. This results in an overdetermined problem, where a slight corruption of the data may result in an inadmissible scattering amplitude. Also, the transmission data required for complete characterization of the scattering amplitude is generally unavailable in inverse seismic problems, since only surface measurements (in the *near* field) are available. Finally, the necessity of computing the entire wave field for each incident direction of probing is clearly inefficient, since the form of the "miracle" equation shows that much of this calculation is redundant.

In this paper a third approach to Schrödinger inverse scattering problems is discussed. A *layer-stripping* procedure is derived, which recursively reconstructs the wave field and the potential simultaneously. To see how this procedure works, consider the plasma wave equation [Eq. (2.5) below]. As the probing wave penetrates the inhomogeneous region, the jump in the wave field at the wave front yields the potential along the wave front. This was first noted by Morawetz,[6] and is called the "fundamental identity" by DeFacio and Rose.[7] However, the reconstructed potential now can be used to propagate the wave field deeper into the inhomogeneous region, and the jump in the wave field at the new position of the wave front yields the potential along this new position. In this way, the potential for the entire inhomogeneous region is reconstructed recursively, rather than in one huge batch operation as with the "miracle" equation.

There are several advantages to using a layer-stripping technique. Only one direction of probing is required, and only backscattered data is used. This makes the procedure more applicable to inverse seismic problems, and also removes the problems of overdetermination and possibly inconsistent data. The procedure is in principle exact, since all multiple reflection, refraction, and diffraction effects are accounted for. Approximation is inherent only in the discretization necessary to implement the algorithm numerically, and data at all frequencies are used. Finally, the algorithm requires less computation than would the solution of Newton's Marchenko integral equation for even one incident direction, if this were possible [note that the coupling of p_s in Eq. (2.8) below precludes this possibility]. This is because the layer-stripping procedure exploits the structure of the inverse scattering problem itself. This structure is manifested as the Hankel structure of the Marchenko integral equation, which can be exploited to reduce the amount of computation required to solve it. However, it should be noted that Newton's procedure allows bound states (square-integrable

0022-2488/86/061701-10$02.50

solutions with negative energy), whereas the applicability of layer stripping to problems with bound states is still unsettled.

The layer-stripping concept has been used to obtain fast algorithm solutions for the one-dimensional Schrödinger inverse scattering problem by Corones *et al.*,[8] Symes,[9] Bruckstein *et al.*,[10] and Yagle and Levy.[11] This approach has also been applied to various inverse seismic problems by Bube and Burridge[12] and Yagle and Levy.[13–15] Similar approaches have been used by other authors. Results of computer runs of these algorithms have been encouraging (see Bube and Burridge[12] and Yagle[16]). Previous application of this concept to multidimensional inverse seismic problems has been limited to Yagle[16] and Symes.[17]

It should be noted that the algorithms proposed in this paper have not yet been numerically tested, and their numerical stability is presently unknown. However, the performance of the one-dimensional problem algorithms is encouraging. In any event, the insight gained into the inversion process is interesting in its own right.

The structure of this paper is as follows. In Sec. II Newton's integral equation procedure, including the "miracle," is quickly reviewed and interpreted using results from Rose *et al.*[2,18] This allows relationships between this approach and the layer-stripping approach to be noted later. In Sec. III a layer-stripping procedure for solving the three-dimensional Schrödinger inverse scattering problem is derived and discussed. In Sec. IV this algorithm is applied to the 2½-dimensional inverse acoustic problem with a harmonic source, which was considered by Coen *et al.*[1] This results in a solution procedure requiring only surface data, in contrast to the procedure of Coen *et al.*,[1] which requires transmission data that are difficult to obtain for an inverse seismic problem. In Sec. V some Born approximation results are quickly reviewed and are then related to a simplification of the layer-stripping algorithm. Section VI concludes the paper with a discussion and summary of results.

II. INTEGRAL EQUATION METHODS

The inverse scattering problem considered in this paper is as follows. The wave field $\hat{p}(\mathbf{x},k)$ satisfies the Schrödinger equation

$$(\nabla^2 + k^2 - V(\mathbf{x}))\hat{p}(\mathbf{x},k) = 0, \tag{2.1}$$

where the potential $V(\mathbf{x})$ is real valued, smooth, and has compact support. It is also assumed that $V(\mathbf{x})$ does not induce bound states; a sufficient condition for this is for $V(\mathbf{x})$ to be non-negative.

Scattering solutions of Eq. (2.1) are given by the Lippman–Schwinger equation

$$\hat{p}(\mathbf{x};k;\mathbf{e}_i) = e^{-ik\mathbf{e}_i\cdot\mathbf{x}} - \int (4\pi|\mathbf{x}-\mathbf{y}|)^{-1} e^{-ik|\mathbf{x}-\mathbf{y}|} \times V(\mathbf{y})\hat{p}(\mathbf{y},k;\mathbf{e}_i)\,d^3\mathbf{y}, \tag{2.2}$$

where the incident wave is a plane wave in the direction of the unit vector $\mathbf{e}_i$. Letting $\mathbf{x} = |\mathbf{x}|\mathbf{e}_s$ and taking $|\mathbf{x}|\to\infty$, we have, in the far field,

$$\hat{p}(\mathbf{x};k\mathbf{e}_i) = e^{-ik\mathbf{e}_i\cdot\mathbf{x}} + (e^{-ik|\mathbf{x}|}/4\pi|\mathbf{x}|)A(k,\mathbf{e}_s,\mathbf{e}_i) + O(|\mathbf{x}|^{-2}), \tag{2.3}$$

where

$$A(k,\mathbf{e}_s,\mathbf{e}_i) \triangleq -\int e^{-ik\mathbf{e}_s\cdot\mathbf{y}}V(\mathbf{y})\hat{p}(\mathbf{y},k;\mathbf{e}_i)\,d^3\mathbf{y} \tag{2.4}$$

is the scattering amplitude for incident direction $\mathbf{e}_i$ and scattered direction $\mathbf{e}_s$.

Taking the inverse Fourier transform of Eq. (2.1) with respect to k yields the plasma wave equation

$$\left(\nabla^2 - \frac{\partial^2}{\partial t^2} - V(\mathbf{x})\right)p(\mathbf{x},t) = 0, \tag{2.5}$$

where

$$p(\mathbf{x},t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{p}(\mathbf{x},k)e^{ikt}\,dk. \tag{2.6}$$

Equation (2.5) models the propagation of electromagnetic waves in the ionosphere, as noted by DeFacio and Rose[7]; in two dimensions, it can also be interpreted as the equation for an elastically braced membrane. The inverse Fourier transform of Eq. (2.3) is

$$\hat{p}(\mathbf{x},t;\mathbf{e}_i) = \delta(t - \mathbf{e}_i\cdot\mathbf{x}) + (4\pi|\mathbf{x}|)^{-1}R(t - \mathbf{e}_s\cdot\mathbf{x},\mathbf{e}_s,\mathbf{e}_i) + O(|\mathbf{x}|^{-2}), \tag{2.7}$$

where $R(t,\mathbf{e}_s,\mathbf{e}_i)$ is the inverse Fourier transform of $A(k,\mathbf{e}_s,\mathbf{e}_i)$ and therefore represents the observed time response to the probing impulsive plane wave $\delta(t - \mathbf{e}_i\cdot\mathbf{x})$ in the far field. As an alternative, of course, the near-field response could be measured. So far, we have followed Rose *et al.*[2]

Newton's procedure[5] for recovering the potential $V(\mathbf{x})$ from the scattering amplitude $A(k,\mathbf{e}_s,\mathbf{e}_i)$ is as follows. First, solve the *Marchenko integral equation* [Eq. (4.19) of Rose *et al.*[2]]

$$p_s(\mathbf{x},t;\mathbf{e}_i) = \int_{S^2}\int_{-\mathbf{e}_s\cdot\mathbf{x}}^{\infty} M(t+\tau,\mathbf{e}_s,\mathbf{e}_i)p_s(\mathbf{x},\tau,\mathbf{e}_s)\,d\tau\,d^2\mathbf{e}_s + \int_{S^2} M(t - \mathbf{e}_s\cdot\mathbf{x},\mathbf{e}_s,\mathbf{e}_i)\,d^2\mathbf{e}_s \tag{2.8}$$

for the scattered field $p_s(\mathbf{x},t;\mathbf{e}_i)$, which is simply

$$p_s(\mathbf{x},t;\mathbf{e}_i) = p(\mathbf{x},t;\mathbf{e}_i) - \delta(t - \mathbf{e}_i\cdot\mathbf{x}), \tag{2.9}$$

and where S^2 denotes the unit sphere in $\mathbf{R}^3$.

The quantity $M(t,\mathbf{e}_s,\mathbf{e}_i)$ is obtained from $A(k,\mathbf{e}_s,\mathbf{e}_i)$ using

$$M(t,\mathbf{e}_s,\mathbf{e}_i) = -\frac{1}{8\pi^2}\frac{\partial}{\partial t}R(t,\mathbf{e}_s,\mathbf{e}_i), \tag{2.10}$$

where $R(\cdot)$ is the inverse Fourier transform of $A(\cdot)$. Here the work of Rose *et al.*[2] has been used to interpret the various quantities in the integral equation (2.8). The potential $V(\mathbf{x})$ is then recovered from the scattered field using the "miracle" equation

$$V(\mathbf{x}) = -2\mathbf{e}_i\cdot\nabla p_s(\mathbf{x},t = \mathbf{e}_i\cdot\mathbf{x}^+;\mathbf{e}_i). \tag{2.11}$$

Two comments are in order here. First, note the Hankel structure of the kernel in Eq. (2.8), which follows since $M(\cdot)$ is a function not of $\mathbf{x}$ and t separately, but only of the delay $t - \mathbf{e}_s \cdot \mathbf{x}$. This suggests that a fast algorithm solution of either the integral equation or the inverse problem itself is possible. Such an algorithm would take advantage of the structure represented by the Hankel kernel to reduce the order of the number of computations required. This is well established in the one-dimensional case (see Bruckstein *et al.*[10] and Yagle[16]). Second, note the redundancy involved in the use of the "miracle" equation (2.11). Newton[5] notes that the right side of Eq. (2.11) characterizes admissible scattering amplitudes: only a subset of all possible $A(k,\mathbf{e}_s,\mathbf{e}_i)$ (five independent variables) can result from all possible $V(\mathbf{x})$ (three independent variables). Even apart from issues of ill-posedness and overdetermination, it is clear that reconstructing the scattered field for each incident direction $\mathbf{e}_i$ involves a large amount of unnecessary computation for the purpose of reconstructing $V(\mathbf{x})$.

The reason all of this computation is necessary is made clear by Rose *et al.*[18] The derivations of Newton's integral equation procedure by Newton[5] and Rose *et al.*,[2] while mathematically rigorous, shed little insight into what is actually occurring during the inversion procedure. However, Rose *et al.*[18] show that Newton's Marchenko integral equation (2.8) is a direct consequence of the representation theorem

$$p(\mathbf{x},t) = \int_{\partial S} d^3\mathbf{x}' \int_{-\infty}^{\infty} dt' \left[p(\mathbf{x}',t') \frac{\partial G}{\partial n}(\mathbf{x},\mathbf{x}',t-t') - G(\mathbf{x},\mathbf{x}',t-t') \frac{\partial p}{\partial n}(\mathbf{x}',t') \right]. \qquad (2.12)$$

This result, which is a consequence of Green's theorem, shows that if a wave field $p(\mathbf{x},t)$ and its normal derivative $\partial p/\partial n$ are known on a closed, simply connected smooth surface ∂S, then the wave field in the interior of ∂S can be reconstructed if the Green's function $G(\mathbf{x},\mathbf{x}',t)$ is known on ∂S.

Insertion of Eq. (2.7) and the inverse Fourier transform of Eq. (2.2) into Eq. (2.12) and its time reversal yields the Marchenko integral equation (2.8) (see Rose *et al.*[18] for details). This shows that the excessive computation required by the solution of the integral equation (2.8) is a consequence of the implicit use of the representation theorem (2.12) and its time reversal [the latter accounts for the coupling between various $p_s(\cdot,\cdot;\mathbf{e}_i)$]. Reducing the amount of computation requires that another means of reconstructing the wave field be found. This is done in the next section.

III. A LAYER-STRIPPING RECONSTRUCTION PROCEDURE

The problem with using the representation theorem (2.12) to reconstruct the wave field is that this integral does not take advantage of the fact that the wave field arises from a scattering experiment. As the probing impulse $\delta(t - \mathbf{e}_i \cdot \mathbf{x})$ penetrates the inhomogeneous region where $V(\mathbf{x})$ differs from zero, it is possible to differentially reconstruct the scattered field. Of course, knowledge of $V(\mathbf{x})$ is necessary to accomplish this. However, it is not necessary to know $V(\mathbf{x})$ for *all* $\mathbf{x}$, but only for $\mathbf{x}$ in the region where the scattered field is being reconstructed: the wave front. And $V(\mathbf{x})$ can be obtained from the jump in the scattered field itself at the wave front.

For convenience, we choose coordinates (x,y,z) such that the direction of the probing impulsive plane wave is in the direction of increasing z, the inhomogeneous region lies in the half-space $z > 0$, and the plane wave passes through the origin at $t = 0$. Specification of the backscattered field $\hat{p}(x,y,z=0,k)$ and/or its inverse Fourier transform $p(x,y,z=0,t)$, together with a radiation condition for large $|\mathbf{x}|$ in the half-space $z > 0$, constitutes boundary conditions for the inverse potential problem for the Schrödinger equation (2.1) and the plasma wave equation (2.5). The experiment geometry is illustrated in Fig. 1.

The plasma wave equation (2.5) may be written as the coupled system

$$\left(\frac{\partial}{\partial z} + \frac{\partial}{\partial t}\right) p(x,y,z,t) \triangleq q(x,y,z,t), \qquad (3.1\text{a})$$

$$\left(\frac{\partial}{\partial z} - \frac{\partial}{\partial t}\right) q(x,y,z,t) = \left(V(x,y,z) - \frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}\right) p(x,y,z,t). \qquad (3.1\text{b})$$

From causality and the form of Eq. (3.1a), p and q have the forms

$$p = \delta(t-z) + \tilde{p}(x,y,z,t)1(t-z), \qquad (3.2\text{a})$$

$$q = \tilde{q}(x,y,z,t)1(t-z), \qquad (3.2\text{b})$$

where $\tilde{p}$ and $\tilde{q}$ are the smooth parts of p and q, respectively, and $1(\cdot)$ is the unit step or Heaviside function.

Inserting Eq. (3.2) into Eq. (3.1) yields

$$\left(\frac{\partial}{\partial z} + \frac{\partial}{\partial t}\right) \tilde{p}(x,y,z,t) = \tilde{q}(x,y,z,t), \qquad (3.3\text{a})$$

$$\left(\frac{\partial}{\partial z} - \frac{\partial}{\partial t}\right) \tilde{q}(x,y,z,t) = \left(V(x,y,z) - \frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}\right) \tilde{p}(x,y,z,t), \qquad (3.3\text{b})$$

$$V(x,y,z) = -2\tilde{q}(x,y,z,t=z^+), \qquad (3.3\text{c})$$

where equating the coefficients of $\delta(t-z)$ in Eq. (3.1b) has been used to obtain Eq. (3.3c). Equations (3.3) suggest a *recursive procedure for reconstructing* $V(\mathbf{x})$: Starting with known $\tilde{p}(x,y,0,t)$ and $\tilde{q}(x,y,0,t)$, Eq. (3.3) may be propagated recursively in z, yielding $V(\mathbf{x})$ recursively in z as the algorithm progresses.

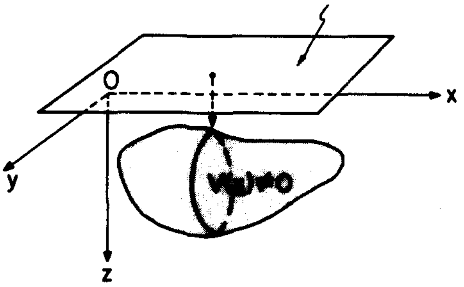

FIG. 1. Setup of the inverse scattering problem, which is solved by a layer-stripping algorithm.

The reconstruction of $V(\mathbf{x})$ takes place along the wave front, with $V(\mathbf{x})$ being obtained from the jump $\tilde{q}(x,y,z,t=z^+)$ in the wave field at the wave front. In this way both the wave field $\tilde{p}$ and potential $V(\mathbf{x})$ are reconstructed recursively and simultaneously as the impulsive probing plane wave passes through the inhomogeneous region. This is in contrast to Newton's procedure, described in Sec. II, in which the entire wave field is reconstructed by solving the Marchenko integral equation (2.8), and the potential $V(\mathbf{x})$ obtained in one big batch operation using the "miracle," Eq. (2.11). Note that the crucial physical principle allowing this simplification of the reconstruction procedure is *causality*, which is manifested in Eqs. (3.2).

The two second-order partial derivatives in Eq. (3.3b) will create high-frequency error problems when Eqs. (3.3) are implemented numerically. These partial derivatives may be eliminated by taking Fourier transforms of Eqs. (3.3) with respect to x and y, yielding

$$\left(\frac{\partial}{\partial z}+\frac{\partial}{\partial t}\right)\hat{\tilde{p}}(k_x,k_y,z,t)=\hat{\tilde{q}}(k_x,k_y,z,t), \tag{3.4a}$$

$$\left(\frac{\partial}{\partial z}-\frac{\partial}{\partial t}\right)\hat{\tilde{q}}(k_x,k_y,z,t)=(k_x^2+k_y^2)\hat{\tilde{p}}(k_x,k_y,z,t) + \hat{V}(k_x,k_y,z)**\hat{\tilde{p}}(k_x,k_y,z,t), \tag{3.4b}$$

$$\hat{V}(k_x,k_y,z)=-2\hat{\tilde{q}}(k_x,k_y,z,t=z^+), \tag{3.4c}$$

where the ** indicates a convolution operation in k_x and k_y. Equations (3.4) are more suitable for numerical implementation; on the other hand, Fourier transforms with respect to x and y must now be performed on the data, and inverse Fourier transforms performed on $\hat{V}(k_x,k_y,z)$. Note that the second-order partial derivatives with respect to x and y appearing in Eq. (3.3b) have now been replaced by the filter $k_x^2+k_y^2$. Since this filter becomes infinite for high wave-number values, to implement it, we should clip the high wave-number components, i.e., we should use $H_L(k_x,k_y) = k_x^2+k_y^2$ for $(k_x^2+k_y^2)^{1/2}<L$, and $H_L(k_x,k_y)=0$ otherwise. Here L is a parameter that determines the degree of smoothing that is applied to the reconstructed potential $V(x,y,z)$. The idea of using a clipped filter of this type was first proposed by Shepp and Logan to implement the filtered backprojection algorithm for the inverse Radon transform. The use of this filter may improve the stability of this algorithm relative to the stability of solving the mixed system (3.3). Note indeed that the system (3.3) corresponds to solving an initial value problem for a mixed partial differential equation, which is likely to be ill posed. By comparison, we expect that the introduction of the filter $H_L(k_x,k_y)$ in Eq. (3.4b) will have the effect of regularizing this problem, since $H_L(k_x,k_y)$ will smooth the variations of the potential $V(x,y,z)$ in the lateral directions x and y, when the medium is probed by a plane wave propagating in the z direction.

If k_x, k_y, z, and t are discretized to integer multiples of Δ, with the integers varying over the interval $[0,N]$, then a forward difference approximation to the partial derivatives in Eq. (3.4) yields

$$\hat{\tilde{p}}(z+\Delta,t+\Delta,k_x,k_y) =\hat{\tilde{p}}(z,t,k_x,k_y)+\hat{\tilde{q}}(z,t,k_x,k_y)\Delta, \tag{3.5a}$$

$$\begin{aligned}\hat{\tilde{q}}(z+\Delta,t-\Delta,k_x,k_y) &=\hat{\tilde{q}}(z,t,k_x,k_y)+(k_x^2+k_y^2)\hat{\tilde{p}}(z,t,k_x,k_y)\Delta\\ &\quad+\sum_m\sum_n V(k_x-m\Delta,k_y-n\Delta,z)\\ &\quad\times\hat{\tilde{p}}(z,t,m\Delta,m\Delta)\Delta,\end{aligned} \tag{3.5b}$$

$$\hat{V}(k_x,k_y,z+\Delta)=-2\hat{\tilde{q}}(k_x,k_y,z+\Delta,t=z+\Delta). \tag{3.5c}$$

The recursion patterns in z and t for $\hat{\tilde{p}}$ and $\hat{\tilde{q}}$ are illustrated in Figs. 2(a) and 2(b). We start off knowing $\hat{\tilde{p}}$ and $\hat{\tilde{q}}$ at z for all k_x, k_y, and t, and we wish to update them to $z+\Delta$ for all k_x, k_y, and t. Although the forms of the recursions may make it seem as though some information is being lost, recall that by causality $\hat{\tilde{p}}$ and $\hat{\tilde{q}}$ are both zero for $t<z$. Note that $O(N^5)$ multiplications-and-adds must be performed at each recursion, so that a total of $O(N^6)$ operations are necessary to reconstruct $V(\mathbf{x})$. However, solution of the discretized Marchenko integral equation (2.8) requires the inversion of an $N^3\times N^3$ matrix in t and $\mathbf{e}_i$ [$O(N^9)$ operations by Gaussian elimination] for each $\mathbf{x}$, for a total of $O(N^{12})$ operations! The numerically unstable gradient in Eq. (2.11) is also necessary for this method.

Equation (3.3c), which allows the recovery of $V(\mathbf{x})$ along the wave front from the jump in the scattered field there, is equivalent to the "fundamental identity"

$$V(\mathbf{x})=-\mathbf{e}_i\cdot\nabla B(\mathbf{x},\mathbf{e}_i) \tag{3.6}$$

of Rose *et al.*[2] and DeFacio and Rose.[7] In Eq. (3.6), which was first noted by Morawetz,[6] $B(\mathbf{x},\mathbf{e}_i)$ is the jump in the scattered field when the wave front passes through $\mathbf{x}$. Equation (3.6) is obtained by inserting the "progressing wave expansion"

$$p(\mathbf{x},t;\mathbf{e}_i)=\delta(t-e_i\cdot\mathbf{x})+B(\mathbf{x},\mathbf{e}_i)1(t-\mathbf{e}_i\cdot\mathbf{x}) + C(\mathbf{x},t;\mathbf{e}_i) \tag{3.7}$$

into the plasma wave equation (2.5). Here $C(\mathbf{x},t;\mathbf{e}_i)$ is smooth and zero for $t<\mathbf{e}_i\cdot\mathbf{x}$. Note that Eq. (3.6) is in turn equivalent to the "miracle," Eq. (2.11). However, it does not seem to have been recognized previously that Eq. (3.6) can be used not only to reconstruct the potential from the scattered field, but also to help and propagate the scattered field itself. The decomposition (3.1) of the plasma wave equation makes this possible by isolating $\tilde{q}(x,y,z,t)$, which is exactly the quantity needed to recover $V(\mathbf{x})$ by Eq. (3.3c).

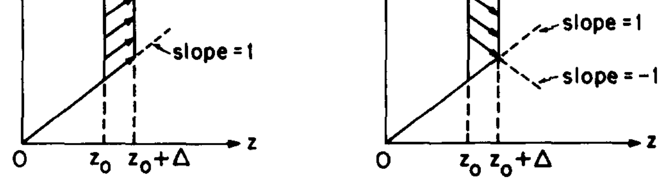

FIG. 2. (a) Recursion pattern for updating $\hat{\tilde{p}}(z,t,k_x,k_y)$. (b) Recursion pattern for updating $\hat{\tilde{q}}(z,t,k_x,k_y)$.

[Note from Eq. (3.7) that $(\partial/\partial t)\tilde{p}(x,y,z,t=z)=0$ in Eq. (3.3a).] The iterative methods of Morawetz[6] and DeFacio and Rose[7] used Eq. (3.6), but did not propagate the scattered field.

A decomposition of the Schrödinger equation into a coupled first-order system similar to Eq. (3.4) was performed by Wilcox[19] in the context of *invariant imbedding.* It should be evident that a layer-stripping procedure can also be interpreted as an invariant imbedding procedure. In fact, the one-dimensional layer-stripping algorithm obtained by Corones *et al.*[8] was derived from an invariant imbedding point of view. However, Wilcox's coupled first-order system[19] was only used to solve the*forward* problem of computing the wave field from a known potential. It is also more complex than Eq. (3.4), since it requires twice as many convolutions for a three-dimensional problem.

In the next section, the layer-stripping method is applied to an inverse seismic problem, illustrating some of its advantages over present methods of solving this problem.

IV. APPLICATION TO AN INVERSE SEISMIC PROBLEM

The inverse seismic problem considered in this section is that of reconstructing the density $\rho(x,z)$ and wave speed $c(x,z)$ of an acoustic medium from measurement of the response of the medium to a harmonic line source. We follow Coen *et al.*[1] in transforming this problem into a two-dimensional Schrödinger equation inverse scattering problem, which is then solved using the layer-stripping algorithm of Sec. III. By performing the experiment twice, at two different source frequencies ω_1 and ω_2, two different potentials $V(\mathbf{x};\omega_1)$ and $V(\mathbf{x};\omega_2)$ are reconstructed, and the density and wave speed are in turn recovered from the two potentials. The one-dimensional version of this procedure was given in Yagle and Levy.[15]

The use of a layer-stripping algorithm implies that the amount of computation required to reconstruct $\rho(x,z)$ and $c(x,z)$ will be less than that required by the integral equation procedure of Coen *et al.*[1] More importantly, only surface measurements in the near field are required to initialize the algorithm. This is in contrast to the procedure of Coen *et al.*,[1] which requires *transmission* data (generally not available for an inverse seismic problem) and a transformation from near-field data to far-field data.

The specifics of the inverse problem are as follows. An inhomogeneous acoustic medium characterized by smooth density $\rho(x,z)$ and wave speed $c(x,z)$ functions is contained in the half-space $z>0$ and bounded by a free (pressure-release) surface at $z=0$. This medium is probed with cylindrical harmonic waves from a harmonic line source extending along the x axis. The strength of this source varies as $\rho(x,z=0)^{1/2}$; see Eq. (4.7a). The vertical acceleration response $\hat{a}_z(x,y,z=0)$ of the medium at the free surface in the sinusoidal steady state is measured for all x and y. The situation is illustrated in Fig. 3. It is assumed that the medium is homogeneous in the y direction [i.e., $\rho=\rho(x,z)$ and $c=c(x,z)$], and that the inhomogeneous region has compact support (i.e., $\rho=\rho_0$ and $c=c_0$ for sufficiently large $|x|$ and z). It is further assumed that ρ_0, c_0, and $\rho(x,z=0)$ are known, and that $\partial\rho/\partial z(x,z=0)=0$. In addition, it is assumed that there are no bound states, a sufficient condition for this is $c_0<c(x,z)$ (see Coen *et al.*[1]).

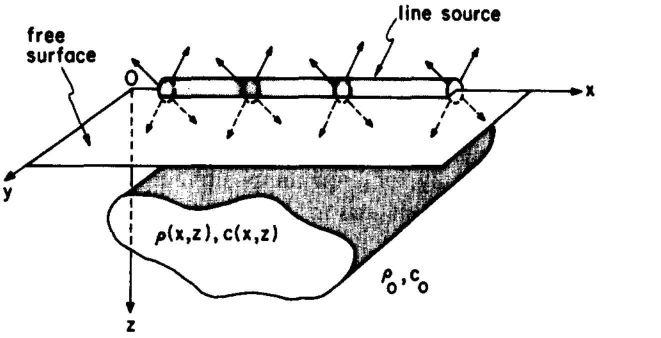

FIG. 3. Setup for the inverse seismic problem, which is solved by a layer-stripping algorithm.

The basic linear equations that describe an acoustic medium are

$$\frac{\partial^2 p}{\partial t^2}=-\rho c^2\nabla\cdot\mathbf{a}, \qquad (4.1a)$$

$$\nabla p=-\rho\mathbf{a}, \qquad (4.1b)$$

where p is the pressure and $\mathbf{a}$ the medium acceleration.

Taking Fourier transforms of Eqs. (4.1) with respect to time and defining

$$\hat{\psi}(\mathbf{x};\omega)=\hat{p}(\mathbf{x},\omega)/\rho(\mathbf{x})^{1/2}, \qquad (4.2)$$

results in the equation

$$(\nabla^2+\omega^2/c_0^2-V(\mathbf{x};\omega))\hat{\psi}(\mathbf{x};\omega)=0, \qquad (4.3)$$

where the potential $V(\mathbf{x};\omega)$ has compact support, is independent of y, and is given by

$$V(\mathbf{x};\omega)=(\omega^2/c_0^2)(1-c_0^2/c(\mathbf{x})^2)+\rho(\mathbf{x})^{1/2}\nabla^2(\rho(\mathbf{x})^{-1/2}). \qquad (4.4)$$

Following Coen *et al.*[1] and Yagle and Levy,[15] the Fourier transform of Eq. (4.3) with respect to y is taken, resulting in the Schrödinger equation

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial z^2}+k^2-V(x,z;\omega)\right)\hat{\hat{\psi}}(x,z,k;\omega)=0, \qquad (4.5)$$

where

$$k^2=\omega^2/c_0^2-k_y^2=k_x^2+k_z^2 \qquad (4.6)$$

is the sum of the squares of the lateral wave number k_x and vertical wave number k_z. The boundary conditions for the Schrödinger equation (4.5) are obtained from

$$\hat{p}(x,y,z=0;\omega)=\rho(x,z=0)^{1/2}\delta(y), \qquad (4.7a)$$

$$\hat{\psi}(x,y,z=0;\omega)=\delta(y), \qquad (4.7b)$$

$$\frac{\partial\hat{p}}{\partial z}=-\rho\hat{a}_z, \qquad (4.7c)$$

which lead to

$$\hat{\hat{\psi}}(x,z=0,k;\omega)=1, \qquad (4.8a)$$

$$\frac{\partial\hat{\hat{\psi}}}{\partial z}=-\rho(x,z=0)^{1/2}\hat{\hat{a}}_z(x,z=0,k;\omega). \qquad (4.8b)$$

Note that it has been assumed that the strength of the harmonic line source varies as $\rho(x,z=0)^{1/2}$, and

$(\partial/\partial z)\rho(x,z=0)=0$. These assumptions simplify the form of the algorithm, although they are not essential. In addition, a radiation condition is imposed on ψ and $\hat{a}_z$ for large $|x|$, $|y|$, and z.

Note that $\hat{a}_z(x,z=0,k;\omega)$ may be obtained from $\hat{a}_z(x,y,z=0;\omega)$ only for k on the positive imaginary axis and on the positive real axis in the interval $[0, \omega/c_0]$; this region is illustrated in Fig. 4. Numerical results in Yagle and Levy[15] and Stickler[20] for the one-dimensional problem suggest that the absence of data for $k>\omega/c_0$ should have little effect on the quality of the reconstructed medium parameters. This is due to the fact that data for *all* k are used to determine these parameters [see Eq. (4.15)], so the lack of data for large k will merely result in a loss of resolution. If this is acceptable, then $\hat{a}_z(x,z=0,k;\omega)$ for $k>\omega/c_0$ may be set equal to zero in the sequel. An alternative is to analytically continue $\hat{a}_z(x,z=0,k;\omega)$ for $k>\omega/c_0$ using a theorem of Van Winter[21] that employs the Mellin transform. This was noted by Coen *et al.*,[1] Yagle and Levy,[15] and Stickler[20].

The layer stripping algorithm for solving this problem is as follows. An inverse Fourier transform of Eq. (4.5) that takes k into the fictitious depth coordinates ζ [recall k is a wave number; see Eq. (4.6)] results in the plasma wave equation

$$\left(\frac{\partial^2}{\partial z^2}+\frac{\partial^2}{\partial x^2}-\frac{\partial^2}{\partial \zeta^2}-V(x,z;\omega)\right)\psi(x,z,\zeta;\omega)=0, \tag{4.9}$$

which can be written as the coupled first-order system

$$\left(\frac{\partial}{\partial z}+\frac{\partial}{\partial \zeta}\right)\psi(x,z,\zeta;\omega)=\phi(x,z,\zeta;\omega), \tag{4.10a}$$

$$\left(\frac{\partial}{\partial z}-\frac{\partial}{\partial \zeta}\right)\phi(x,z,\zeta;\omega)=\left(V(x,z;\omega)-\frac{\partial^2}{\partial x^2}\right)\psi(x,z,\zeta;\omega), \tag{4.10b}$$

with the initial conditions [from Eq. (4.8)]

$$\psi(x,z=0,\zeta;\omega)=\delta(\zeta), \tag{4.11a}$$

$$\phi(x,z=0,\zeta;\omega)=-\rho(x,z=0)^{1/2}a_z(x,z=0,\zeta;\omega). \tag{4.11b}$$

The form of Eqs. (4.10) makes it clear that the impulse in Eq. (4.11a) will propagate in z and ζ, so that ψ and ϕ have the forms

$$\psi(x,z,\zeta;\omega)=\delta(\zeta-z)+\tilde{\psi}(x,z,\zeta;\omega)1(\zeta-z), \tag{4.12a}$$

$$\phi(x,z,\zeta;\omega)=\tilde{\phi}(x,z,\zeta;\omega)1(\zeta-z). \tag{4.12b}$$

Insertion of Eqs. (4.12) in Eqs. (4.10) finally results in the layer-stripping algorithm

$$\left(\frac{\partial}{\partial z}+\frac{\partial}{\partial \zeta}\right)\tilde{\psi}(x,z,\zeta;\omega)=\tilde{\phi}(x,z,\zeta;\omega), \tag{4.13a}$$

$$\left(\frac{\partial}{\partial z}-\frac{\partial}{\partial \zeta}\right)\tilde{\phi}(x,z,\zeta;\omega)=\left(V(x,z;\omega)-\frac{\partial^2}{\partial x^2}\right)\tilde{\psi}(x,z,\zeta;\omega), \tag{4.13b}$$

$$V(x,z;\omega)=-2\tilde{\phi}(x,z,\zeta=z^+;\omega). \tag{4.13c}$$

The coupled set of equations (4.23) can be downward continued in z, as in the method of Sec. III. The algorithm is initialized using

$$\tilde{\psi}(x,z=0,\zeta;\omega)=0, \tag{4.14a}$$

$$\begin{aligned}\tilde{\phi}(x,z=0,\zeta;\omega)&=-\rho(x,z=0)^{1/2}a_z(x,z=0,\zeta;\omega)\\&=\rho(x,z=0)^{1/2}\\&\quad\times F_k^{-1}[\hat{a}_z(x,z=0,k_y^2=\omega^2/c_0^2-k;\omega)].\end{aligned} \tag{4.14b}$$

Note in particular that $V(x,z;\omega)$ is obtained using

$$\begin{aligned}V(x,z;\omega)&=-2\tilde{\phi}(x,z,\zeta=z^+;\omega)\\&=\frac{\rho(x,z)^{1/2}}{\pi}\int_{-\infty}^{\infty}\hat{a}_z(x,z,k;\omega)e^{jkz}\,dk,\end{aligned} \tag{4.15}$$

so that data for all k are used to determine V. This is more stable numerically than obtaining V exclusively from high-k data using the initial value theorem. This is especially important here, since data for large k can only be obtained from the data $\hat{a}_z(x,y,z=0;\omega)$ by analytic continuation, which may be quite unstable numerically. To avoid this, we may set $\hat{a}_z(x,z,k;\omega)$ equal to zero for $k>\omega/c_0$, as discussed earlier; this will result only in a loss of resolution in reconstucting $V(x,z;\omega)$.

After performing the experiment twice, using two different source frequencies ω_1 and ω_2, and reconstructing the potentials $V(x,z;\omega_1)$ and $V(x,z;\omega_2)$, $\rho(x,z)$ and $c(x,z)$ are recovered as follows. We have from Eq. (4.4),

$$\frac{1}{c(x,z)^2}=\frac{1}{c_0^2}-\frac{V(x,z;\omega_1)-V(x,z;\omega_2)}{(\omega_1^2-\omega_2^2)}, \tag{4.16a}$$

$$\begin{aligned}&\rho(x,z)^{1/2}\nabla^2(\rho(x,z)^{-1/2})\\&\quad=(\omega_2^2V(x,z;\omega_1)-\omega_1^2V(x,z;\omega_2))/(\omega_2^2-\omega_1^2),\end{aligned} \tag{4.16b}$$

and Eq. (4.16b) can be solved, since $\rho(x,z)=\rho_0$ for large $|x|$ and z, and $\rho(x,z=0)$ is known.

The algorithm given by Eqs. (4.13) is a generalization of the algorithm for the one-dimensional inverse acoustic problem for a point harmonic source, which was given by Yagle and Levy.[15] In that paper ζ was interpreted as a fictitious depth coordinate, along which *image source* distributions are being computed in order to synthesize the response of the medium below depth z. The causality of ψ and ϕ is then a consequence of the principle that an image source never lies in that part of the medium in which the field is to be synthesized. The first nonzero value of $\phi(x,z,\zeta;\omega)$ synthesizes the primary reflection at the point (x,z), which is $V(x,z;\omega)$. Note that for each x_0, the first nonzero value of $\phi(x_0,z,\zeta;\omega)$ [which is $\phi(x_0,z,\zeta=z^+;\omega)$] can only depend on $V(x_0,z;\omega)$; other values $V(x,z;\omega)$ cannot affect $\phi(x_0,z,\zeta=z^+;\omega)$ by causality in ζ.

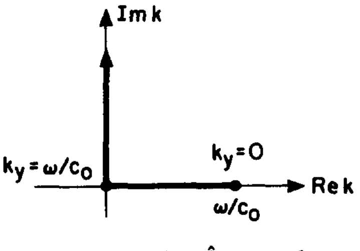

FIG. 4. Regions where $\hat{a}(x,z=0,k)$ may be computed from the data.

An alternative to probing the medium with harmonic, single-frequency waves is to probe it with an impulsive (all frequencies) plane wave. This requires a different mathematical formulation from the present problem. Probing the medium twice, at two different angles of incidence, would allow the recovery of $\rho(x,z)$ and $c(x,z)$ separately. This is a higher-dimensional version of the problem solved by Coen[22] using integral equations, and by Yagle and Levy[13] using a layer-stripping algorithm. However, the lateral variation of $c(x,z)$ implies that the wave front for this problem [which is in (x,z,t) rather than (x,z,ζ)] will not be planar. Since, in a layer-stripping algorithm, the reconstruction of the medium parameters takes place along the wave front, it is necessary to make a transformation to wave-front centered (or ray centered) coordinates, run the algorithm using these coordinates, and then transform the reconstructed parameters back into Cartesian (x,z) coordinates. An algorithm for solving this inverse seismic problem using this procedure is given in Yagle[16]; the numerical performance of this algorithm is unknown.

Rose *et al.*[18] point out that the representation theorem could be used to reconstruct the wave field, and the location of the characteristic surfaces (wave fronts) thus inferred from the various positions of the probing impulse. The eikonal equation could then be used to recover $c(\mathbf{x})$. However, this method uses far-field data and transmitted data, whereas in an actual inverse seismic problem the data are obtained in the near field and transmitted data are not available.

Note that the problem of nonplanar wave fronts does not arise for the plasma wave equation (2.5), since the wave speed is implicitly unity. For the inverse problem with a harmonic source, the wave speed in (z,ζ) space is also implicitly unity. But for time-domain inverse seismic problems, the lateral variation of the wave speed $c(x,z)$ will lead to wave front definition problems, and other problems such as caustics (see Yagle[16] and Rose *et al.*[18]).

V. BORN APPROXIMATION INVERSION METHODS

The coupling of the first-order system (3.3) is necessary to account for multiple scattering events. In this section the effect of neglecting this coupling is compared to several Born approximation inversion techniques.

The (first) Born approximation applied to a Schrödinger equation scattering problem is as follows. In the Lippman–Schwinger equation (2.2) the field $\hat{p}(\mathbf{y},k;\mathbf{e}_i)$ inside the inhomogeneous region is replaced by the incident field. Then Eq. (2.2) becomes

$$\hat{p}^{\mathrm{B}}(\mathbf{x},k;\mathbf{e}_i)=e^{-ik\mathbf{e}_i\cdot\mathbf{x}}-\int(4\pi|\mathbf{x}-\mathbf{y}|)^{-1}\times e^{-ik|\mathbf{x}-\mathbf{y}|}V(\mathbf{y})e^{ik\mathbf{e}_i\cdot\mathbf{y}}\,d^3\mathbf{y},\qquad(5.1)$$

where the superscript B indicates that the Born approximation has been used. Note that the only unknown in Eq. (5.1) is the potential $V(\mathbf{x})$.

The Born approximation is often described as a "weak scattering" approximation. While this is true, it conceals the main point of Eq. (5.1), which is that the scattered field is arising solely from interactions of the incident probing plane wave with the potential. In other words, the Born approximation is a *single scattering approximation*: multiple scattering events are neglected.

To illustrate this explicitly, consider the one-dimensional Schrödinger inverse scattering problem. The Lippman–Schwinger equation for this problem is

$$\hat{p}(z,k)=e^{-ikz}-\int\frac{1}{ik}e^{-ik|z-z'|}V(z')\hat{p}(z',k)dz',\qquad(5.2)$$

which in the Born approximation becomes

$$\hat{p}^{\mathrm{B}}(z,k)=e^{-ikz}-\int\frac{1}{ik}e^{-ik|z-z'|}V(z')e^{-ikz'}\,dz'.\qquad(5.3)$$

If backscattered data (i.e., data for negative values of z) are being measured, the equation for the scattered field $\hat{p}_s^B(z,k)$ is

$$p_s^{\mathrm{B}}(z,k)=-e^{ikz}\frac{1}{ik}\int V(z')e^{-2ikz'}\,dz'.\qquad(5.4)$$

Taking the partial derivative of Eq. (5.4) with respect to z, and following with an inverse Fourier transform with respect to k yields, at $z=0$,

$$\frac{\partial}{\partial z}p_s^{\mathrm{B}}(0,t)=-\int V(z')\delta(t-2z')dz'=-\frac{1}{2}V\left(\frac{t}{2}\right).\qquad(5.5)$$

Equation (5.5) depicts that the Born approximation is really doing: *imaging* the potential profile $V(z)$ from the response $\partial p_s^{\mathrm{B}}/\partial z$, which arises from the jump in the scattered field at z. The probing wave requires time $t/2$ to reach location $z=t/2$ and the response $\partial p_s^{\mathrm{B}}/\partial z$ caused by the jump in the field at $z=t/2$ [due to $V(z=t/2)$] requires time $t/2$ to make it back to the surface $z=0$. Hence examining $(\partial/\partial z)p_s^{\mathrm{B}}(0,t)$ images $V(t/2)$. Conversely, every nonzero value of $(\partial/\partial z)p_s^{\mathrm{B}}(0,t)$ is interpreted as a nonzero value of $V(t/2)$, i.e., multiple scattering events are being neglected. Note also that Eq. (5.5) is the fundamental identity with the scattered field back-propagated to $z=0$ without any additional scattering (i.e., using the Born approximation).

Two other comments should be made. First, iterative substitution of $\hat{p}(z,k)$ in the Lippman–Schwinger equation (5.2) will yield contributions to the scattered field due to multiple scattering events. The resulting Neumann series will converge if $V(z)$ is small (see Simon[23]). The nth term of this series accounts for scattering events of order n, making it the Fourier transform of the Bremmer series. Second, the inhomogeneous variable-velocity wave equation can be treated as a Schrödinger equation with potential $k^2(c_0^{-2}-c(z)^{-2})$. For this problem, an inverse Fourier transform of Eq. (5.4) with respect to k yields at $z=0$.

$$c_0^{-2}-c(z')^{-2}=2\int_0^{2z'}p_s^{\mathrm{B}}(0,t')dt',\qquad(5.6)$$

where t has been replaced by $2z'$ for clarity.

These ideas all generalize to higher dimensions. For example, if the Born approximation is applied to the definition of the scattering amplitude Eq. (2.4), the result is

$$A^{\mathrm{B}}(k,\mathbf{e}_x\mathbf{e}_i) = -\int V(\mathbf{y})e^{ik(\mathbf{e}_s-\mathbf{e}_i)\cdot\mathbf{y}}\,d^3\mathbf{y}, \tag{5.7}$$

and an inverse Fourier transform with respect to k yields

$$\begin{aligned} R\,[V(\mathbf{x})] &\triangleq \int V(\mathbf{y})\delta(t+(\mathbf{e}_s-\mathbf{e}_i)\cdot y)d^3\mathbf{y} \\ &= -R^{\mathrm{B}}(t,\mathbf{e}_s,\mathbf{e}_i), \end{aligned} \tag{5.8}$$

where the left side of Eq. (5.8) is the *Radon transform* of $V(\mathbf{y})$ (see Deans[24] for a discussion of the Radon transform and inversion techniques for it). In particular, for backscattered data ($\mathbf{e}_s = -\mathbf{e}_i$) we have [cf. Eq. (5.5)]

$$\int V(\mathbf{y})\delta(t-2\mathbf{e}_i\cdot\mathbf{y})d^3\mathbf{y} = -R^{\mathrm{B}}(t,-\mathbf{e}_i,\mathbf{e}_i), \tag{5.9}$$

while for transmission data ($\mathbf{e}_s = \mathbf{e}_i$), we have, from Eq. (5.7),

$$\int V(\mathbf{y})d^3\mathbf{y} = -A^{\mathrm{B}}(k,\mathbf{e}_i,\mathbf{e}_i). \tag{5.10}$$

These equations are all generalizations of the one-dimensional results, and illustrate that the Born approximation is a single scattering approximation.

Following Rose *et al.*,[2] we also have from Eq. (5.7) that

$$V(\mathbf{x}) = -F_{\mathbf{q}}^{-1}[A^{\mathrm{B}}(k,\mathbf{e}_s,\mathbf{e}_i)], \tag{5.11}$$

where the inverse Fourier transform is taken with respect to

$$\mathbf{q} = k(\mathbf{e}_s-\mathbf{e}_i). \tag{5.12}$$

Equations (5.8) and (5.11) are related by the well-known relation (Deans,[24] p. 97)

$$F_t R_{(\mathbf{e}_s-\mathbf{e}_i)}[V(\mathbf{x})] = F_x F_y[V(\mathbf{x})]\big|_{(k_x,k_y)=\mathbf{q}=k(\mathbf{e}_s-\mathbf{e}_i)}. \tag{5.13}$$

Having discussed the Born (single scattering) approximation, we now consider the effect of a single scattering approximation for the layer-stripping procedure (3.3). The purpose of the coupled system (3.3a) and (3.3b) is to account for all multiple scattering events. Thus the single scattering approximation to the procedure is simply Eq. (3.3c). Note that the fundamental identity, Eq. (3.3c), yields $V(\mathbf{x})$ without requiring an inverse Radon transform. But when the jump in the scattered field $q(\mathbf{x},t)$ is propagated from the wave front back to the receiving surface on which data are taken, the scattered fields from various points on the wave front interfere with each other, and an inverse Radon transform becomes necessary to sort out the various contributions.

To illustrate this point, we now consider two different Born approximation inversion methods. The first uses far-field data, in the form of the scattering amplitude, and consists of solving Eq. (5.8) for $V(\mathbf{x})$. The other, due to Rose *et al.*,[2] uses near-field data, and also requires an inverse Radon transform. We now show that both of these Born approximation methods are in fact direct consequences of the fundamental identity

$$V(\mathbf{x}) = -2\tilde{q}(\mathbf{x},t=\mathbf{e}_i\cdot\mathbf{x}), \tag{5.14}$$

which is the single scattering approximation to the procedure (3.3). It should be noted here that since both methods rely exclusively on high-k data, the Born (weak scattering) approximation becomes exact.

To see this explicitly for the far-field data Born approximation method we note that when the Born approximation is used inside the Marchenko integral equation (2.8), the first term on the right-hand side of (2.8) disappears, and we obtain

$$p_s^{\mathrm{B}}(\mathbf{x},t;\mathbf{e}_i) = -\frac{1}{8\pi^2}\int \frac{d}{dt}R^{\mathrm{B}}(t-\mathbf{e}_s\cdot\mathbf{x},\mathbf{e}_s,\mathbf{e}_i)d^2\mathbf{e}_s. \tag{5.15}$$

This leads to (recall $z=\mathbf{e}_i\cdot\mathbf{x}$)

$$\begin{aligned} q^{\mathrm{B}}(\mathbf{x},t;\mathbf{e}_i) &= \left(\frac{\partial}{\partial z}+\frac{\partial}{\partial t}\right)p_s^B(\mathbf{x},t;\mathbf{e}_i) \\ &= -\frac{1}{8\pi^2}\int \frac{d^2}{dt^2}R^{\mathrm{B}}(t-\mathbf{e}_s\cdot\mathbf{x},\mathbf{e}_s,\mathbf{e}_i) \\ &\quad\cdot(1-\mathbf{e}_s\cdot\mathbf{e}_i)d^2\mathbf{e}_s, \end{aligned} \tag{5.16}$$

and substituting $t=\mathbf{e}_i\cdot\mathbf{x}$ and changing the variable of integration to $\delta\mathbf{e}=\mathbf{e}_i-\mathbf{e}_s$ results in

$$\begin{aligned} &2q^{\mathrm{B}}(\mathbf{x},t=\mathbf{e}_i\cdot\mathbf{x};\mathbf{e}_i) \\ &\quad= \frac{1}{8\pi^2}\int \frac{d^2}{dt^2}R^{\mathrm{B}}(t=\delta\mathbf{e}\cdot\mathbf{x})d^2(\delta\mathbf{e}), \end{aligned} \tag{5.17}$$

since the Jacobian of the transformation is $|\mathbf{e}_i-\mathbf{e}_s|^2 = 2(1-\mathbf{e}_i\cdot\mathbf{e}_s)$. But the right-hand side of Eq. (5.17) is an inverse Radon transform (see Deans,[24] p. 111), and from Eq. (5.8) we see that it is $-V(\mathbf{x})$. The left-hand side of Eq. (5.17) can be identified to $2\tilde{q}(\mathbf{x},t=\mathbf{e}_i\cdot\mathbf{x};\mathbf{e}_i)$ since in high-frequency limit the Born approximation is exact, and Eq. (5.4) shows that it is equal to $-V(\mathbf{x})$. This links the two approaches together, and proves that they are equivalent.

The near-field data Born approximation method considered next was given by Rose *et al.*[2] A simple derivation of their result is now given. Let $B(\mathbf{x})$ be the jump in the scattered field when the wave front passes through $\mathbf{x}$, so that

$$B(\mathbf{x}) = p_s(\mathbf{x},t=\mathbf{e}_i\cdot\mathbf{x}^+;\mathbf{e}_i). \tag{5.18}$$

Here $B(\mathbf{x})$ is measured over the surface of a sphere that contains the support of $V(\mathbf{x})$ (this is the near-field data). Let S be the disk that is the intersection of this sphere with the plane $\mathbf{x}\cdot\mathbf{e}^\perp = h$, where $\mathbf{e}^\perp \perp \mathbf{e}_i$, and h varies, and let $\mathbf{e}^* = \mathbf{e}^\perp\times\mathbf{e}_i$. The situation is illustrated in Fig. 5. We then have, using a well-known identity,

$$\int_{\partial S} B\,d\mathbf{r} = \int_S d\,\mathbf{S}\times\nabla B, \tag{5.19}$$

and taking the dot product with $-2e^*$ gives

$$\begin{aligned} -2\int_{\partial S} B\mathbf{e}^*\cdot d\mathbf{r} &= -2\int_S \mathbf{e}^*\cdot(d\,\mathbf{S}\times\nabla B) \\ &= -2\int_S \nabla B\cdot(\mathbf{e}^*\times d\,\mathbf{S}) \\ &= \int_S (-2\nabla B\cdot\mathbf{e}_i)dS \\ &= \int_S V(\mathbf{x})dS = R\,[V(\mathbf{x})], \end{aligned} \tag{5.20}$$

where $d\,\mathbf{S} = dS\,\mathbf{e}^\perp$ is a differential surface area in the direction $\mathbf{e}^\perp$, and the "miracle" equation (2.11) has been used. By letting $\mathbf{e}_i$ vary over a half-plane, enough information is gained to invert $R[V(\mathbf{x})]$. Note that the Born approximation is implicitly used in the assumption that all nonzero values of B arise from a nonzero value of $V(\mathbf{x})$. However, since this method uses *only* data for large k, multiple scattering events ae negligible and the method is exact.

To interpret this result, define the positive x, y, and z axes to be in the direction $\mathbf{e}^{\perp}$, $\mathbf{e}^*$, and $\mathbf{e}_i$, respectively [see Fig. 5(b)]. The probing plane wave passes through the origin at $t = 0$, as before. Now, note that $\mathbf{e}^* \cdot d\mathbf{r}$ in Eq. (5.20) is the projection of $d\mathbf{r}$ on a line L parallel to the y axis. Hence, the left side of Eq. (5.20) can be interpreted as minus twice the integral of $B(\mathbf{x})$ (the jump in the scattered field at $\mathbf{x}$ as the wave front passes) over the line L if L lies in the homogeneous region beyond (i.e., for large z) the support of $V(\mathbf{x})$. To see this, note that from the fundamental identity Eq. (3.6) that $B(\mathbf{x})$ does not vary in the z direction in the homogeneous region. Therefore, each point $\mathbf{x}$ on ∂S^+, the far side of ∂S, can be projected to an image point $\mathbf{x}^*$ on L, as shown in Fig. 5(b), and we will have $B(\mathbf{x}) = B(\mathbf{x}^*)$. We also have that $B(\mathbf{x})$ on the near side of ∂S is zero.

This same result follows directly if the fundamental identity Eq. (3.6) is integrated over S. The result is

$$\int_S V(\mathbf{x})dy\,dz = -2\int_S \frac{\partial B}{\partial z}dy\,dz$$

$$= -2\int_{\partial S^+} B(\mathbf{x})dy = -2\int_L B(\mathbf{x})dy, \qquad (5.21)$$

which, by the interpretation given above, is the same as the left side of Eq. (5.20). Thus the near-field data Born inversion method is also related to the single scattering approximation to the layer-stripping procedure (viz., the fundamental identity). The exact nature of this relation depends on how $B(\mathbf{x})$ is computed. If $B(\mathbf{x})$ is computed using exclusively large-k data (by means of the initial value theorem), then multiple scattering events are negligible and both methods are exact. However, if data at *all* values of k are to be used to compute $B(\mathbf{x})$ (by means of an inverse Fourier transform), then the near-field Born method cannot be so used, and the layer-stripping procedure single scattering approximation is in fact only approximate.

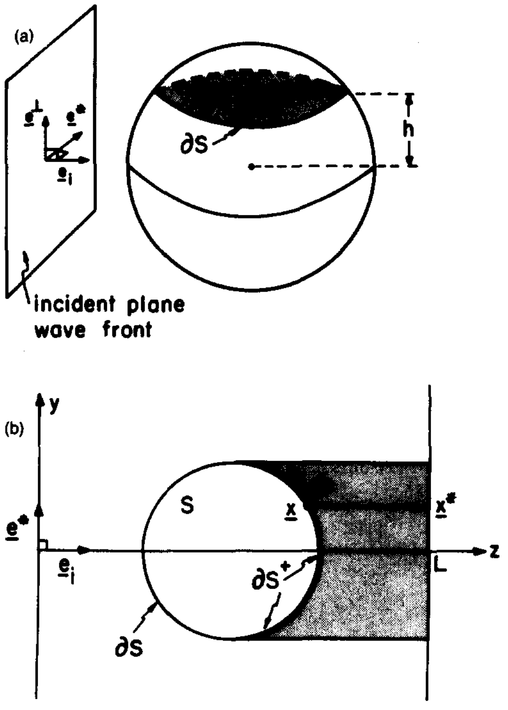

FIG. 5. (a) Setup for the near-field, Born, approximation inverse scattering procedure (side view). (b) Setup for the near-field, Born-approximation inverse scattering procedure (top view).

As a final note, if source–receiver configurations different from those considered so far in this paper are employed, then Born approximation inversion techniques may involve projections other than the Radon transform. For example, if an impulsive *point source* is used to probe the inhomogeneous region, it might be expected that the response at the point source at time $2t$ might be the *spherical mean* of $V(\mathbf{x})$ over a sphere of radius t centered on the point source. This is indeed the case, as is now demonstrated, following Norton and Linzer.[25] Application of the Born approximation to the Lippman–Schwinger equation for this problem results in

$$\hat{p}_s^B(\mathbf{x},k) = \int \frac{e^{-ik|\mathbf{x}_0-\mathbf{y}|}}{|\mathbf{x}_0-\mathbf{y}|} V(\mathbf{y}) \frac{e^{-ik|\mathbf{y}-\mathbf{x}_0|}}{|\mathbf{y}-\mathbf{x}_0|} d^3\mathbf{y}$$

$$= \int V(\mathbf{y}) \frac{e^{-2ik|\mathbf{y}-\mathbf{x}_0|}}{|\mathbf{y}-\mathbf{x}_0|^2} d^3\mathbf{y}, \qquad (5.22)$$

where the point source and receiver are located at $\mathbf{x}_0$. An inverse Fourier transform with respect to $2k$ results in

$$2p_s^B(\mathbf{x},2t) = \int V(\mathbf{y}) \frac{\delta(t-|\mathbf{y}-\mathbf{x}_0|)}{|\mathbf{y}-\mathbf{x}_0|^2} d^3\mathbf{y}$$

$$= \frac{1}{t^2}\int V(\mathbf{y})\delta(t-|\mathbf{y}-\mathbf{x}_0|)d^3\mathbf{y}, \qquad (5.23)$$

which is the spherical mean of $V(\mathbf{x})$ over a sphere of radius t centered on $\mathbf{x}_0$, as expected. Norton and Linzer[25] and Fawcett[26] have obtained analytic inversion formulas for Eq. (5.23). Cohen and Bleistein[3] obtained an equation similar to Eq. (5.23) for the inhomogeneous wave equation in two dimensions, and they also derived an analytic inversion formula.

VI. CONCLUSION

A fast algorithm for solving Schrödinger equation inverse scattering problems has been obtained by utilizing the concept of layer stripping. This procedure differentially reconstructs both the scattered wave field and the potential, with the potential being obtained from the jump in the scattered field at the wave front. Thus the potential is differentially reconstructed along the wave front as the probing plane wave penetrates the inhomogeneous region. The layer-stripping procedure is "fast" in that it requires fewer computations to reconstruct the potential than the integral equation method of Newton. This results from the fact that the procedure takes full advantage of the inherent, causality-induced structure of the inverse scattering problem.

The fact that near-field, backscattered data are used to initiate the procedure makes it more suitable for application to inverse seismic problems than Newton's integral equation method, which requires far-field data measured in all direction for incident waves from all directions. This was illustrated and discussed in Sec. IV. This choice of data also avoids the problems of overdetermination and ill-posedness that arise in Newton's method, although numerical instability might be a problem. Of course, data for several directions of incidence could be used, and a three-dimensional least-squares fit applied to the resulting potentials as they are computed. This should reduce the effect of noise on the reconstructed potentials. The numerical performance of this algorithm, and the effects of noisy and band-limited data on its operation is an important topic that requires more research.

The coupling in the layer-stripping algorithm (3.3) accounts for multiple scattering events. Neglecting this coupling left Eq. (3.3c), which is also the fundamental identity Eq. (2.11). This equation yields the potential directly from the jump in the scattered field at the wave front. Propagation of this field to the surfaces on which data are collected leads to complications that are manifested in Born approximation inversion methods by the appearance of the Radon transform. Two different Born approximation inversion methods were derived and interpreted, and were then related to the wave field induced by the fundamental identity. This illustrated the single scattering assumptions inherent in the uncoupled layer stripping procedure and in the Born inversion procedures themselves. These latter procedures are exact only if high-energy data are used exclusively.

It is important to note that the layer-stripping algorithm is in principle *exact*, since it is equivalent to the Schrödinger equation itself. In addition, it uses data of all energies, instead of relying only on high-energy data as does the Born approximation. It is in fact a differential method that is dual to Newton's integral equation method, but it is considerably more efficient, since it avoids solving infinitely many three-dimensional integral equations.

ACKNOWLEDGMENT

The authors wish to thank the reviewer for many helpful comments, and for identifying several issues requiring clearer exposition in the text.

The work of A.E.Y. was suported by the Exxon Education Foundation.

The work of B.C.L. was supported by the Army Research Office under Grant No. DAAG-29-84-K-0555, and by the National Science Foundation under Grant No. ECS-83-12921.

[1] S. Coen, M. Cheney, and A. Weglein, "Velocity and density of a two-dimensional acoustic medium from point source surface data," J. Math. Phys. **25**, 1857 (1984).

[2] J. H. Rose, M. Cheney, and B. DeFacio, "The connection between time- and frequency-domain three-dimensional inverse scattering methods," J. Math. Phys. **25**, 2995 (1984).

[3] J. K. Cohen and N. Bleistein, "Velocity inversion procedure for acoustic waves," Geophys. **44**, 1077 (1979).

[4] A. J. Devaney, "Geophysical diffraction tomography," IEEE Trans. Geosci. Electron. **22**, 3 (1984).

[5] R. G. Newton, "Inverse scattering. II. Three dimensions," J. Math. Phys. **21**, 1698 (1980).

[6] C. Morawetz, "A formulation for higher-dimensional inverse problems for the wave equation," Comput. Math. Appl. **7**, 319 (1981).

[7] B. De Facio and J. H. Rose, "Inverse scattering theory for the non-spherically-symmetric three-dimensional plasma wave equation," Phys. Rev. A **31**, 897 (1985).

[8] J. P. Corones, M. E. Davison, and R. J. Krueger, "Direct and inverse scattering in the time domain via invariant imbedding equations," J. Acoust. Soc. Am. **74**, 1535 (1983).

[9] W. Symes, "Stable solution of the inverse reflection problem for a smoothly stratified medium," SIAM J. Math. Anal. **12**, 421 (1981).

[10] A. M. Bruckstein, B. C. Levy, and T. Kailath, "Differential methods in inverse scattering," SIAM J. Appl. Math. **45**, 312 (1985).

[11] A. E. Yagle and B. C. Levy, "The Schur algorithm and its applications," Acta Appl. Math. **3**, 255 (1985).

[12] K. P. Bube and R. Burridge, "The one-dimensional problem of reflection seismology," SIAM Rev. **23**, 497 (1983).

[13] A. E. Yagle and B. C. Levy, "Application of the Schur algorithm to the inverse problem for a layered acoustic medium," J. Acoust. Soc. Am. **76**, 301 (1984).

[14] A. E. Yagle and B. C. Levy, "A layer-stripping solution of the inverse problem for a one-dimensional elastic medium," Geophys. **50**, 425 (1985).

[15] A. E. Yagle and B. C. Levy, "A fast algorithm solution of the inversion problem for a layered acoustic medium probed by spherical hamonic waves," J. Acoust. Soc. Am. **78**, 729 (1985).

[16] A. E. Yagle, "Layer-stripping solutions of inverse seismic problems, Ph.D. Thesis, Dept. of Electrical Engineering and Computer Science, MIT, 1985.

[17] W. Symes, "An inversion method for a three-dimensional wave equation problem," presented at the Workshop on Computational Methods in Ill-Posed and Inverse Problems, Cornell University, Ithaca, NY, July 1983.

[18] J. H. Rose, M. Cheney, and B. DeFacio, "Three-dimensional inverse scattering: Plasma and variable velocity wave equations," J. Math. Phys. **26**, 2803 (1985).

[19] R. Wilcox, "Wave propagation through longitudinally and transversally inhomogeneous slabs—I," in *Invariant Imbedding*, edited by R. E. Bellman and E. D. Denman (Springer, New York, 1970).

[20] D. Stickler, "Inverse scattering in a stratified medium," J. Acoust. Soc. Am. **74**, 994 (1983).

[21] C. Van Winter, "Fredholm equations on a Hilbert space of analytic functions," Trans. Am. Math. Soc. **162**, 103 (1971).

[22] S. Coen, "Density and compressibility profiles of a layered acoustic medium from precritical incidence data," Geophys. **46**, 1244 (1981).

[23] B. Simon, *Quantum Mechanics for Hamiltonians Defined as Quadratic Forms* (Princeton U. P. Princeton, NJ, 1971).

[24] S. R. Deans, *The Radon Transform and Some of its Applications*(Wiley, NY, 1983).

[25] S. J. Norton and M. Linzer, "Ultrasonic reflectivity imaging in three dimensions: Exact inverse-scattering solutions for plane, cylindrical, and spherical apertures," IEEE Trans. Biomed. Eng. **28**, 202 (1981).

[26] J. A. Fawcett, "Inversion of N-dimensional spherical averages," SIAM J. Appl. Math. **45**, 336 (1985).

Feynman rules in the Anderson model: Generalization to the $(2J+1)$-component case

H. Matsumoto and H. Umezawa
Department of Physics, The University of Alberta, Edmonton, Alberta, T6G 2J1 Canada

J. P. Whitehead
Department of Physics Memorial University of Newfoundland, St. John's, Newfoundland, A1B 3X7 Canada

(Received 28 October 1985; accepted for publication 5 February 1986)

The Feynman rules for the Anderson model with a $(2J+1)$-component localized spin are formulated by means of a generalized Wick's expansion together with the reduction formulas in the thermo-field-dynamics. In the $U \to \infty$ limit, Feynman rules for arbitrary J correspond closely to those for the case $J = \frac{1}{2}$ and may be obtained from them by a simple replacement rule.

I. INTRODUCTION

The many-body effects arising from interactions between localized electrons and conduction electrons contain many field-theoretical problems, such as a strong coupling theory, a localized–delocalized problem, and the Kondo effect. There are many publications on these problems to study using the diagrammatic method.[1-5] Although the perturbation scheme may break down in some parameter regions, we hope to extend its applicability if we combine the self-consistency[4,5] or the renormalization group idea.[6] Despite the fact that the Bethe ansatz method[7] gives us an exact solvable method for the single impurity problems in the Kondo model[8] or the Anderson model (in the $U \to \infty$ limit), diagrammatic approaches are still actively studied because a related but more difficult problem of the Kondo and the Anderson lattices still poses a considerable challenge to theoretical physicists. Analysis of the lattice is quite important since the lattice Anderson model, for example, is considered to be a suitable model for the lattice Kondo effect, valence instabilities,[9] moment formation in metals,[10] metal insulator transition,[11] and recently discovered heavy fermion superconductivity.[12] Recently a systematic diagrammatic method[13] for the Anderson model was proposed based on the real time finite temperature field theory, thermofield dynamics (TFD).[14] Feynman rules presented in Ref. 13 are different from those in the conventional field theory owing to the fact that free field operators form a closed algebra different from the harmonic oscillator type and that certain bosonic operators exist, in that algebra, that commute with the unperturbed Hamiltonian (let us call these operators zero-energy bosonic operators). A systematic treatment of zero-energy bosonic operators was discussed in Ref. 13 for the first time. In this paper we extend the results of Ref. 13 for the case of spin-$\frac{1}{2}$ to the general case of a $(2J+1)$-spin, clarifying the mathematical structure to arrive at the generalized Feynman rules. As the following argument will demonstrate, the diagram method is based on a systematic use of the projection operators $P_l(n)$ (see the following) and the thermal state condition in TFD.[15]

II. THE $(2J+1)$-COMPONENT ANDERSON MODEL

The $(2J+1)$-component Anderson model is given by the following Hamiltonian, $H = H_0 + H_I$:

$$H_0 = \epsilon_f n + \frac{U}{2} n(n-1) + \int d^3x\, c^\dagger(x)\epsilon(-i\nabla)c(x)\,, \tag{1}$$

$$H_I = \int d^3x \sum_{m,\sigma} V_{m\sigma}(\mathbf{x})\,[f_m^\dagger\, c_\sigma(\mathbf{x}) + c_\sigma^\dagger\,(\mathbf{x})f_m]\,, \tag{2}$$

where $c(\mathbf{x})$ is the conduction electron field, $f_m\,(m = -J, -J+1,\ldots,J)$ is the localized electron operator, $n = \Sigma_m f_m^\dagger f_m$, and $V_{m\sigma}(\mathbf{x})$ is the potential created by the localized spin at $\mathbf{x} = 0$ and projects out the total angular momentum J state from the conduction electron field. Since f is fermionic, the eigenvalues of n can only assume values in the range 0 to $(2J+1)$.

We introduce the operator ξ_m by

$$\xi_m = P_0(n)f_m\,, \tag{3}$$

where $P_0(n)$ is the projection operator for the $n = 0$ state and is given by

$$P_0(n) = (1-n)(2-n)\cdots(2J+1-n)/(2J+1)!\,. \tag{4}$$

In general we denote the projection operator for $n = l$ state by $P_l(n)$. Noting that

$$P_l(n)P_{l'}(n) = 0\,, \quad \text{for } l \neq l'\,, \tag{5a}$$

$$P_l(n)^2 = P_l(n)\,, \tag{5b}$$

it is easy to see that

$$[H_0, P_l(n)f_m] = -(\epsilon_f + lU)P_l(n)f_m\,. \tag{6}$$

On the other hand, we have

$$P_0(n)f_m = f_m P_1(n)\,, \tag{7a}$$

$$P_0(n)f_m^\dagger = 0\,, \tag{7b}$$

and

$$P_l(n)f_m^\dagger f_{m'} = f_m^\dagger f_{m'} P_l(n)\,, \tag{7c}$$

0022-2488/86/061711-05$02.50

which leads to the result

$$\xi_m \xi_{m'} = 0\,, \tag{8a}$$

$$\xi_{m'}\, \xi^\dagger_m = \delta_{mm'} P_0(n)\,, \tag{8b}$$

$$X_{mm'} \equiv \xi^\dagger_m\, \xi_{m'} = P_1(n) F_{mm'}\,, \tag{8c}$$

with

$$F_{mm'} \equiv f^\dagger_m f_{m'}\,. \tag{8d}$$

Thus, we obtain the following algebra

$$[H_0, \xi_m] = -\epsilon_f \xi_m\,, \tag{9}$$

$$\{\xi_m, \xi_{m'}\} = 0\,, \tag{10}$$

$$M_{mm'} \equiv \{\xi_m, \xi^\dagger_{m'}\} = \delta_{mm'} P_0(n) + P_1(n) F_{m'm}\,. \tag{11}$$

We also have

$$\xi_l M_{mm'} = \delta_{m'l} \xi_m\,, \tag{12a}$$

$$M_{mm'} \xi_l = \delta_{mm'} \xi_l\,, \tag{12b}$$

which leads to the result

$$[\xi_l, M_{mm'}] = c^{ll'}_{mm'}\, \xi_{l'}\,, \tag{13}$$

with

$$c^{ll'}_{mm'} = -\delta_{mm'} \delta_{ll'} + \delta_{lm'} \delta_{ml'}\,. \tag{14}$$

In order to treat the finite temperature regime we must associate with each operator a thermal doublet[14]; for example, ξ is replaced by ξ^α $(\alpha = 1,2)$. The generalization of (9)–(11) in TFD is given by

$$[\hat{H}_0, \xi^\alpha_m] = -\epsilon_f\, \xi^\alpha_m\,, \tag{15}$$

with $\hat{H}_0 = H^1_0 - H^2_0$, and

$$\{\xi^\alpha_m, \xi^{\alpha'}_{m'}\} = 0\,, \tag{16}$$

$$\{\xi^\alpha_m, \xi^{\alpha'\dagger}_m\} = \delta^{\alpha\alpha'} M^\alpha_{mm'}\,, \tag{17}$$

$$[\xi^\alpha_l, M^\gamma_{mm'}] = \epsilon^\alpha \delta^{\alpha\gamma} c^{ll'}_{mm'}\, \xi_{l'}\,, \tag{18}$$

where

$$\begin{aligned} \epsilon^\alpha &= 1 \quad (\alpha = 1)\,, \\ &= -1 \quad (\alpha = 2)\,, \end{aligned} \tag{19}$$

and

$$M^\alpha_{mm'} = P^\alpha_0(n) \delta_{mm'} + P^\alpha_1(n)\, F^\alpha_{m'm}, F^\alpha_{m'm} = \epsilon^\alpha f^{\dagger\alpha}_{m'} f^\alpha_m\,. \tag{20}$$

The relation (8c) now reads as

$$X^\alpha_{mm'} = \epsilon^\alpha\, \xi^{\alpha\dagger}_m\, \xi^\alpha_{m'} = P^\alpha_1(n) F^\alpha_{mm'}\,. \tag{21}$$

A rather remarkable feature of the result contained in Eqs. (16)–(18) is the fact that the operators ξ and M form a closed algebra. This considerably simplifies the calculation of time ordered products of the operators ξ_m,

$$\left\langle T \xi^{\alpha'_1}_{m'_1}(t'_1) \cdots \xi^{\alpha'_l}_{m'_l}(t'_l)\, \xi^{\alpha^\dagger_1}_{m_1}(t_1) \cdots \xi^{\alpha^\dagger_l}_{m_l}(t_l) \right\rangle,$$

where T denotes the time-ordered products in TFD. This is of particular importance when we consider the limiting case $U \to \infty$, since in this limit only the ξ contribution to f remains. However, it should be noted that the results contained in this paper involve no assumptions about the size of U. Furthermore the additional terms involving $(f - \xi)$, which occur in the $U \neq \infty$, may be evaluated by methods analogous to those described here for the ξ products.

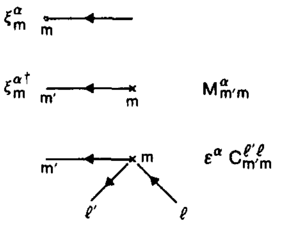

FIG. 1. Diagrammatic expression for ξ and $\xi^\dagger$.

By use of the algebra contained in Eqs. (17) and (18), we can remove the ξ operators successively by means of the reduction formulas given in Ref. 13. Namely $\xi^{\alpha'}(t')$ and $\xi^{\alpha\dagger}(t)$ are contracted to give the propagator $S^{\alpha'\alpha}(t'-t)$ defined by

$$S^{\alpha'\alpha}(t'-t) = \frac{i}{2\pi} \int d\omega\; e^{-i\omega(t-t')} \times U_F(\omega) [\omega - \epsilon_f + i\delta\tau]^{-1}\, U^\dagger_F(\omega)\,, \tag{22}$$

with the appropriate weight associated with the end points. In Eq. (22), $U_F(\omega)$ denotes the fermionic thermal transformation matrix given by

$$U_F(\omega) = \frac{1}{\sqrt{e^{\beta\omega} + 1}} \begin{pmatrix} e^{\beta\omega/2} & 1 \\ -1 & e^{\beta\omega/2} \end{pmatrix}. \tag{23}$$

The appropriate weights are summarized diagrammatically in Fig. 1; ξ^α behaves as an annihilation operator, while $\xi^{\alpha\dagger}$ behaves both as a creation operator with weight M and as the vertex with the weight $\epsilon^\alpha c$, where c is the matrix given in (14). After all the ξ operators have been removed in this way we are left with the time-ordered products involving only M; $\langle T M^{\alpha_1}(t_1) \cdots M^{\alpha_l}(t_l) \rangle$. The multi-$M$ function may be evaluated in a variety of ways. In the following, we present one repesentative method.

First, we note the relation

$$\langle T M^{\alpha_1}(t_1) \cdots M^{\alpha_l}(t_l) \rangle = \langle P_0(n) \rangle + \langle P_1(n) T F^{\alpha_1}(t_1) \cdots F^{\alpha_l}(t_l) \rangle\,. \tag{24}$$

Here we have used (5b), (7c), (8d), (11), and $[P_l(n), H_0] = 0$. In the derivation of Eq. (24) we made successive use of the thermal state condition $\langle P^\alpha_l \cdots \rangle = \langle P_l \cdots \rangle$. We can write

$$F_{mm'} = \frac{1}{2J+1} \left[\delta_{mm'} n + \sum_i (\lambda_i)_{m'm} F_i \right], \tag{25}$$

where

$$\lambda_i \quad (i = 1,2,\ldots,(2J+1) \times (2J+1) - 1)$$

are $(2J+1) \times (2J+1)$ matrices, which are both Hermitian and traceless and which satisfy

$$\mathrm{Tr}[\lambda_i \lambda_j] = \delta_{ij}\,, \tag{26}$$

and the F_i are given by

$$F_i = \sum_{m,m'} f^\dagger_m (\lambda_i)_{mm'} f_{m'}\,. \tag{27}$$

With this notation, we have

$$c^{ll'}_{mm'} = \frac{1}{2J+1}\left[-2J\delta_{mm'}\delta_{ll'} + \sum_i (\lambda_i)_{mm'}(\lambda_i)_{ll'} \right]. \tag{28}$$

Then from (21), (24), and $P_1(n)n = P_1(n)$, it follows that
$\langle TM^{\alpha_1}(t_1)\cdots M^{\alpha_l}(t_l)\rangle$

$$= \langle P_0(n)\rangle + \frac{1}{(2J+1)^l}\Big\langle P_1(n)T\Big\{1 + \sum_i \lambda_i F_i^{\alpha_i}(t_i) + \sum_{i,j}\lambda_i F_i^{\alpha_i}(t_i)\lambda_j F_j^{\alpha_j}(t_j) + \cdots + \lambda_{i_1}F_{i_1}^{\alpha_1}(t_1)\cdots \lambda_{i_l}F_{i_l}^{\alpha_l}(t_l)\Big\}\Big\rangle \tag{29a}$$

$$= \langle P_0(n)\rangle + \frac{1}{(2J+1)^l}\langle P_1(n)\rangle + \frac{1}{(2J+1)^l} \times\Big\{\sum_i \langle \lambda_i X_i^{\alpha_i}(t_i)\rangle + \sum_{ij}\langle T\lambda_i X_i^{\alpha_i}(t_i)\lambda_j X_j^{\alpha_j}(t_j)\rangle + \cdots\Big\}. \tag{29b}$$

Here use was made of Eq. (21). It now remains to calculate products of the form $\langle TX_{i_1}^{\alpha_1}(t_1)\cdots X_{i_l}^{\alpha_l}(t_l)\rangle$.

To evaluate such terms diagrammatically, we look for a reduction formula for the quantities of the form $\langle TX_{A_1}^{\alpha_1}(t_1)\cdots X_{A_i}^{\alpha_i}(t_i)\rangle$ with

$$X_A^\alpha(t) \equiv \epsilon^\alpha \xi^{\dagger\alpha}A\xi^\alpha = P_1^\alpha(n)\epsilon^\alpha f^{\dagger\alpha}Af^\alpha, \tag{30}$$

where A stands for any $(2J+1)\times(2J+1)$ matrix. Note that we have

$$[\xi^\alpha, X_A^\gamma] = \delta^{\alpha\gamma}\epsilon^\gamma A\xi^\gamma. \tag{31a}$$

In particular, when $A = \lambda_i$, we have

$$[\xi_m^\alpha, X_i^\gamma] = \delta^{\alpha\gamma}\epsilon^\gamma\lambda_i\xi^\gamma. \tag{31b}$$

Then Eqs. (16)–(21) together with the relation $P_0(n)$ $X_{mm'} = 0$ give the following reduction formulas:

$$\langle T\xi_m^\alpha(t')X_{A_1}^{\alpha_1}(t_1)\cdots X_{A_i}^{\alpha_i}(t_i)\xi_l^{\beta\dagger}(t)\rangle$$

$$= S^{\alpha\beta}(t'-t)\delta_{ml}[\langle P_0(n)\rangle + (2J+1)^{-1}\langle P_1(n)\rangle], \quad \text{for } i=0, \tag{32}$$

$$= S^{\alpha\beta}(t'-t)\langle TX_{lm}^\beta(t)X_{A_1}^{\alpha_1}(t_1)\cdots X_{A_i}^{\alpha_i}(t_i)\rangle + \sum_{k,j}S^{\alpha\alpha_j}(t'-t_j)(A_j)_{mk}\epsilon^{\alpha_j} \times\langle TX_{A_1}^{\alpha_1}(t_1)\cdots \xi_k^{\alpha_j}(t_j)\cdots X_{A_i}^{\alpha_i}(t_i)\,\xi_l^{\beta\dagger}(t)\rangle, \quad \text{for } i\neq 0, \tag{33a}$$

$$= \langle TX_{lm}^\alpha(t')X_{A_1}^{\alpha_1}(t_1)\cdots X_{A_i}^{\alpha_i}(t_i)\rangle S^{\alpha\beta}(t'-t) + \sum_{k,j}\langle T\xi_m^\alpha(t')X_{A_1}^{\alpha_1}(t_1)\cdots \xi_k^{\alpha_j\dagger}(t_j)\cdots X_{A_i}^{\alpha_i}(t_i)\rangle \times\epsilon^{\alpha_j}(A_j)_{km}S^{\alpha_j\beta}(t_j-t), \quad \text{for } i\neq 0. \tag{33b}$$

Multiplying both sides of (33) with A_{lm}, we obtain

$$\langle TX_A^{\beta\alpha}(t,t')X_{A_1}^{\alpha_1}(t_1)\cdots X_{A_i}^{\alpha_i}(t_i)\rangle = -\epsilon^\alpha S^{\alpha\beta}(t'-t)\langle TX_A^\beta(t)X_{A_1}^{\alpha_1}(t_1)\cdots X_{A_i}^{\alpha_i}(t_i)\rangle + \sum_j \epsilon^\alpha S^{\alpha\alpha_j}(t'-t_j)\langle TX_{AA_j}^{\beta\alpha_j}(t,t_j)X_{A_1}^{\alpha_1}(t_1)\overset{j}{\cdots}X_{A_i}^{\alpha_i}(t_i)\rangle \tag{34a}$$

$$= -\langle TX_A^\alpha(t')X_{A_1}^{\alpha_1}(t_1)\cdots X_{A_i}^{\alpha_i}(t_i)\rangle\epsilon^\alpha S^{\alpha\beta}(t'-t) + \sum_j \langle TX_{A_jA}^{\alpha_j\alpha}(t_j,t')X_{A_1}^{\alpha_1}(t_1)\overset{j}{\cdots}X_{A_i}^{\alpha_i}(t_i)\rangle\epsilon^{\alpha_j}S^{\alpha_j\beta}(t_j-t), \tag{34b}$$

when $i\geqslant 2$. Here, use was made of the notation $X_A^{\alpha_\beta}(t,t') = \epsilon^\alpha\xi^\alpha(t)^\dagger A\xi^\beta(t')$. The notation $\overset{j}{\smile}$ in Eqs. (34) indicates that the operator $X_{A_j}^{\alpha_j}(t_j)$ has been omitted from the time ordered product.

Since (25) gives

$$X_A = (A)P_1(n) + \sum_a (A\lambda_a)X_a \tag{35}$$

with the definition $(A)\equiv(2J+1)^{-1}\operatorname{tr}A$, (34) gives
$\langle TX_A^{\beta\alpha}(t,t')X_{A_1}^{\alpha_1}(t_1)\cdots X_{A_i}^{\alpha_i}(t_i)\rangle$

$$= -\epsilon^\alpha S^{\alpha\beta}(t'-t)\Big[(A)\langle TX_{A_1}^{\alpha_1}(t_1)\cdots X_{A_i}^{\alpha_i}(t_i)\rangle + \sum_a (A\lambda_a)\langle TX_a^\beta(t)X_{A_1}^\alpha(t_1)\cdots X_{A_i}^{\alpha_i}(t_i)\rangle\Big] + \sum_j \epsilon^\alpha S^{\alpha\alpha_j}(t'-t_j)\langle TX_{AA_j}(t,t_j)X_{A_1}^{\alpha_1}(t_1)\overset{j}{\cdots}X_{A_i}^{\alpha_i}(t_i)\rangle \tag{36a}$$

$$= -\Big[\langle TX_{A_1}^{\alpha_1}(t_1)\cdots X_{A_i}^{\alpha_i}(t_i)\rangle(A) + \sum_a \langle TX_a^\alpha(t')X_{A_1}^{\alpha_1}(t_1)\cdots X_{A_i}^{\alpha_i}(t_i)\rangle(\lambda_a A)\Big]\epsilon^\alpha S^{\alpha\beta}(t'-t) + \sum_j \langle TX_{A_jA}^{\alpha_j\alpha}(t_j,t')X_{A_1}^{\alpha_1}(t_1)\overset{j}{\cdots}X_{A_i}^{\alpha_1}(t_i)\rangle\epsilon^{\alpha_j}S^{\alpha_j\beta}(t_j-t), \tag{36b}$$

for $i\neq 0$. In particular, when we assume $\alpha=\beta$ and $t'=t-\epsilon^\alpha 0$, Eq. (34) gives

$$\langle TX^\alpha_A(t)X^{\alpha_1}_{A_1}(t_1)\cdots X^{\alpha_i}_{A_i}(t_i)\rangle$$

$$=\frac{1}{\cos^2\theta_f}\sum_j \epsilon^\alpha S^{\alpha\alpha_j}(t-t_j)\langle TX^{\alpha\alpha_j}_{AA_j}(t,t_j)$$

$$\times X^{\alpha_1}_{A_1}(t_1)\overset{\overset{j}{\smile}}{\cdots}X^{\alpha_1}_{A_1}(t_i)\rangle \qquad (37a)$$

$$=\frac{1}{\cos^2\theta_f}\sum_j \langle TX^{\alpha_j\alpha}_{A_jA}(t_j,t)$$

$$\times X^{\alpha_1}_{A_1}(t_1)\overset{\overset{j}{\smile}}{\cdots}X^{\alpha_i}_{A_i}(t_i)\rangle\epsilon^{\alpha_j}S^{\alpha_j\alpha}(t_j-t)\,, \qquad (37b)$$

where use was made of the relation

$$\epsilon^\alpha S^{\alpha\alpha}(-\epsilon_\alpha 0)=\sin^2\theta_f=1/(e^{\beta\epsilon_f}+1)\,.$$

With the particular choice, $A=\lambda_a$ and $A_j=\lambda_j$, (37) reads as

$$\langle TX^\alpha_a(t)X^{\alpha_1}_1(t_1)\cdots X^{\alpha_i}_i(t_i)\rangle$$

$$=\frac{1}{\cos^2\theta_f}\sum_j \epsilon^\alpha S^{\alpha\alpha_j}(t-t_j)\langle TX^{\alpha\alpha_j}_{aj}(t,t_j)$$

$$\times X^{\alpha_1}_1(t_1)\overset{j}{\cdots}X^{\alpha_i}_i(t_i)\rangle\,, \qquad (38a)$$

$$=\frac{1}{\cos^2\theta_f}\sum_k \langle TX^{\alpha_j\alpha}_{ja}(t_j,t)$$

$$\times X^{\alpha_1}_1(t_1)\overset{\overset{j}{\smile}}{\cdots}X^{\alpha_1}_i(t_i)\rangle\epsilon^{\alpha_k}S^{\alpha_j\alpha}(t_j-t)\,, \qquad (38b)$$

where $X^{\alpha\beta}_{aj}$ means $\epsilon^\alpha\xi^{\alpha\dagger}\lambda_a\lambda_j\xi^\beta$.

The structure of the reduction formula presented in Eqs. (34) and (35) are analogous to those obtained in Ref. 13. We can therefore simplify the reduction formula in an analogous manner to that presented in Ref. 13. Further reduction can be achieved by substitution of (36) into (38). If we introduce the notation $(ij\cdots)=(\lambda_i\lambda_j\cdots)$ together with the relation $(ikj)=-(jki)$ we obtain, for $i\geqslant 2$,

$$\langle TX^\alpha_a(t)X^{\alpha_1}_1(t_1)\cdots X^{\alpha_i}_i(t_i)\rangle$$

$$=(\cos^2\theta_f)^{-1}\Big\{-\sum_j \epsilon^\alpha S^{\alpha\alpha_j}(t-t_j)\epsilon^{\alpha_j}$$

$$\times S^{\alpha_j}(t_j-t)(a\,j)\langle TX^{\alpha_1}_1(t_1)\overset{\overset{j}{\smile}}{\cdots}X^{\alpha_i}_i(t_i)\rangle$$

$$+\sum_{k,j}\epsilon^{\alpha_j}S^{\alpha_j\alpha}(t_j-t)\epsilon^\alpha S^{\alpha\alpha_k}(t-t_k)$$

$$\times\langle TX^{\alpha_j\alpha_k}_{jak}(t_j,t_k)X^{\alpha_1}_1(t_1)\overset{\overset{k}{\smile}}{\cdots}\overset{\overset{j}{\smile}}{\cdots}X^{\alpha_i}_i(t_i)\rangle\Big\}\,. \qquad (39)$$

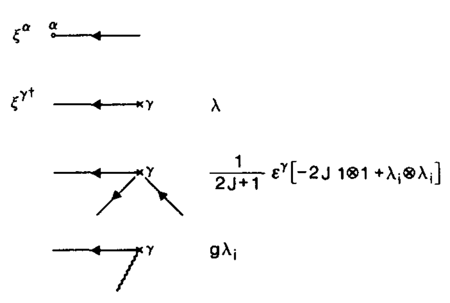

FIG. 2. Feynman rules for ξ and $\xi^\dagger$.

TABLE I. Values of λ and g.

	$\langle P_0\rangle$	$\langle P_1\rangle$	$\langle P_{01}\rangle$
λ	1	$1/(2J+1)$	$1/(2J+1)$
g	0	0	$1/(2J+1)$

The reduction formulas (32)–(37) are equivalent to those presented in Ref. 13 (in which $J=\frac{1}{2}$ was considered) when we make the following replacements:

$$\langle n\rangle\to\langle P_1(n)\rangle,\quad \langle 1-n\rangle\to\langle P_0(n)\rangle,\quad \langle 2-n\rangle\to\langle P_{01}\rangle\,, \qquad (40a)$$

$$\sigma_i\to\lambda_i,\quad \tfrac{1}{2}\to 1/(2J+1)\,, \qquad (40b)$$

where the projection operator P_{01} is defined as

$$P_{01}(n)\equiv(2J+1)P_0(n)+P_1(n)\,. \qquad (41)$$

Therefore the generalized Feynman rules given in Ref. 13 may be extended to higher values of the multiplicity J when we make the above replacements. The rules are summarized in Fig. 2 and the weight factors λ and g are given in Table I.

III. CONCLUDING REMARKS

We close this paper emphasizing again that if we restrict ourselves to time-ordered products involving only ξ operators, then all of the formulas presented in this paper are true for any value of J and U. In the limit $U\to\infty$, we have

$$\langle P_0\rangle=1/(1+(2J+1)e^{-\beta\epsilon_f})\,,$$

$$\langle P_1\rangle=(2J+1)e^{-\beta\epsilon_f}/(1+(2J+1)e^{-\beta\epsilon_f})\,, \qquad (42)$$

which gives

$$\langle P_{01}\rangle=(2J+1)(1+e^{-\beta\epsilon_f})/(1+(2J+1)e^{-\beta\epsilon_f})\,.$$

As pointed out earlier in the limit $U=\infty$, we require the multipoint functions consisting only of ξ. When $U\neq\infty$, there appear other eigenoperators besides ξ. The calculation of their multipoint functions may be evaluated by using diagrammatic rules similar to the ones presented in this paper. For example, the operators η_m and $\eta^\dagger_m$ defined by

$$\eta_m=P_{2J}f_m,\quad \eta^\dagger_m=f^\dagger_m P_{2J}\,, \qquad (43)$$

also form a closed algebra as

$$\{\eta^\dagger_m,\eta_{m'}\}=P_{2J+1}\delta_{mm'}+P_{2J}f_m f^\dagger_{m'}$$

$$=\phi_{mm'} \qquad (44)$$

and

$$[\eta_l,\phi_{mm'}]=(\delta_{mm'}\delta_{ll'}-\delta_{lm'}\delta_{ml'})\eta_{l'}\,. \qquad (45)$$

Thus nearly all of the present analysis may be used to calculate a time ordered product of the form

$$\langle T\eta^{\alpha_1}_{m_1}(t_1)\cdots\eta^{\alpha_1}_{m_i}(t_i)\,\eta^{+\gamma_1}_{m'_1}\cdots\eta^{+\gamma_i}_{m'_i}(t'_i)\rangle$$

with very little modification. The application of the diagrammatic techniques described here and in Ref. 13 to specific problems of interest is currently in progress.

ACKNOWLEDGMENTS

This work was partly supported by the Natural Science and Engineering Research Council, Canada and grants from the Dean of the Faculty of Science, The University of Alberta, Edmonton, Alberta, Canada.

[1] V. G. Vaks, A. I. Larkin, and S. A. Pikin, Zh. Eksp. Teor. Fiz. **53**, 281 (1967) [Sov. Phys. JETP **26**, 188 (1968)]; R. O. Zaitsev, Zh. Eksp. Teor. Fiz. **68**, 207 (1975) [Sov. Phys. JETP **41**, 100 (1975)]; A. F. Barabanov and A. M. Tsvelik, Fiz. Tverd. Tela (Leningrad) **21**, 3214 (1979) [Sov. Phys. Solid State **21**, 1855 (1979)]; R. O. Zaitev, Zh. Eksp. Teor. Fiz. **70**, 1100 (1976) [Sov. Phys. JETP **43**, 574 (1976)]; M. Roberts and K. W. H. Steevens, J. Phys. C **13**, 5941 (1980); E. V. Anda, J. Phys. C **14**, L1037 (1981).

[2] H. Keiter and J. C. Kimball, Intern. J. Magn. **1**, 233 (1971); N. Grewe and H. Keiter, Phys. Rev. B **24**, 4420 (1981).

[3] S. E. Barnes, J. Phys. F **6**, 1375 (1976); **7**, 2637 (1977).

[4] Y. Kuramoto, Z. Phys. B **53**, 37 (1983); H. H. Kojima, Y. Kuramoto, and M. Tachiki, Z. Phys. B **54**, 293 (1984).

[5] P. Coleman, Phys. Rev. B **29**, 3035 (1984).

[6] H. R. Krishna-murthy, J. W. Wilkins, and K. G. Wilson, Phys. Rev. B **21**, 1003, 1044 (1980).

[7] See, for example, the review articles, N. Andrei, K. Furuya, and J. H. Lowenstein, Rev. Mod. Phys. **55**, 331 (1983); A. M. Tsvelick and P. B. Wiegmann, Adv. Phys. **32**, 453 (1983).

[8] J. Kondo, Prog. Theor. Phys. **32**, 37 (1964).

[9] See, for example, the reviews, C. M. Varma, Rev. Mod. Phys. **48**, 219 (1976); J. M. Lawrence, P. S. Riseborough, and R. D. Parks, Rep. Prog. Phys. **44**, 1 (1981).

[10] See, for example, *Magnetic Moment Formation in Metals*, edited by W. J. L. Buyers (Plenum, New York, 1984).

[11] W. F. Brinkman and T. M. Rice, Phys. Rev. B **2**, 4302 (1972).

[12] See, for example, the recent review G. R. Stewart, Rev. Mod. Phys. **56**, 755 (1984).

[13] H. Matsumoto and H. Umezawa, Phys. Rev. B **31**, 4433 (1985).

[14] See, for example, H. Umezawa, H. Matsumoto, and M. Tachiki, *Thermofield Dynamics and Condensed States* (North-Holland, Amsterdam, 1982).

[15] H. Matsumoto, Y. Nakano, and H. Umezawa, Phys. Rev. D **31**, 1495 (1985).