Exact solutions of unitarily invariant matrix models in zero dimensions

P. A. Amundsen and P. H. Damgaard
Nordita, Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark

(Received 26 April 1983; accepted for publication 18 October 1983)

Explicit constructions of Green functions for $U(N)$ invariant-matrix Φ^4 theories are given in zero space-time dimensions. The results are expressed in terms of a new class of orthogonal polynomials, which are orthogonal on the interval $[-\infty, \infty]$ with respect to the weight function $\exp\{-m^2\phi^2 - \lambda\phi^4\}$. An explicit construction of the coefficients of these generalized Hermite polynomials is presented. We briefly discuss the "thermodynamic" limit $N \rightarrow \infty$ of the corresponding one-dimensional statistical system of N classical particles with logarithmic pair interactions and subjected to an external anharmonic potential.

PACS numbers: 02.10.Sp

1. INTRODUCTION

Interest in unitarily invariant-matrix field theories based on some internal $U(N)$ invariance has up to now mainly been confined to the formal $N \rightarrow \infty$ limit, in which only the planar sector is probed. Using an elegant steepest-descent technique, Brezin, Itzykson, Parisi and Zuber[1] managed to solve exactly, in the $N \rightarrow \infty$ limit, the unitary-matrix Φ^4 theory in zero space-time dimensions. The same method has subsequently been applied to a host of other model field theories, such as zero-dimensional two-matrix models[2] and two-dimensional $U(N)$ lattice gauge theories.[3]

Recently, $U(N)$ matrix models for large N have received renewed attention. With the important discovery[4] that the so-called quenched momentum prescription may represent a valid prescription even in the continuum, new possibilities for large-N theories have suddenly opened up. Perhaps a particularly interesting aspect is the hope it has given of obtaining the hypothesized "master field" of large-N theories,[5] but many other interesting applications are clearly possible.

Simultaneously with these developments in quenched field theories, the large-N behavior of the four-dimensional version of the $U(N)$ matrix model we consider here (at negative couplings to make it asymptotically free) was shown by 't Hooft to be "finite" in the sense that the perturbation expansion is Borel summable.[6] This puts the planar approximation on a very solid theoretical foundation.

Yet, we are not of course really interested in $N = \infty$ theories. Even though it may seem almost stupefyingly difficult to extend in a simple and calculable way the quenched momentum prescription to the case of finite N, it is not altogether inconceivable that some of the lessons learned recently for $N = \infty$ theories may be carried over to the physical case of finite N. For this reason it is important to have available a class of exactly soluble models on which to test the methods. Furthermore, with such exact solutions available one can directly see to what degree of approximation the $N \rightarrow \infty$ limit makes sense.

Apart from these considerations, the unitarily invariant matrix models in zero dimensions that we shall be concerned with in this paper have an interest in their own right. For example, the generating functional of this zero-dimensional theory can be used to directly obtain the weight factors of the perturbative expansions associated with the corresponding d-dimensional theory. (Loosely speaking, the weight factor gives the sum of all symmetry factors, color factors, etc. to each order in perturbation theory, i.e., the result one would obtain if one assigned the value 1 to each Feynman diagram.) These zero-dimensional matrix models are also of interest from the point of view of statistical mechanics. We shall show that one obtains directly the partition function of a one-dimensional statistical ensemble of N classical particles interacting through a logarithmic-pair potential, and subjected to an external anharmonic-oscillator potential. Taking the limit in which the coupling to the external field approaches zero, one obtains the partition function whose form was conjectured several years ago by Dyson.[7] Keeping the coupling to the external potential nonzero, we can apply the steepest-descent model of Ref. 1 (now also treating the temperature as a dynamical variable) to obtain the exact partition function in the "thermodynamic" limit $N \rightarrow \infty$.

The paper is organized as follows. In Sec. 2 the Lagrangian is defined, and we show how the vacuum-to-vacuum amplitude can be calculated exactly by relating it to the normalization factors of a set of orthogonal polynomials. From this we show how all relevant Green functions, as well as the generating functional itself, can be constructed. Section 3 is devoted to a discussion of the set of orthogonal polynomials needed to compute the vacuum-to-vacuum amplitude. We show how in this case the standard orthogonalization procedure reduces to a set of algebraic equations, which in turn make it possible to construct the coefficients of nth polynomial directly and without reference to the lower-order polynomials. The $N \rightarrow \infty$ "thermodynamic" limit of the corresponding statistical-mechanics system is calculated in Sec. 4. Finally, Sec. 5 contains a short summary.

2. THE VACUUM-TO-VACUUM AMPLITUDE

The theory we will be considering here is defined by a matrix field $\Phi_{ij}(x)$ and its corresponding Euclidean Lagrangian

$$\mathscr{L} = \mathrm{Tr}[\partial_\mu \Phi \partial_\mu \Phi^+] + m^2 \,\mathrm{Tr}[\Phi\Phi^+] + \lambda \,\mathrm{Tr}[(\Phi\Phi^+)^2]. \tag{2.1}$$

Obviously, if $\Phi_{ij}(x)$ belongs to the space of complex Hermitian matrices, the Lagrangian (2.1) will be globally invariant under the group $U(N)$. It was this theory which was solved

 0022-2488/84/061633-05$02.50

exactly in zero dimensions for $N \to \infty$ in Ref. 1 by a steepest-descent method. We shall here give the solution for finite N.

In zero space-time dimensions the propagators all become simple constants, and the kinetic term in the Lagrangian drops out. The vacuum-to-vacuum amplitude

$$\langle 0^+ | 0^- \rangle = Z_N(\lambda) = \int D\Phi e^{-S[\Phi]}, \tag{2.2}$$

where $S[\Phi]$ is the Euclidean action,

$$S[\Phi] = \int d^D x \, \mathscr{L}[\Phi(x)] \tag{2.3}$$

becomes simply

$$Z_N(\lambda) = \int D\Phi \exp(-m^2 \operatorname{Tr}[\Phi\Phi^+] - \lambda \operatorname{Tr}[(\Phi\Phi^+)^2]). \tag{2.4}$$

Here $D\Phi$ represents the integration measure over complex Hermitian matrices. A more tractable representation of the path integral (2.4) is obtained in a standard manner by diagonalization Φ_{ij} and the inclusion of the corresponding "Jacobian".[7,8] The result is (apart from an overall unitary group-volume factor which is irrelevant for our purpose)

$$Z_N(\lambda) = \int_{-\infty}^{\infty} \prod_{i=1}^{N} d\phi_i \prod_{i<j} (\phi_i - \phi_j)^2 e^{-m^2 \Sigma \phi_i^2 - \lambda \Sigma \phi_i^4}, \tag{2.5}$$

in which equation the N integration variables ϕ_i are the eigenvalues of the field Φ_{ij}.

First note that $Z_N(\lambda)$ can be solved by an exact cluster decomposition in which one assigns $(1 - F_{ij})$ to the ijth link and a factor $\phi_i^{2(i-1)}$ to the ith vertex. We have introduced the link variables

$$F_{ij} \equiv f_{ij}^2 - 2f_{ij}, \tag{2.6}$$

where the f's are simply given by the ratio of the eigenvalues, $f_{ij} = \phi_i/\phi_j$. However, such a graphical solution becomes extremely tedious for moderately large values of N.

The clue to a simpler solution, which will be useful for all values of N, lies in noting that the factor $\Pi_{i<j}(\phi_i - \phi_j)^2$ appearing in Eq. (2.5) is the square of the Vandermonde determinant

$$D(\phi_1,\dots,\phi_N) = \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \phi_1 & \phi_2 & \cdots & \phi_N \\ \phi_1^{N-1} & \phi_2^{N-1} & \cdots & \phi_N^{N-1} \end{vmatrix} \tag{2.7}$$

A brute-force solution can now be obtained by simply expanding the determinant (2.7) out in rows and columns. Upon squaring the result, obviously only even powers of the ϕ_i's will contribute to the integral (2.5). One can therefore group the answer into blocks of ϕ_i-products, each ϕ_i appearing with only even powers, and such that the sum of the exponents equal $N(N-1)$. There are $[(N/2]!]^2$ such blocks if N is even, and $[((N+1)/2)!((N-1)/2)!]$ blocks if N is odd. Each block appears with a multiplicity given by standard combinatorics, and there is finally an overall factor of $N!$ due to the permutation symmetry. This brute-force method also becomes quite tedious for moderately large values of N, although it is probably easier than the cluster decomposition method.

Instead, we shall present a solution which makes use of a fundamental property of the Vandermonde determinant of Eq. (2.7), namely its invariance under the replacement of its lower $(N-1)$ rows by arbitrary $(N-1)$th-order polynomials in the ϕ_i's, provided we choose to normalize our polynomials such that the coefficient of the highest power is unity:

$$D(\phi_1,\dots,\phi_N) = \begin{vmatrix} P_0(\phi_1) & P_0(\phi_2)\cdots P_0(\phi_N) \\ P_1(\phi_1) & P_1(\phi_2)\cdots P_1(\phi_N) \\ P_{N-1}(\phi_1) & P_{N-1}(\phi_2)\cdots P_{N-1}(\phi_N) \end{vmatrix}. \tag{2.8}$$

According to our normalization condition $P_0(\phi_i) = 1$, $P_1(\phi_i) = \phi_i + \text{const}$, and so on. Note that Eq. (2.8) is exact for any arbitrary choice of the $(N-1)$ polynomials $P_1(\phi_i),\dots,P_{N-1}(\phi_i)$. In particular, we can choose these polynomials such that they are orthogonal on the interval $[-\infty, \infty]$ with respect to the weight function

$$W(\phi_i) = e^{-m^2\phi_i^2 - \lambda\phi_i^4}. \tag{2.9}$$

With this particular choice of the polynomials $P_n(\phi_i)$ all crossed terms in the square of the Vandermonde determinant vanish upon integration, and one is left with the simple result

$$Z_N(\lambda) = N! \prod_{i=0}^{N-1} \langle P_i | P_i \rangle, \tag{2.10}$$

where we have introduced the notation

$$\langle P_i | P_j \rangle = \int_{-\infty}^{\infty} d\phi_l \, P_i(\phi_l) P_j(\phi_l) e^{-m^2\phi_l^2 - \lambda\phi_l^4}. \tag{2.11}$$

The orthogonality condition requires this inner product to be proportional to δ_{ij}. We shall return to the question of how to construct the polynomials $P_n(\phi_i)$ in the next section.

Having calculated with vacuum-to-vacuum amplitude in the form of Eq. (2.10), it is now trivial to obtain any Green functions. For example, for the 2-point function

$$\langle \operatorname{Tr}\Phi^2 \rangle = \frac{\int D\Phi \operatorname{Tr}[\Phi^2] \exp\{-m^2 \operatorname{Tr}[\Phi^2] - \lambda \operatorname{Tr}[\Phi^4]\}}{\int D\Phi \exp\{-m^2 \operatorname{Tr}[\phi^2] - \lambda \operatorname{Tr}[\Phi^4]\}} \tag{2.12}$$

one has

$$\langle \operatorname{Tr}\Phi^2 \rangle = \frac{\partial}{\partial m^2} \ln[Z_N(\lambda)], \tag{2.13}$$

with $Z_N(\lambda)$ given explicitly by Eq. (2.10). Similarly, all other $U(N)$ invariant-vacuum expectation values can be generated from $Z_N(\lambda)$ by differentiations with respect to m^2. This gives the implicit construction of the generating functional of $U(N)$ invariants.

3. THE POLYNOMIALS $P_n(\phi_i)$

In the previous section we expressed the vacuum-to-vacuum amplitude $Z_N(\lambda)$ as the product of the normalization factors of the polynomials $P_n(\phi_i)$. In this section we shall construct these polynomials, and their normalization factors, explicitly.

First, let us define

$$I_\beta(\alpha) = \int_{-\infty}^{\infty} dx \, e^{-\alpha x^2 - \beta x^4} = \tfrac{1}{2}\sqrt{\frac{\alpha}{\beta}} \, e^{\alpha^2/8\beta} K_{1/4}\left(\frac{\alpha^2}{8\beta}\right), \tag{3.1}$$

where K_ν is a modified Bessel function of the second kind. We thus have

$$I'_\beta(\alpha) = \left(\frac{1}{2\alpha} + \frac{\alpha}{4\beta}\right) I_\beta(\alpha) + \frac{\alpha^{3/2}}{8\beta^{3/2}} e^{\alpha^2/8\beta} K'_{1/4}\left(\frac{\alpha^2}{8\beta}\right),$$

the prime denoting differentiation with respect to the argument. Exploiting the differential equation for Bessel functions, we find the useful result

$$I_\beta^{(n+2)}(\alpha) = (\alpha/2\beta) I_\beta^{(n+1)}(\alpha) + (2n + 1/\beta) I^{(n)}(\alpha), \quad (3.2)$$

which gives us the explicit expression

$$I_\beta^{(n)}(\alpha) = \frac{(-1)^n \Gamma(n+\frac{1}{2})}{2^n \beta^{n/2} \alpha^{1/2}} e^{\alpha^2/8\beta} W_{-n/2,-1/4}\left(\frac{\alpha^2}{8\beta}\right). \quad (3.3)$$

Here $W_{\mu,\nu}$ represents the irregular Whittaker function, and Γ is the standard gamma function. (Our conventions follow those of Ref. 9.) From the explicit representation (3.1) it is clear that $W_{-n/2,-1/4}$ always can be represented in terms of $K_{1/4}$ and its first derivative, or in terms of $K_{1/4}$ and $K_{5/4}$, or $K_{1/4}$ and $K_{3/4}$.

We write the polynomials

$$P_n(x) = \sum_{j=0}^{n} p_{nj}(\alpha,\beta) x^j,$$

$$p_{nn} = 1, \; p_{nj} = 0 \quad \text{for } n-j \text{ odd},$$

$$\int_{-\infty}^{\infty} dx\, P_n(x) P_m(x) e^{-\alpha x^2 - \beta x^4} = A_n(\alpha,\beta)\delta_{mn}. \quad (3.4)$$

The coefficients $p_{nj}(\alpha,\beta)$ can be determined, for instance, by the usual Gram–Schmidt orthogonalization procedure. We note, however, that starting with $P_0(x) = 1$ ($P_1(x) = x$) for n even (odd), the condition of orthogonality with respect to x^2, x^4,...(x^3,x^5,...) can be represented by differentiations with respect to α. Thus we get a set of algebraic equations,

$$\sum_\nu M_{\mu\nu}^{(n)} P_\nu^{(n)} = I_\nu^{(n)}, \quad (3.5)$$

where $\mu, \nu = 1,...,[n/2]$, and

$$I_\nu^{(n)} = I_\beta^{(\nu + [n-1)/2]}(\alpha),$$

$$M_{\mu\nu}^{(n)} = I_\beta^{(\mu+\nu+e_n-2)}(\alpha), \quad e_n = \begin{cases} 0 \text{ for } n \text{ even} \\ 1 \text{ for } n \text{ odd} \end{cases}$$

$$P_\mu^{(n)} = (-1)^{\mu + [n/2] - 1} p_{n,\mu+e_n}. \quad (3.6)$$

Here $[y]$ stands for the integral part of y. From Eq. (3.5) we immediately have the coefficients $P_{nj}(\alpha,\beta)$ for any given n,

$$P_\mu^{(n)} = \sum_\nu (M^{(n)^{-1}})_{\mu\nu} I_\nu^{(n)}. \quad (3.7)$$

The normalization factor $A_n(\alpha,\beta)$ then follows from

$$A_n = \langle P_n | P_n \rangle = \sum_{j,k=0}^{n} p_{nj} p_{nk} (-1)^{j+k/2} I_\beta^{(j+k/2)}(\alpha), \quad (3.8)$$

and the polynomials $P_n(x)$ are now completely determined. We list the coefficients of some of the lowest-order polynomials in Table I.

The polynomials $P_n(x)$ provide the natural generalization of Hermite polynomials, to which they reduce as $\beta \to 0$,

$$P_n(x) \to (2\sqrt{\alpha})^{-n} H_n(\sqrt{\alpha}x) \quad \text{as } \beta \to 0, \quad (3.9)$$

as can also be checked explicitly from Table 1.

4. THE CONNECTION TO STATISTICAL MECHANICS

The vacuum-to-vacuum amplitude $Z_N(\lambda)$ of Eq. (2.5) can, as mentioned in the introduction, be given the interpretation of the partition function of a particular one-dimensional gas, which we shall now describe.

First note that we can add to the exponent in Eq. (2.5) a term

$$H_{\text{kin}} = \sum_{i=1}^{N} \frac{P_i^2}{2m_i} \quad (4.1)$$

and integrate all P_i's from $-\infty$ to $+\infty$. These are all simple Gaussian integrals which will add an overall factor to $Z_N(\lambda)$. We can interpret this modified vacuum-to-vacuum amplitude as the partition function for a classical one-dimensional gas of N particles having phase-space coordinates (p_i, ϕ_i). It follows from Eq. (2.5) that these N particles interact with an external potential V_{ext} given by

$$V_{\text{ext}} = m^2 \sum_{i=1}^{N} \phi_i^2 + \lambda \sum_{i=1}^{N} \phi_i^4, \quad (4.2)$$

and have a logarithmic pair interaction,

$$V_{\text{pair}} = -\sum_{i \neq j} \ln|\phi_i - \phi_j|. \quad (4.3)$$

TABLE I. The coefficients p_{nj} [from Eq. (3.4)] of the polynomials $P_n(x)$ for low values of n. Note that by definition $p_{nn} = 1$. Higher-order coefficients can be computed by the method outlined in Sec. 3.

j \ n	0	1	2	3	4	5
0	1	0	$\frac{I'_\beta(\alpha)}{I_\beta(\alpha)}$	0	$\frac{I''_\beta(\alpha)^2 - I'''_\beta(\alpha)I'_\beta(\alpha)}{I'_\beta(\alpha)^2 - I''_\beta(\alpha)I_\beta(\alpha)}$	0
1		1	0	$\frac{I''_\beta(\alpha)}{I'_\beta(\alpha)}$	0	$\frac{I'''_\beta(\alpha)^2 - I^{(4)}_\beta(\alpha)I''_\beta(\alpha)}{I''_\beta(\alpha)^2 - I'''_\beta(\alpha)I'_\beta(\alpha)}$
2			1	0	$\frac{I''_\beta(\alpha)I'_\beta(\alpha) - I'''_\beta(\alpha)I_\beta(\alpha)}{I'_\beta(\alpha)^2 - I''_\beta(\alpha)I_\beta(\alpha)}$	0
3				1	0	$\frac{I'''_\beta(\alpha)I''_\beta(\alpha) - I^{(4)}_\beta(\alpha)I'_\beta(\alpha)}{I''_\beta(\alpha)^2 - I'''_\beta(\alpha)I'_\beta(\alpha)}$
4					1	0
5						1

The matrix eigenvalues ϕ_i, in suitable units, simply play the role of position coordinates of the N particles.

If we let both $m^2 \to 0$ and $\lambda \to 0$, such that the coupling to the external potential vanishes, then $Z_N(\lambda)$ reduces to the form of the famous conjecture of Dyson.[7] This original conjecture of Dyson was shortly afterwards given a beautiful proof by Wilson.[10]

Note that if we keep the external potential (4.2), $Z_N(\lambda)$ will, once the parameters m and λ have been chosen, only produce the partition function for the one-dimensional gas at fixed temperature T. To find the full partition function we must allow the pair interaction V_{pair} to depend on an overall constant which can be chosen independently of m and λ. We will call this coupling constant σ:

$$V_{\text{pair}} \to -\sigma \sum_{i \neq j} \ln|\phi_i - \phi_j|. \tag{4.4}$$

The equivalent problem for $\lambda = 0$ was studied years ago by Mehta and Dyson.[7] They conjectured the solution for the partition function Z_N for all N. This conjecture has, to our knowledge, never been proved. We shall here study the generalized problem in which $\lambda \neq 0$, in the thermodynamic limit $N \to \infty$. Wishing to retain as $N \to \infty$ a smooth limit for the partition function, we are allowed to rescale the parameters appropriately. In particular, we will choose to rescale $\lambda \to \lambda / N$ and let $\phi_i \to \sqrt{N}\phi(i/N)$, as in Ref. 1. The matrix eigenvalues ϕ_i then become continuous variables on the interval $[0,1]$. The parameters m and σ are left untouched. It is, as we shall see shortly, this particular rescaling which will yield an appropriate $N \to \infty$ limit.

Now, as $N \to \infty$ the steepest-descent appproximation becomes exact, and the solution to our partition function will follow completely analogous to the treatment in Ref. 1. Here we just briefly recall the argument in order to define our notation.

The steepest-descent solution selects the stationary point

$$m^2\phi_i + 2\frac{\lambda}{N}\phi_i^3 = \sigma \sum_{i \neq j} \frac{1}{\phi_i - \phi_j}. \tag{4.5}$$

Rearranging the sum over eigenvalues such that the sum becomes ordered we can, in the limit $N \to \infty$, turn the sum into an integral. After the rescaling mentioned above we therefore obtain

$$m^2\phi(x) + 2\lambda\phi(x)^3 = \sigma \fint \frac{dy}{\phi(x) - \phi(y)}, \tag{4.6}$$

where $\fint$, as usual, denotes the principal value of the integral.

If we introduce the $N \to \infty$ density of eigenvalues $\rho(\phi) = dx/d\phi$ normalized on a finite interval 2Λ such that

$$\int_{-\Lambda}^{\Lambda} d\phi \rho(\phi) = 1, \tag{4.7}$$

then Eq. (4.5), with a simple change of variables, becomes

$$m^2\phi + 2\lambda\phi^3 = \sigma \int_{-\Lambda}^{\Lambda} d\mu \frac{\rho(\mu)}{\phi - \mu}. \tag{4.8}$$

This integral equation can be solved by the method of Ref. 1, and the solution is

$$\rho(\phi) = \frac{1}{\pi\sigma}\left[(m^2\phi + 2\lambda\phi^3 + \lambda\Lambda^2\phi)\sqrt{\frac{\Lambda^2}{\phi^2} - 1}\right], \tag{4.9}$$

subject to the constraint

$$3\lambda\Lambda^4 + 2m^2\Lambda^2 - 4\sigma = 0 \tag{4.10}$$

and defined for $\phi \in [-\Lambda, \Lambda]$. We note that we cannot now take the cutoff Λ to infinity, since in fact Λ is fixed by equation (4.10). This implies that in fact the whole system is only in equilibrium when confined to the interval $[-\Lambda, \Lambda]$. The size of the system depends on σ such that $\Lambda \to 0$ as $\sigma \to 0$. The fact that the system becomes restricted to a finite interval is a property which holds for any analytic choice of the external potential $V_{\text{ext}}(\phi)$ with a nontrivial $N \to \infty$ limit.

The saddle-point solution itself can now be determined immediately. Writing the partition function

$$Z_N(\lambda) = e^{-N^2 H/\beta}, \tag{4.11}$$

it follows that

$$\begin{aligned}\frac{H}{\beta} &= \int_0^1 dx[m^2\phi(x)^2 + \lambda\phi(x)^4] \\ &\quad - \sigma\int_0^1\int_0^1 dx\, dy \ln|\phi(x) - \phi(y)| \\ &= \int_{-\Lambda}^{\Lambda} d\phi\, \rho(\phi)[m^2\phi^2 + \lambda\phi^4] \\ &\quad - \sigma\int_{-\Lambda}^{\Lambda}\int_{-\Lambda}^{\Lambda} d\mu\, d\phi\, \rho(\phi)\rho(\mu)\ln|\phi - \mu|.\end{aligned} \tag{4.12}$$

We can use the saddle-point equation (4.8) to rewrite this as

$$\frac{H}{\beta} = \int_0^{\Lambda} d\phi\, \rho(\phi)[m^2\phi^2 + \lambda\phi^4 - 2\sigma\ln\phi], \tag{4.13}$$

which, upon inserting $\rho(\phi)$ from Eq. (4.9) becomes

$$\begin{aligned}\frac{H}{\beta} &= \frac{1}{\sigma}\left[\frac{m^4}{16}\Lambda^4 + \frac{5\lambda m^2}{32}\Lambda^6 + \frac{9\lambda^2}{128}\Lambda^8\right] \\ &\quad + \frac{m^2}{2}\Lambda^2\left(\tfrac{1}{2} - \ln\left(\frac{\Lambda}{2}\right)\right) + \frac{3\lambda}{4}\Lambda^4\left(\tfrac{1}{4} - \ln\left(\frac{\Lambda}{2}\right)\right),\end{aligned} \tag{4.14}$$

and the full partition function [Eq. (4.11)] is then known. Λ and the pair coupling σ are of course related through the constraint (4.10).

5. SUMMARY

In this note we have shown that the zero-dimensional Φ^4 Hermitian-matrix model, which previously was solved by a saddle-point technique in the large N limit, in fact can be solved exactly for all N. We have given the explicit form of the vacuum-to-vacuum amplitude and shown how to construct all Green functions of the theory. To do this, a set of orthogonal polynomials, which provide a natural generalization of Hermite polynomials, was introduced. All relevant quantities can be expressed in terms of the normalization constants of these polynomials. We note that this method of solution is completely general for matrix models, and can be applied to other effective Lagrangian methods as well. We showed that the analog of this zero-dimensional matrix model in statistical mechanics is a $(1+1)$-dimensional system of N classical particles with logarithmic-pair interactions and subjected to an external anharmonic-oscillator potential. The partition function for this system was computed in the "thermodynamic" limit $N \to \infty$.

ACKNOWLEDGMENT

We thank Paul Mansfield for useful discussions.

Note: It has been pointed out to us that the equivalent of the polynomial method used in this paper has been used by Y. Y. Goldschmidt [J. Math. Phys. **21**, 1842 (1980)] in an investigation of the large N limit of two-dimensional lattice gauge theories. We thank Michael Peskin for drawing our attention to this reference.

[1]E. Brezin, C. Itzykson, G. Parisi, and J. B. Zuber, Commun. Math. Phys. **59**, 35 (1978).
[2]C. Itzykson and J. B. Zuber, J. Math. Phys. **21**, 411 (1980).
[3]D. J. Gross and E. Witten, Phys. Rev. D **21**, 446 (1980).
[4]G. Parisi, Phys. Lett. B **112**, 463 (1982). D. J. Gross and Y. Kitazawa, Nucl. Phys. B **206**, 440 (1982).
[5]J. Greensite and M. B. Halpern, Nucl. Phys. B **211**, 343 (1983). J. Greensite, Phys. Lett. B **121**, 169 (1983).
[6]G. 't Hooft, Phys. Lett. B **119**, 369 (1982).
[7]F. J. Dyson, J. Math. Phys. **3**, 140 (1962). M. L. Mehta and F. J. Dyson, J. Math. Phys. **4**, 713 (1963).
[8]M. L. Mehta, *Random Matrices* (Academic, New York, 1967).
[9]I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products* (Academic, New York, 1980), 2nd ed.
[10]K. G. Wilson, J. Math. Phys. **3**, 1040 (1962).

Dynamical group of microscopic collective states. III. Coherent state representations in d dimensions

J. Deenen and C. Quesne[a)]
Physique Théorique et Mathématique CP 229, Université Libre de Bruxelles, Bd. du Triomphe, B 1050 Brussels, Belgium

(Received 31 October 1983; accepted for publication 16 December 1983)

The present series of papers deals with various realizations of the dynamical group $\mathscr{S}\!p_c(2d,R)$ of microscopic collective states for an A nucleon system in d ($=1,2$, or 3) dimensions, when these collective states are assumed to be invariant under the orthogonal group O(n) associated with the $n = A - 1$ relative Jacobi vectors. In this paper, we further study the Barut–Girardello representation proposed in the first two papers of the present series to show that it may be reformulated in terms of some coherent states by generalizing to $\mathscr{S}\!p_c(2d,R)$ a class of Sp(2,R) coherent states introduced by Barut and Girardello. For such purpose, our starting point is another coherent state representation, namely the Perelomov one, previously considered by Kramer. We also propose a third, new class of coherent states leading to Holstein–Primakoff representation in a straightforward way. We review various properties of these three classes of coherent states, such as their reproducing kernel and measure explicit forms, their generating function properties, and the representations they lead to for both the collective states and their dynamical group.

PACS numbers: 02.20. + b, 21.60.Fw

1. INTRODUCTION

In various recent works,[1–3] the microscopic collective states of an A nucleon system are discussed in terms of the orthogonal group O(n) associated with the $n = A - 1$ relative Jacobi vectors of the system. These states are assumed to transform under a definite O(n) irreducible representation (IR) $(\lambda_1\lambda_2\lambda_3)$, which can be derived from shell model considerations. It is actually obtained by filling compactly with the A nucleons all the single-particle states in an oscillator well up to a given level and by considering the most symmetrical IR $[\lambda_1\lambda_2\lambda_3]$ of U(3) in the last unfilled level.

Although complete sets of states are now available for the general case wherein λ_1, λ_2, and λ_3 are not all equal,[3,4] many works have been devoted up to now to the simpler case of the O(n) scalar IR for which $\lambda_1 = \lambda_2 = \lambda_3 = 0$.[1,5–8] Strictly speaking, the scalar case only applies to s-shell nuclei. Its consideration interest, however, goes much beyond this very restricted class of nuclei. With slight modifications (actually by replacing n by $n + 2\lambda$ in all the formulas), the results for the scalar case can indeed be extended to the case where $\lambda_1 = \lambda_2 = \lambda_3 = \lambda \neq 0$, corresponding to closed-shell nuclei. Moreover, the detailed study of the scalar case may provide us with some useful hints to deal with the more complicated case of open-shell nuclei, for which λ_1, λ_2, and λ_3 are not all equal.

The purpose of the present series of papers is to study various realizations of the dynamical group of microscopic collective states in the scalar case. The analysis is carried out in a d-dimensional space. The physical relevant case of course corresponds to $d = 3$, but it is also useful to consider simplified models for which $d = 1$ or 2.

In the first paper[7] (henceforth referred to as I), we did show that in d dimensions the dynamical group of microscopic collective states is the restriction $\mathscr{S}\!p_c(2d,R)$ to the collective subspace of a real symplectic group in $2d$ dimensions $\mathscr{S}\!p(2d,R)$. We then introduced two new realizations of $\mathscr{S}\!p_c(2d,R)$, that we studied in detail only in the one-dimensional case. The first one corresponds to a representation of collective states in the O(n) invariant subspace of a Bargmann Hilbert space: this is the Barut–Girardello (BG) representation (improperly called Barut representation in Paper I), wherein the collective states are represented by analytic functions in $\nu = d(d+1)/2$ complex variables $w_{ij} = w_{ji}$, $i,j = 1,\ldots,d$. It was shown in Paper I that in one dimension, the BG representation can be formulated in terms of a class of coherent states (CS), introduced by Barut and Girardello[9] for the Sp(2,R) group. The second realization of $\mathscr{S}\!p_c(2d,R)$ proposed in I is the so-called generalized Holstein–Primakoff (HP) representation,[10] wherein the $\mathscr{S}\!p_c(2d,R)$ generators are expressed in terms of ν boson creation operators and the corresponding annihilation operators.

Boson representations of $\mathscr{S}\!p_c(2d,R)$ for $d > 1$ were studied in the second paper of the present series (henceforth, referred to as II).[8] We did show there that the BG representation is equivalent to a generalized Dyson representation[11] of $\mathscr{S}\!p_c(2d,R)$, and we explicitly obtained the HP representation in d dimensions.

In the present paper, we would like to complete our analysis as outlined in I by generalizing the Barut–Girardello coherent states (BGCS) to d dimensions where $d > 1$, and by showing that they lead to BG representation in a straightforward way. For such purpose, it is convenient to use a different approach from that adopted in I for the one-dimensional case. Our present starting point is the Perelomov coherent state (PCS) representation[12] of $\mathscr{S}\!p_c(2d,R)$ as considered by Kramer,[6] whose main properties are reviewed

[a)] Maître de recherches F.N.R.S.

 0022-2488/84/061638-13$02.50

in Sec. 2. In Sec. 3, the BGCS of $\mathcal{Sp}_c(2d,R)$ are defined, their representation in the PCS basis obtained, and some of their properties listed. In Sec. 4, the overlap of two BGCS is calculated to derive explicit expressions of BGCS. A convenient procedure to determine the measure for BGCS is described and applied to the two-dimensional case in Sec. 5. In Sec. 6, a third class of CS leading to HP representation is introduced. Finally, Sec. 7 contains some concluding remarks.

2. PERELOMOV COHERENT STATE REPRESENTATION

In Paper I, it was shown that under the assumption that the collective states are scalar under $O(n)$, they belong to a single IR $\langle (n/2)^d \rangle$ of an $\mathcal{Sp}(2d,R)$ group, whose generators are defined by

$$\mathscr{D}_{ij}^{\dagger} = \mathscr{D}_{ji}^{\dagger} = \sum_{s=1}^{n} \eta_{is}\, \eta_{js}\,, \quad 1 \leqslant i \leqslant j \leqslant d\,, \tag{2.1a}$$

$$\mathscr{D}_{ij} = \mathscr{D}_{ji} = \sum_{s=1}^{n} \xi_{is}\xi_{js}\,, \quad 1 \leqslant i \leqslant j \leqslant d\,, \tag{2.1b}$$

and

$$\mathscr{E}_{ij} = \mathscr{C}_{ij} + \frac{n}{2}\delta_{ij}\,, \quad \mathscr{C}_{ij} = \sum_{s=1}^{n} \eta_{is}\xi_{js}\,, \quad i,j = 1,...,d\,. \tag{2.1c}$$

Here η_{is} and ξ_{is}, $i = 1,..., d$, $s = 1,..., n$, respectively, denote the boson creation and annihilation operators associated with the relative Jacobi coordinates x_{is}, $i = 1,..., d$, $s = 1,..., n = A - 1$. The dynamical group of collective states is then the restriction $\mathcal{Sp}_c(2d,R)$ of $\mathcal{Sp}(2d,R)$ to this single IR.

According to Perelomov,[12] generalized CS of $\mathcal{Sp}_c(2d,R)$ based on the lowest weight state of the IR $\langle (n/2)^d \rangle$, i.e., the boson vacuum state $|0\rangle$, can be defined by the following relation

$$|\mathbf{u}\rangle = \exp(\tfrac{1}{2}\,\mathrm{tr}\,\mathbf{u}^*\mathscr{D}^{\dagger})|0\rangle\,, \tag{2.2}$$

where $\mathscr{D}^{\dagger}$ is the $d \times d$ matrix of the $\mathscr{D}_{ij}^{\dagger}$ generators introduced in Sec. 2 of Paper II, $\mathbf{u}$ is some $d \times d$ symmetrical complex matrix, and $\mathbf{u}^*\mathscr{D}^{\dagger}$ denotes usual matrix multiplication. Since these states were already considered in Ref. 6, we shall merely list their main properties.

One leading property of PCS is that they form an overcomplete, nonorthogonal family of states. The overlap of two PCS, characterized by some parameters $\mathbf{u}$ and $\mathbf{u}'$, respectively, is given by

$$\langle \mathbf{u}'|\mathbf{u}\rangle = [\det(\mathbf{I} - \mathbf{u}'\mathbf{u}^*)]^{-n/2}\,, \tag{2.3}$$

where $\mathbf{I}$ denotes the $d \times d$ unit matrix. The completeness of the set is expressed by the unity resolution relation within the representation space of $\langle (n/2)^d \rangle$,

$$\int d\hat{\sigma}(\mathbf{u})|\mathbf{u}\rangle\langle\mathbf{u}| = I_c\,. \tag{2.4}$$

Here the measure $d\hat{\sigma}(\mathbf{u})$ is given by

$$d\hat{\sigma}(\mathbf{u}) = \hat{f}(\mathbf{u},\mathbf{u}^*)d\mathbf{u}\, d\mathbf{u}^*\,, \tag{2.5}$$

where

$$\hat{f}(\mathbf{u},\mathbf{u}^*) = \hat{A}\,[\det(\mathbf{I} - \mathbf{u}\mathbf{u}^*)]^{n/2-d-1}\,, \tag{2.6}$$

$$d\mathbf{u}\, d\mathbf{u}^* = \prod_{i \leqslant j} d\,\mathrm{Re}\, u_{ij}\; d\,\mathrm{Im}\, u_{ij}\,, \tag{2.7}$$

and the parameter space is the origin-centered unit ball. In Eq. (2.6), $\hat{A}$ is a normalization constant, determined by the condition

$$\int d\hat{\sigma}(\mathbf{u}) = 1\,. \tag{2.8}$$

Consequently to the unity resolution relation, we may use PCS as a (continuous) basis of the collective subspace.

In Eqs. (8.6) and (8.10) of Paper I, we defined, up to some normalization constant, discrete basis states of the collective subspace $|\phi_{N_{11}N_{12}\cdots N_{dd}}\rangle$, where $N_{ij} = 0,1,...,$ for $1 \leqslant i \leqslant j \leqslant d$, hereafter quoted as $|\mathbf{N}\rangle$ for short, with $\mathbf{N}$ representing the whole set of quantum numbers $N_{11},N_{12},...,N_{dd}$. It will prove convenient to choose their normalization in such a way that

$$|\phi_{N_{11}N_{12}\cdots N_{dd}}\rangle = |\mathbf{N}\rangle = \prod_{i \leqslant j} (N_{ij}!)^{-1/2}[(1 + \delta_{ij})^{-1/2}\mathscr{D}_{ij}^{\dagger}]^{N_{ij}}|0\rangle\,. \tag{2.9}$$

In the following, we shall call them oscillator basis states. As is well known, they do not form an orthogonal set. With the normalization chosen in Eq. (2.9), they are neither normalized to unity.

The usefulness of PCS for practical purposes mainly stems from the fact that they are generating functions for the oscillator basis states. By expanding the exponential in Eq. (2.2), we indeed obtain

$$|\mathbf{u}\rangle = \sum_{\mathbf{N}} F_{\mathbf{N}}(\mathbf{u}^*)|\mathbf{N}\rangle\,, \tag{2.10}$$

where the summation over $N_{11},N_{12},...,N_{dd}$ runs over all nonnegative integers, and

$$F_{\mathbf{N}}(\mathbf{u}^*) = \prod_{i \leqslant j} (N_{ij}!)^{-1/2}[(1 + \delta_{ij})^{-1/2}u_{ij}^*]^{N_{ij}}\,. \tag{2.11}$$

Since PCS form a basis of the collective subspace, they can be used to get a representation of the collective states and of their dynamical group, where any collective state $|\psi\rangle$ is represented by an analytic function

$$\hat{\hat{\psi}}(\mathbf{u}) = \langle \mathbf{u}|\psi\rangle \tag{2.12}$$

in the ν complex variables $u_{ij} = u_{ji}$, $1 \leqslant i \leqslant j \leqslant d$. In particular, the oscillator basis states are represented by the functions

$$\hat{\hat{\phi}}_{\mathbf{N}}(\mathbf{u}) = \langle \mathbf{u}|\mathbf{N}\rangle = \sum_{\mathbf{N}'} F_{\mathbf{N}'}(\mathbf{u})M_{\mathbf{N}',\mathbf{N}}\,, \tag{2.13}$$

where

$$M_{\mathbf{N}',\mathbf{N}} = \langle \mathbf{N}'|\mathbf{N}\rangle \tag{2.14}$$

is an element of the oscillator basis overlap matrix $\mathbf{M}$. The presence of the latter in Eq. (2.13) makes Perelomov representation of collective states quite complicated. We shall not therefore derive an explicit expression for $\hat{\hat{\phi}}_{\mathbf{N}}(\mathbf{u})$.

The Perelomov representation of any operator X is a differential operator $\hat{X}^c$ with respect to the u_{ij} variables, defined by the following relation:

$$\langle \mathbf{u}|X|\psi\rangle = \hat{X}^c\langle \mathbf{u}|\psi\rangle\,, \tag{2.15}$$

where $|\psi\rangle$ is any collective state. For instance, it is shown in Appendix A that the Perelomov representation of the $\mathcal{Sp}_c(2d,R)$ generators is given by

$$\hat{\mathscr{D}}^{c\dagger} = [\hat{\mathscr{C}}^c + (n-d-1)\mathbf{I}]\mathbf{u} = \mathbf{u}[\hat{\tilde{\mathscr{C}}}^c + n\mathbf{I}]\,, \quad (2.16a)$$

$$\hat{\mathscr{D}}^c = \Delta_{\mathbf{u}}\,, \quad (2.16b)$$

$$\hat{\mathscr{E}}^c = \hat{\mathscr{C}}^c + \frac{n}{2}\mathbf{I}\,, \quad \hat{\mathscr{C}}^c = \mathbf{u}\Delta_{\mathbf{u}}\,, \quad (2.16c)$$

where we use the matrix notation introduced in Sec. 2 of Paper II, and $\Delta_{u_{ij}}$ is defined by

$$\Delta_{u_{ij}} = (1+\delta_{ij})\frac{\partial}{\partial u_{ij}}\,. \quad (2.17)$$

It is important to note that the operators (2.16) are first-order differential operators and therefore provide us with a quite convenient realization of the $\mathscr{S}\!p_c(2d,R)$ generators.

3. BARUT–GIRARDELLO COHERENT STATE REPRESENTATION

An alternative definition of generalized coherent states, only valid for noncompact groups, was proposed by Barut and Girardello[9] for SO(2,1) and its locally isomorphic groups SU(1,1), SL(2,R), and Sp(2,R), and used in I in connection with the BG representation of collective states in one dimension. The present section purpose is to extend to d dimensions the definition of such coherent states and to study some of their properties.

According to Ref. 9, and Eq. (6.6) in Paper I, the BG coherent states of $\mathscr{S}\!p_c(2,R)$ are the eigenstates of the operator $\mathscr{D}_{11} = \mathscr{D}$ belonging to the collective subspace. To any complex number w, there corresponds (up to some normalization factor) one and only one BG coherent state $|w)$ whose eigenvalue is w^*,

$$\mathscr{D}|w) = w^*|w)\,. \quad (3.1)$$

It is therefore opportune to question whether in d dimensions it is possible to find common eigenstates $|\mathbf{w})$ of the set of commuting operators $\mathscr{D}_{ij}$, $1 \leqslant i \leqslant j \leqslant d$, corresponding to some complex eigenvalues $w^*_{ij}(=w^*_{ji})$,

$$\mathscr{D}_{ij}|\mathbf{w}) = w^*_{ij}|\mathbf{w})\,. \quad (3.2)$$

Here $\mathbf{w}$ denotes, as usual, the symmetrical complex $d \times d$ matrix whose elements are w_{ij}.

To answer this question, let us calculate the scalar product of both sides of Eq. (3.2) with a PCS $|\mathbf{u}\rangle$,

$$\langle\mathbf{u}|\mathscr{D}_{ij}|\mathbf{w}) = w^*_{ij}\langle\mathbf{u}|\mathbf{w})\,. \quad (3.3)$$

From Eqs. (2.15) and (2.16b), we obtain the following result:

$$\Delta_{u_{ij}}\langle\mathbf{u}|\mathbf{w}) = w^*_{ij}\langle\mathbf{u}|w)\,. \quad (3.4)$$

When i and j run from 1 to d, Eq. (3.4) gives rise to a system of ν independent first-order partial differential equations for $\langle\mathbf{u}|\mathbf{w})$, whose solution is given by

$$\langle\mathbf{u}|\mathbf{w}) = G(\mathbf{w}^*)\exp(\tfrac{1}{2}\operatorname{tr}\mathbf{u}\mathbf{w}^*)\,, \quad (3.5)$$

where $G(\mathbf{w}^*)$ is an arbitrary function of $\mathbf{w}^*$. Let us fix the normalization of $|\mathbf{w})$ in such a way that

$$\langle 0|\mathbf{w}) = 1\,. \quad (3.6)$$

Since the PCS $|\mathbf{0}\rangle$, whose parameters are all equal to zero, reduces to the vacuum state $|0\rangle$, Eq. (3.6) imposes that

$$G(\mathbf{w}^*) = 1\,, \quad (3.7)$$

so that Eq. (3.5) becomes

$$\langle\mathbf{u}|\mathbf{w}) = \exp(\tfrac{1}{2}\operatorname{tr}\mathbf{u}\mathbf{w}^*)\,. \quad (3.8)$$

We have therefore proved that for any complex values of the set of parameters $w_{ij} = w_{ji}$, the system of equations (3.2) has [up to some normalization factor we choose in accordance with Eq. (3.6)] one and only one solution to be written as

$$|\mathbf{w}) = \int d\hat{\sigma}(\mathbf{u})\exp(\tfrac{1}{2}\operatorname{tr}\mathbf{u}\mathbf{w}^*)|\mathbf{u}\rangle \quad (3.9)$$

in the PCS basis. The collection of states $|\mathbf{w})$ will be called the BGCS of $\mathscr{S}\!p_c(2d,R)$. Note the slight change of our notations with respect to I as a round bracket is now used instead of an angular one for the BGCS to distinguish them from the PCS corresponding to the same set of parameters.

Let us now review some properties of the BGCS in d dimensions. In I, we established [cf. Eq. (8.19) in Paper I] that in the BG representation the oscillator basis states (2.9) are represented by the analytic functions

$$\bar{\phi}_{\mathbf{N}}(\mathbf{w}) = \prod_{i\leqslant j}(N_{ij}!)^{-1/2}[(1+\delta_{ij})^{-1/2}w_{ij}]^{N_{ij}} \quad (3.10)$$

in the complex collective variables

$$w_{ij} = \sum_{s=1}^{n} z_{is}z_{js}\,, \quad (3.11)$$

defined in terms of the Bargmann representation z_{is} of the boson creation operators η_{is}. On the other hand, if we take the scalar product of the oscillator basis states (2.9) with BGCS and use the definition (3.2) of the latter, as well as the Hermiticity property $(\mathscr{D}_{ij})^\dagger = \mathscr{D}^\dagger_{ij}$, we obtain for $(\mathbf{w}|\mathbf{N}\rangle$ the same expression as the right-hand side of Eq. (3.10). We therefore conclude that the BG representation of the oscillator basis states may be regarded as their scalar product with the BGCS,

$$\bar{\phi}_{\mathbf{N}}(\mathbf{w}) = (\mathbf{w}|\mathbf{N}\rangle\,. \quad (3.12)$$

Hence, the BG representation of any collective state is given by

$$\bar{\psi}(\mathbf{w}) = (\mathbf{w}|\psi\rangle\,. \quad (3.13)$$

We have simultaneously proved that the set of complex parameters $\mathbf{w}$, used in Eq. (3.2), has a well-defined meaning in Bargmann space, as shown in Eq. (3.11). Note incidentally that the above considerations justify *a posteriori* the notation $(\mathbf{w}|\psi\rangle$ somewhat loosely used in II to denote the BG representation $\bar{\psi}(\mathbf{w})$ of $|\psi\rangle$ [see, e.g., Eqs. (2.4) and (2.5) in Paper II].

By using the definition of BG representation given in I, it is now trivial to show that the BGCS give rise to a unity resolution relation in the collective subspace and therefore form a continuous basis of the latter. Since the BG space $\mathscr{F}_c$ is the O(n) invariant subspace of Bargmann space $\mathscr{F}$, the scalar product of two oscillator basis states $|\mathbf{N}\rangle$ and $|\mathbf{N}'\rangle$ can be calculated by taking the scalar product in Bargmann space of their BG representations $\bar{\phi}_{\mathbf{N}}(\mathbf{w})$ and $\bar{\phi}_{\mathbf{N}'}(\mathbf{w})$,

$$\langle\mathbf{N}'|\mathbf{N}\rangle = \int\left[\prod_{i=1}^{d}\prod_{s=1}^{n}d\mu(z_{is})\right][\bar{\phi}_{\mathbf{N}'}(\mathbf{w})]^*\bar{\phi}_{\mathbf{N}}(\mathbf{w})\,. \quad (3.14)$$

Here $d\mu(z_{is})$ is the usual Bargmann measure[13]

$$d\mu(z_{is}) = \pi^{-1}\exp(-z_{is}z^*_{is})d\operatorname{Re}z_{is}\,d\operatorname{Im}z_{is}\,. \quad (3.15)$$

By integrating over the $dn - \nu$ noncollective variables, which together with the ν collective variables w_{ij} form a set equivalent to the dn complex variables z_{is}, we can, in principle, express $\langle \mathbf{N}'|\mathbf{N}\rangle$ as an integral over the w_{ij} variables only,

$$\langle \mathbf{N}'|\mathbf{N}\rangle = \int d\sigma(\mathbf{w})[\bar{\phi}_{\mathbf{N}'}(\mathbf{w})]^* \bar{\phi}_{\mathbf{N}}(\mathbf{w}), \tag{3.16}$$

with some measure

$$\begin{aligned} d\sigma(\mathbf{w}) &= f(\mathbf{w},\mathbf{w}^*) d\mathbf{w}\, d\mathbf{w}^*, \\ d\mathbf{w}\, d\mathbf{w}^* &= \prod_{i \leqslant j} d\,\mathrm{Re}\, w_{ij}\, d\,\mathrm{Im}\, w_{ij}, \end{aligned} \tag{3.17}$$

directly deriving from Bargmann one. Since Eq. (3.16) is valid for any $\mathbf{N}$ and $\mathbf{N}'$, we obtain from Eq. (3.12) the sought for unity resolution relation

$$\int d\sigma(\mathbf{w})|\mathbf{w})(\mathbf{w}| = I_c, \tag{3.18}$$

where the integration is carried out over the whole complex plane for each variable w_{ij}. Eq. (3.18) implies that for any two collective states $|\phi\rangle$ and $|\psi\rangle$

$$\langle \phi|\psi\rangle = \int d\sigma(\mathbf{w})\langle \phi|\mathbf{w})(\mathbf{w}|\psi\rangle = \int d\sigma(\mathbf{w})[\bar{\phi}(\mathbf{w})]^* \bar{\psi}(\mathbf{w}). \tag{3.19}$$

As it was already noted in I for $d = 1$, the above-mentioned procedure to determine the weight function $f(\mathbf{w},\mathbf{w}^*)$ would be quite inconvenient in practice. In Sec. 5, we shall present an alternative method and apply it to the two-dimensional case.

Since the BG representation considered in I and II actually coincides with BGCS representation, we may apply to the latter all the results previously derived for the former. Let us denote by X^c the BGCS representation of an operator X, so that

$$(\mathbf{w}|X|\psi\rangle = X^c(\mathbf{w}|\psi\rangle \tag{3.20}$$

for any collective state $|\psi\rangle$. From Eq. (2.11) in II, the representation of the $\mathcal{S}p_c(2d,R)$ generators is given by

$$\mathcal{D}^{c\dagger} = \mathbf{w}, \tag{3.21a}$$

$$\mathcal{D}^c = \Delta_{\mathbf{w}}[\mathcal{C}^c + (n-d-1)\mathbf{I}] = [\tilde{\mathcal{C}}^c + n\mathbf{I}]\Delta_{\mathbf{w}}, \tag{3.21b}$$

$$\mathcal{E}^c = \mathcal{C}^c + (n/2)\mathbf{I}, \quad \mathcal{C}^c = \mathbf{w}\Delta_{\mathbf{w}}, \tag{3.21c}$$

where $\Delta_{w_{ij}}$ is defined by

$$\Delta_{w_{ij}} = (1+\delta_{ij})\frac{\partial}{\partial w_{ij}}. \tag{3.22}$$

When comparing Eq. (3.21) with the PCS representation given in Eq. (2.16), we note some striking similarities. In a way we may consider that one goes from the former to the latter by replacing $\mathbf{w}$ and $\Delta_{\mathbf{w}}$ by $\Delta_{\mathbf{u}}$ and $\mathbf{u}$, respectively, and by interchanging $\mathcal{D}^\dagger$ with $\mathcal{D}$. However, these similarities are misleading. As a matter of fact, there is one essential difference between both representations of $\mathcal{S}p_c(2d,R)$, which turns out to have far-reaching consequences for the determination of the measure: namely the $\mathcal{S}p_c(2d,R)$ generators are represented in BGCS representation by second-order differential operators instead of first-order ones in PCS representation.

An important leftover question is the determination of the BGCS explicit form in terms of the $\mathcal{D}^\dagger_{ij}$ generators, i.e., the analog of Eq. (2.2) for the PCS. Since this is a rather hard problem, we postpone its solution to the next section.

We conclude the present section by studying the properties of BGCS as generating functions. As noted in the previous section, the oscillator basis (2.9) is not orthogonal; hence we may consider its dual basis whose elements (for which we use a round bracket instead of an angular one) are defined by the following relation:

$$|\mathbf{N}) = \sum_{\mathbf{N}'} |\mathbf{N}'\rangle(\mathbf{M}^{-1})_{\mathbf{N}',\mathbf{N}}, \tag{3.23}$$

in terms of the inverse $\mathbf{M}^{-1}$ of the oscillator basis overlap matrix $\mathbf{M}$. At this point, it is worth noting that since

$$M_{\mathbf{N}',\mathbf{N}} = 0 \quad \text{if} \sum_{i \leqslant j}(N'_{ij} - N_{ij}) \neq 0, \tag{3.24}$$

the infinite-dimensional matrix $\mathbf{M}$ is block diagonal, and the submatrices on the diagonal are finite-dimensional; its inverse $\mathbf{M}^{-1}$ is therefore well defined. The oscillator basis and its dual one form a biorthogonal system, i.e.,

$$\langle \mathbf{N}'|\mathbf{N}) = \delta_{\mathbf{N}',\mathbf{N}}, \tag{3.25}$$

and give rise to the following unity resolution relation:

$$\sum_{\mathbf{N}} |\mathbf{N})\langle \mathbf{N}| = I_c. \tag{3.26}$$

As a consequence, the BGCS can be written as

$$|\mathbf{w}) = \sum_{\mathbf{N}} |\mathbf{N})\langle \mathbf{N}|\mathbf{w}). \tag{3.27}$$

By using Eqs. (3.10) and (3.12), the latter expression becomes

$$|\mathbf{w}) = \sum_{N} F_{\mathbf{N}}(\mathbf{w}^*)|\mathbf{N}), \tag{3.28}$$

where $F_{\mathbf{N}}(\mathbf{w}^*)$ is defined by an equation similar to Eq. (2.11). We therefore conclude that the BGCS are generating functions for the dual basis states, with the same expansion coefficients as those appearing in the expansion (2.10) of the PCS in terms of the oscillator basis states.

4. EXPLICIT FORM OF BARUT–GIRARDELLO COHERENT STATES

In the present section, we wish to determine the function K in the $\mathcal{D}^\dagger_{ij}$ generators giving rise to a BGCS $|\mathbf{w})$ when applied to the vacuum state. Since this function also depends upon the eigenvalues w^*_{ij} of $\mathcal{D}_{ij}$ parametrizing the BGCS, we may write

$$|\mathbf{w}) = K(\mathcal{D}^\dagger,\mathbf{w}^*)|0\rangle. \tag{4.1}$$

To start with, we note that the overlap or reproducing kernel of two BGCS is given by the following relation:

$$(\mathbf{w}'|\mathbf{w}) = (\mathbf{w}'|K(\mathcal{D}^\dagger,\mathbf{w}^*)|0\rangle = K(\mathcal{D}^{c\dagger},\mathbf{w}^*)(\mathbf{w}'|0\rangle, \tag{4.2}$$

which becomes

$$(\mathbf{w}'|\mathbf{w}) = K(\mathbf{w}',\mathbf{w}^*), \tag{4.3}$$

when using Eqs. (3.6) and (3.21a). From Eqs. (4.1) and (4.3), it is then clear that the determination of the BGCS explicit form and the calculation of their overlap are equivalent problems. The latter is however easier to solve since it only requires the determination of an ordinary function $K(\mathbf{w}',\mathbf{w}^*)$ instead of an operator function $K(\mathscr{D}^\dagger,\mathbf{w}^*)$ as it is the case for the former.

The function $K(\mathbf{w}',\mathbf{w}^*)$ construction is based upon the Hermiticity properties of the $\mathscr{S}p(2d,R)$ generators

$$\mathscr{E}_{ij} = \mathscr{E}_{ji}^\dagger , \tag{4.4a}$$

$$\mathscr{D}_{ij} = (\mathscr{D}_{ij}^\dagger)^\dagger , \tag{4.4b}$$

which imply that

$$(\mathbf{w}'|\mathscr{E}_{ij}|\mathbf{w}) = (\mathbf{w}|\mathscr{E}_{ji}|\mathbf{w}')^* , \tag{4.5a}$$

and

$$(\mathbf{w}'|\mathscr{D}_{ij}|\mathbf{w}) = (\mathbf{w}|\mathscr{D}_{ij}^\dagger|\mathbf{w}')^* . \tag{4.5b}$$

By using Eqs. (3.20) and (3.21), Eqs. (4.5a) and (4.5b), respectively, become

$$(\mathbf{w}'\Delta_{\mathbf{w}'})_{ij}K(\mathbf{w}',\mathbf{w}^*) = (\mathbf{w}^*\Delta_{\mathbf{w}^*})_{ji}K(\mathbf{w}',\mathbf{w}^*) , \tag{4.6a}$$

$$\{\Delta_{\mathbf{w}'}[\mathbf{w}'\Delta_{\mathbf{w}'} + (n-d-1)\mathbf{I}]\}_{ij}K(\mathbf{w}',\mathbf{w}^*) = w_{ij}^*K(\mathbf{w}',\mathbf{w}^*) , \tag{4.6b}$$

and represent two systems of d^2 first-order and $d(d+1)/2$ second-order partial differential equations to be satisfied by $K(\mathbf{w}',\mathbf{w}^*)$. We shall proceed to show that among the solutions of Eq. (4.6), there is one and only one function, analytic in both $\mathbf{w}'$ and $\mathbf{w}^*$, and normalized in accordance with Eq. (3.6), i.e., satisfying the conditions

$$K(\mathbf{0},\mathbf{w}^*) = K(\mathbf{w}',\mathbf{0}) = 1 . \tag{4.7}$$

This unique function will therefore be the function $K(\mathbf{w}',\mathbf{w}^*)$ we are looking for.

Starting with Eq. (4.6a), we shall proceed to prove that its general solution is an arbitrary function in the d independent invariants of the $d\times d$ matrix

$$\mathbf{W} = \mathbf{w}'\mathbf{w}^* . \tag{4.8}$$

The demonstration of this property is sketched below, additional details being given in Appendix B. It is essentially based upon two changes of variables. Owing to the symmetry of $\mathbf{w}'$ and $\mathbf{w}^*$, K is a function in $d(d+1)$ independent variables that we can take as w'_{ij} and w^*_{ij}, $i\leqslant j$. We first replace the latter by the d^2 variables W_{ij}, $i,j = 1,\ldots,d$, and the d variables

$$X_i = w'_{ii} + w^*_{ii} , \quad i = 1,\ldots,d . \tag{4.9}$$

In the new variables, Eq. (4.6a) can be written as

$$\left[\sum_k \left(W_{ik}\frac{\partial}{\partial W_{jk}} - W_{kj}\frac{\partial}{\partial W_{ki}}\right) + 2w'_{ij}\frac{\partial}{\partial X_j} - 2w^*_{ij}\frac{\partial}{\partial X_i}\right]K(\mathbf{W},\mathbf{X}) = 0 , \tag{4.10}$$

where w'_{ij} and w^*_{ij} are some functions of $\mathbf{W}$ and $\mathbf{X}$ which could in principle be obtained by inverting Eqs. (4.8) and (4.9).

By linearly combining the d^2 equations (4.10), it is possible to get d very simple equations. Let us indeed multiply Eq. (4.10) by $(\mathbf{W}^m)_{ji}$, where $\mathbf{W}^m$ denotes the mth power of the $\mathbf{W}$ matrix for some fixed m value belonging to the set $0,1,\ldots,d-1$, and let us sum the result over i and j. In this way, we obtain the following d equations:

$$\sum_i [(\mathbf{W}^m\mathbf{w}')_{ii} - (\mathbf{w}^*\mathbf{W}^m)_{ii}]\frac{\partial}{\partial X_i}K(\mathbf{W},\mathbf{X}) = 0 , \quad m = 0,1,\ldots,d-1 , \tag{4.11}$$

whose solution is given by

$$\frac{\partial}{\partial X_i}K(\mathbf{W},\mathbf{X}) = 0 , \quad i = 1,\ldots,d . \tag{4.12}$$

Hence, K only depends upon the $\mathbf{W}$ matrix. Eq. (4.10) now reduces to

$$\sum_k \left(W_{ik}\frac{\partial}{\partial W_{jk}} - W_{kj}\frac{\partial}{\partial W_{ki}}\right)K(\mathbf{W}) = 0 , \tag{4.13}$$

where we only have $d(d-1)$ independent equations when i and j run from 1 to d. Let us therefore, restrict ourselves to the $d(d-1)$ equations for which $i\neq j$.

We then replace the d^2 variables W_{ij} by some new variables T_{ij}, defined by

$$T_{ii} = \operatorname{tr}\mathbf{W}^i , \tag{4.14a}$$

$$T_{ij} = W_{ij} , \quad i\neq j . \tag{4.14b}$$

In the variables T_{ij}, Eq. (4.13) can be rewritten in the following form

$$\left[\sum_{k\neq i,j}\left(W_{ik}\frac{\partial}{\partial T_{jk}} - W_{kj}\frac{\partial}{\partial T_{ki}}\right) + (W_{ii} - W_{jj})\frac{\partial}{\partial T_{ji}}\right]K(\mathbf{T}) = 0 , \quad i\neq j , \tag{4.15}$$

which can be shown to be equivalent to

$$\frac{\partial}{\partial T_{ij}}K(\mathbf{T}) = 0 , \quad i\neq j . \tag{4.16}$$

In conclusion, the general solution of Eq. (4.6a) is an arbitrary function in $T_{11},\ldots,T_{dd}$, which are a set of d independent invariants of the $\mathbf{W}$ matrix. We shall henceforth denote the latter by $t_i, i = 1,\ldots,d$.

Let us now turn to the solution of Eq. (4.6b), which can be rewritten as

$$\{\mathbf{w}'\Delta_{\mathbf{w}'}[\mathbf{w}'\Delta_{\mathbf{w}'} + (n-d-1)\mathbf{I}]\}_{ij}K(t_1,\ldots,t_d) = W_{ij}K(t_1,\ldots,t_d) , \tag{4.17}$$

where $\mathbf{W}$ now appears on the right-hand side instead of $\mathbf{w}^*$. In the solution of Eq. (4.17), it is advantageous to choose for the invariants of $\mathbf{W}$ the traces of the compound matrices $\mathbf{W}^{(i)}$,[14]

$$t_i = \operatorname{tr}\mathbf{W}^{(i)} , \quad i = 1,\ldots,d, \tag{4.18}$$

instead of the traces of the powers $\mathbf{W}^i$, as it was done in Eq. (4.14a). In the following, we shall use the fact that $t_1,\ldots,t_d$, as defined in Eq. (4.18), are the coefficients of the characteristic polynomial of $\mathbf{W}$,

$$\det(\mathbf{W} - \lambda\,\mathbf{I}) = \sum_{k=0}^{d} t_{d-k}(-\lambda)^k , \tag{4.19}$$

where we set $t_0 = 1$. For $d = 1$, 2, and 3, they are given by

$$t_1 = \operatorname{tr}\mathbf{W} = W , \tag{4.20a}$$

$$t_1 = \operatorname{tr}\mathbf{W} , \quad t_2 = \det\mathbf{W} , \tag{4.20b}$$

and

$$t_1 = \operatorname{tr}\mathbf{W} , \quad t_2 = \operatorname{tr}\mathbf{W}^{(2)} = \tfrac{1}{2}[(\operatorname{tr}\mathbf{W})^2 - \operatorname{tr}\mathbf{W}^2] , \quad t_3 = \det\mathbf{W} , \tag{4.20c}$$

respectively.

We are interested in those solutions $K(t_1,...,t_d)$ of Eq. (4.17) which are analytic in all the variables w'_{ij} and w^*_{ij}. Since the matrix elements W_{ij} are linear functions in each one of the w'_{ij}'s and w^*_{ij}'s, and the same is true for the invariants t_i with respect to each one of the W_{ij}'s, it is clear that the functions $K(t_1,...,t_d)$ must also be analytic in $t_1,...,t_d$. Let us therefore expand K in powers of the latter as follows:

$$K(t_1,...,t_d)=\sum_{m_1m_2\cdots m_d} a_{m_1m_2\cdots m_d}\left[\prod_{i=1}^{d} t_i^{m_i}\right]. \tag{4.21}$$

It remains to determine the constants $a_{m_1m_2\cdots m_d}$ so that Eqs. (4.7) and (4.17) are satisfied. The former simply imposes that

$$a_{00\ldots0}=1. \tag{4.22}$$

Let us now introduce Eq. (4.21) into Eq. (4.17). To write the detailed expression of the latter, we have to know the action of the operators $(\mathbf{w}'\Delta_{\mathbf{w}'})_{ij}$ and $(\mathbf{w}'\Delta_{\mathbf{w}'}\mathbf{w}'\Delta_{\mathbf{w}'})_{ij}$ upon the invariants t_i. Straightforward differentiations lead to the following relations:

$$\begin{aligned}&(\mathbf{w}'\Delta_{\mathbf{w}'})_{ij}\ \mathrm{tr}\,\mathbf{W}=2W_{ij}\,,\\ &(\mathbf{w}'\Delta_{\mathbf{w}'})_{ij}\ \mathrm{tr}\,\mathbf{W}^{(2)}=2[W_{ij}\ \mathrm{tr}\,\mathbf{W}-(\mathbf{W}^2)_{ij}]\,,\\ &(\mathbf{w}'\Delta_{\mathbf{w}'})_{ij}\ \det\mathbf{W}=2\delta_{ij}\det\mathbf{W}\,,\end{aligned} \tag{4.23}$$

and

$$\begin{aligned}&(\mathbf{w}'\Delta_{\mathbf{w}'}\mathbf{w}'\Delta_{\mathbf{w}'})_{ij}\ \mathrm{tr}\,\mathbf{W}=2(d+1)W_{ij}\,,\\ &(\mathbf{w}'\Delta_{\mathbf{w}'}\mathbf{w}'\Delta_{\mathbf{w}'})_{ij}\ \mathrm{tr}\,\mathbf{W}^{(2)}=2d\,[W_{ij}\ \mathrm{tr}\,\mathbf{W}-(\mathbf{W}^2)_{ij}]\,,\\ &(\mathbf{w}'\Delta_{\mathbf{w}'}\mathbf{w}'\Delta_{\mathbf{w}'})_{ij}\ \det\mathbf{W}=4\delta_{ij}\det\mathbf{W}\,.\end{aligned} \tag{4.24}$$

In the one-dimensional case, Eq. (4.17) becomes

$$\sum_m 2m(2m+n-2)a_m t_1^m=\sum_m a_m t_1^{m+1}. \tag{4.25}$$

By equating equal powers of t_1 on both sides, we obtain a recursion relation for a_m, whose solution, normalized according to Eq. (4.22), is given by

$$a_m=[2^{2m}m!(n/2)_m]^{-1}, \tag{4.26}$$

where $(n/2)_m=n/2(n/2+1)\ldots(n/2+m-1)$ is Pochhammer's symbol. Hence the reproducing kernel is equal to the following expression:

$$K(W)=\sum_m[2^{2m}m!(n/2)_m]^{-1}W^m={}_0F_1\left(\frac{n}{2};\frac{W}{4}\right), \tag{4.27}$$

which coincides with Eq. (5.19b) in I.

In the two-dimensional case, Eq. (4.17) is transformed into the following relation:

$$\begin{aligned}\sum_{m_1m_2} a_{m_1m_2}\big[&2m_2(2m_2+n-3)\delta_{ij}t_1^{m_1}\\ &+2m_1(4m_2+n)W_{ij}t_1^{m_1-1}\\ &+4m_1(m_1-1)(\mathbf{W}^2)_{ij}t_1^{m_1-2}\big]t_2^{m_2}\\ =W_{ij}&\sum_{m_1m_2}a_{m_1m_2}t_1^{m_1}t_2^{m_2}.\end{aligned} \tag{4.28}$$

In this equation, $\mathbf{W}^2$ can be reexpressed in terms of $\mathbf{I}$ and $\mathbf{W}$. Any matrix satisfying its characteristic equation, we indeed obtain from Eq. (4.19) that for $d=2$

$$\mathbf{W}^2=-\mathbf{I}t_2+\mathbf{W}t_1. \tag{4.29}$$

After this transformation, Eq. (4.28) becomes a relation between the independent matrices $\mathbf{I}$ and $\mathbf{W}$. Their coefficients must therefore be set equal to zero. In this way, we obtain the two following recursion relations for $a_{m_1m_2}$:

$$m_2(2m_2+n-3)a_{m_1m_2}=2(m_1+2)(m_1+1)a_{m_1+2,m_2-1}, \tag{4.30a}$$

$$2m_1(2m_1+4m_2+n-2)a_{m_1m_2}=a_{m_1-1,m_2}. \tag{4.30b}$$

The solution of Eq. (4.30b) is given by

$$a_{m_1m_2}=[2^{2m_1}m_1!(n/2+2m_2)_{m_1}]^{-1}a_{0m_2}, \tag{4.31}$$

where a_{0m_2} has to be determined from Eq. (4.30a). By setting $m_1=0$ in the latter, we obtain the following recursion relation for a_{0m_2}:

$$\begin{aligned}2m_2(2m_2+n-3)(4m_2+n-2)(4m_2+n-4)a_{0m_2}\\ =a_{0,m_2-1},\end{aligned} \tag{4.32}$$

whose solution satisfying Eq. (4.22) is given by

$$a_{0m_2}=[2^{4m_2}m_2!([n-1]/2)_{m_2}(n/2)_{2m_2}]^{-1}. \tag{4.33}$$

For the reproducing kernel, we finally obtain

$$\begin{aligned}&K(\mathrm{tr}\,\mathbf{W},\det\mathbf{W})\\ &=\sum_{m_1m_2}\left\{2^{2m_1+4m_2}m_1!m_2!\left(\frac{[n-1]}{2}\right)_{m_2}(n/2)_{m_1+2m_2}\right\}^{-1}\\ &\quad\times(\mathrm{tr}\,\mathbf{W})^{m_1}(\det\mathbf{W})^{m_2},\end{aligned} \tag{4.34a}$$

or equivalently

$$\begin{aligned}&K(\mathrm{tr}\,\mathbf{W},\det\mathbf{W})\\ &=\sum_m\left\{m!\left(\frac{[n-1]}{2}\right)_m\left(\frac{n}{2}\right)_{2m}\right\}^{-1}\\ &\quad\times{}_0F_1\left(\frac{n}{2}+2m;\frac{1}{4}\,\mathrm{tr}\,\mathbf{W}\right)\left(\frac{1}{16}\det\mathbf{W}\right)^m.\end{aligned} \tag{4.34b}$$

In the three-dimensional case, a similar procedure, detailed in Appendix C, leads to the following relation:

$$\begin{aligned}&K(\mathrm{tr}\,\mathbf{W},\mathrm{tr}\,\mathbf{W}^{(2)},\det\mathbf{W})\\ &=\sum_{m_1m_2m_3}(m_1+2m_2+3m_3+n-3)_{m_3}\\ &\quad\times\{2^{2m_1+4m_2+6m_3}m_1!m_2!m_3!\\ &\quad\times([n-2]/2)_{m_3}([n-1]/2)_{m_2+2m_3}\\ &\quad\times(n/2)_{m_1+2m_2+3m_3}\}^{-1}(\mathrm{tr}\,\mathbf{W})^{m_1}\\ &\quad\times(\mathrm{tr}\,\mathbf{W}^{(2)})^{m_2}(\det\mathbf{W})^{m_3}.\end{aligned} \tag{4.35}$$

From the reproducing kernels (4.27), (4.34), and (4.35), it is now trivial to obtain the explicit expression of BGCS in one, two, and three dimensions. For such purpose, we only have to make the following replacement:

$$\mathbf{W}=\mathbf{w}'\mathbf{w}^*\to\mathscr{D}^\dagger\mathbf{w}^*, \tag{4.36}$$

which implies the substitution

$$\mathbf{W}^{(i)}\to(\mathscr{D}^\dagger\mathbf{w}^*)^{(i)}=(\mathscr{D}^\dagger)^{(i)}(\mathbf{w}^*)^{(i)}, \tag{4.37}$$

where in the last step we used Binet–Cauchy theorem on the product of compound matrices.[14]

5. UNITY RESOLUTION FOR THE BARUT-GIRARDELLO COHERENT STATES

In Sec. 3, we proved the (over) completeness of BGCS by showing that they give rise in the collective subspace to a unity resolution relation, expressed in Eq. (3.18). In this section, we shall propose a procedure to determine the weight function $f(\mathbf{w},\mathbf{w}^*)$. We shall prove that $f(\mathbf{w},\mathbf{w}^*)$ is a solution of a system of partial differential equations, that we shall write in full generality for any d value, and we shall then illustrate its solution by explicitly treating the two-dimensional case.

As it was done for the one-dimensional case in Sec. 4 of I, we start with the Hermiticity properties of the $\mathscr{S}p(2d,R)$ generators, given in Eq. (4.4). By introducing the unity resolution relation (3.18) into Eq. (4.4), and by successively taking Eqs. (3.20) and (3.21) into account, Eqs. (4.4a) and (4.4b) are transformed into

$$\int d\mathbf{w}\, d\mathbf{w}^* f(\mathbf{w},\mathbf{w}^*)|\mathbf{w})\mathscr{E}^c_{ij}(\mathbf{w}| = \int d\mathbf{w}\, d\mathbf{w}^* f(\mathbf{w},\mathbf{w}^*)[\mathscr{E}^c_{ji}(\mathbf{w}|]^\dagger(\mathbf{w}| , \tag{5.1a}$$

$$\int d\mathbf{w}\, d\mathbf{w}^* f(\mathbf{w},\mathbf{w}^*)|\mathbf{w})\mathscr{D}^c_{ij}(\mathbf{w}| = \int d\mathbf{w}\, d\mathbf{w}^* f(\mathbf{w},\mathbf{w}^*)[\mathscr{D}^{c\dagger}_{ij}(\mathbf{w}|]^\dagger(\mathbf{w}| , \tag{5.1b}$$

or

$$\int d\mathbf{w}\, d\mathbf{w}^* f(\mathbf{w},\mathbf{w}^*)|\mathbf{w})(\mathbf{w}\Delta_\mathbf{w})_{ij}(\mathbf{w}| = \int d\mathbf{w}\, d\mathbf{w}^* f(\mathbf{w},\mathbf{w}^*)(\mathbf{w}^*\Delta_{\mathbf{w}^*})_{ji}|\mathbf{w})(\mathbf{w}| , \tag{5.2a}$$

$$\int d\mathbf{w}\, d\mathbf{w}^* f(\mathbf{w},\mathbf{w}^*)|\mathbf{w})\{\Delta_\mathbf{w}[\mathbf{w}\Delta_\mathbf{w} + (n-d-1)\mathbf{I}]\}_{ij}(\mathbf{w}| = \int d\mathbf{w}\, d\mathbf{w}^* f(\mathbf{w},\mathbf{w}^*)w^*_{ij}|\mathbf{w})(\mathbf{w}| . \tag{5.2b}$$

We next transfer the differential operators from $|\mathbf{w})$ or $(\mathbf{w}|$ to the weight function $f(\mathbf{w},\mathbf{w}^*)$ by integrating by parts, once on both sides of Eq. (5.2a), and twice on the left-hand side of Eq. (5.2b). This can be done because $|\mathbf{w})$ and $(\mathbf{w}|$ only depend upon w^*_{ij} or w_{ij}, respectively, and the latter are independent variables. We obtain the following system of partial differential equations for $f(\mathbf{w},\mathbf{w}^*)$:

$$(\mathbf{w}\Delta_\mathbf{w})_{ij} f(\mathbf{w},\mathbf{w}^*) = (\mathbf{w}^*\Delta_{\mathbf{w}^*})_{ji} f(\mathbf{w},\mathbf{w}^*) , \tag{5.3a}$$

$$\{\Delta_\mathbf{w}[\mathbf{w}\Delta_\mathbf{w} - (n-d-1)\mathbf{I}]\}_{ij} f(\mathbf{w},\mathbf{w}^*) = w^*_{ij} f(\mathbf{w},\mathbf{w}^*) . \tag{5.3b}$$

At this point, it is important to note the formal analogy existing between these equations and Eqs. (4.6a) and (4.6b) satisfied by the reproducing kernel $K(\mathbf{w}',\mathbf{w}^*)$ of BGCS if in the latter we set $\mathbf{w}' = \mathbf{w}$. As a matter of fact, Eq. (5.3a) is identical with Eq. (4.6a), whereas Eqs. (5.3b) and (4.6b) only differ by a sign change inside the square brackets.

The results of Sec. 4 show that the general solution of Eq. (5.3a) is an arbitrary function in the invariants $t_i = \operatorname{tr} \mathbf{W}^{(i)}$, $i = 1,\ldots,d$, of the matrix $\mathbf{W} = \mathbf{w}\mathbf{w}^*$. We are then left to solve Eq. (5.3b), which can be rewritten as

$$\{\mathbf{w}\Delta_\mathbf{w}[\mathbf{w}\Delta_\mathbf{w} - (n-d-1)\mathbf{I}]\}_{ij} f(t_1,\ldots,t_d) = W_{ij} f(t_1,\ldots,t_d) . \tag{5.4}$$

It differs from the solution of Eq. (4.17) not only in the abovementioned sign change, but also in the additional restrictions on $f(t_1,\ldots,t_d)$ necessary to make it an acceptable weight function: f must be real, positive (except at the boundaries where it may go to zero), and integrable. Moreover, by taking the expectation value of both sides of the unity resolution relation with respect to the vacuum state, we find from Eq. (3.6) that the weight function must also satisfy the normalization condition

$$\int d\mathbf{w}\, d\mathbf{w}^* f(t_1,\ldots,t_d) = 1 . \tag{5.5}$$

In the one-dimensional case, f is a function in the single real variable $t_1 = W = |w|^2$, whose range is given by $0 \leqslant t_1 \leqslant +\infty$. Since

$$w\Delta_w f(t_1) = 2t_1 \frac{d}{dt_1} f(t_1) , \tag{5.6}$$

Eq. (5.4) becomes

$$\left[4t_1 \frac{d^2}{dt_1^2} - 2(n-4)\frac{d}{dt_1} - 1\right] f(t_1) = 0 . \tag{5.7}$$

It was shown in I [cf. Eq. (4.17) in I] that the solution of this equation, which goes to zero when $t_1 \to \infty$, is given by

$$f(t_1) = A t_1^{(n-2)/4} K_{(n-2)/2}(\sqrt{t_1}) , \tag{5.8}$$

where K is a modified Bessel function, and A is some normalization constant determined by Eq. (5.5).

Let us now turn to the two-dimensional case. The weight function is then a function in the two real variables

$$t_1 = \operatorname{tr} \mathbf{W} = \sum_{ij} |w_{ij}|^2 ,$$

and

$$t_2 = \det \mathbf{W} = |\det \mathbf{w}|^2 . \tag{5.9}$$

Let us first determine the domain of this function. Since $\mathbf{W}$ is a Hermitian, positive definite matrix, its eigenvalues λ_1,λ_2 are real and lie in the range

$$0 \leqslant \lambda_1,\lambda_2 \leqslant +\infty . \tag{5.10}$$

In terms of these two independent variables, t_1 and t_2 can be expressed as

$$t_1 = \lambda_1 + \lambda_2 , \quad t_2 = \lambda_1\lambda_2 . \tag{5.11}$$

We now question what is the domain of variation of t_1 and t_2 when λ_1 and λ_2 are real and satisfy Eq. (5.10). To answer this, let us consider the characteristic equation

$$\lambda^2 - t_1\lambda + t_2 = 0 , \tag{5.12}$$

whose solutions are λ_1 and λ_2. The reality condition for the latter can be written as

$$t_1^2 - 4t_2 \geqslant 0 . \tag{5.13}$$

When it is fulfilled, λ_1 and λ_2 are given by

$$\lambda_{1,2} = \tfrac{1}{2}[t_1 \pm (t_1^2 - 4t_2)^{1/2}] . \tag{5.14}$$

It is clear from Eqs. (5.11) and (5.14) that t_1 and t_2 are nonnegative whenever λ_1 and λ_2 are so, and vice versa. The domain of variation of t_1 and t_2 is therefore given by

$$0 \leqslant t_1 \leqslant +\infty\,, \quad 0 \leqslant t_2 \leqslant \frac{t_1^2}{4}\,. \tag{5.15}$$

Let us now rewrite Eq. (5.4) for $d = 2$ in terms of the variables t_1 and t_2. For such purpose, we start with the relation

$$(\mathbf{w}\Delta_{\mathbf{w}})_{ij} f(t_1,t_2) = 2\Big(W_{ij}\frac{\partial}{\partial t_1} + \delta_{ij} t_2 \frac{\partial}{\partial t_2}\Big) f(t_1,t_2)\,, \tag{5.16}$$

which directly follows from Eq. (4.23). By repeated applications of Eq. (5.16) and by taking Eqs. (4.23) and (4.24) into account, it is straightforward to transform Eq. (5.4) into the following equation:

$$\begin{aligned}\Big[&2\delta_{ij}\Big(2t_2^2\frac{\partial^2}{\partial t_2^2} - (n-5)t_2\frac{\partial}{\partial t_2}\Big)\\ &+ 2W_{ij}\Big(4t_2\frac{\partial^2}{\partial t_1 \partial t_2} - (n-6)\frac{\partial}{\partial t_1}\Big)\\ &+ 4(\mathbf{W}^2)_{ij}\frac{\partial^2}{\partial t_1^2}\Big] f(t_1,t_2) = W_{ij} f(t_1,t_2)\,,\end{aligned} \tag{5.17}$$

where $\mathbf{W}^2$ can be reexpressed in terms of $\mathbf{I}$ and $\mathbf{W}$ through Eq. (4.29). Equation (5.17) is then converted into a relation between the independent matrices $\mathbf{I}$ and $\mathbf{W}$. By equating their coefficients to zero, we obtain the following system of two partial differential equations for $f(t_1,t_2)$:

$$\Big[2\frac{\partial^2}{\partial t_1^2} - 2t_2\frac{\partial^2}{\partial t_2^2} + (n-5)\frac{\partial}{\partial t_2}\Big] f(t_1,t_2) = 0\,, \tag{5.18a}$$

$$\Big[4t_1\frac{\partial^2}{\partial t_1^2} + 8t_2\frac{\partial^2}{\partial t_1 \partial t_2} - 2(n-6)\frac{\partial}{\partial t_1} - 1\Big] f(t_1,t_2) = 0\,. \tag{5.18b}$$

Equation (5.18a) being separable with respect to t_1 and t_2 is easy to solve. It is shown in Appendix D that it has a solution integrable in the domain (5.15), given by

$$\begin{aligned}f(t_1,t_2) = &\int_0^\infty d\lambda\, \chi(\lambda)\\ &\times \exp(-\tfrac{1}{2}\lambda t_1)(\lambda\sqrt{t_2})^{(n-3)/2} K_{(n-3)/2}(\lambda\sqrt{t_2})\,,\end{aligned} \tag{5.19}$$

where $\chi(\lambda)$ is a function in λ, only subject to the condition that the integral in Eq. (5.19) does exist. This function χ is now determined by imposing that $f(t_1,t_2)$ is also a solution of Eq. (5.18b). For such purpose, it is convenient to use the following relation:

$$\begin{aligned}&\Big[4t_1\frac{\partial^2}{\partial t_1^2} + 8t_2\frac{\partial^2}{\partial t_1\partial t_2} - 2(n-6)\frac{\partial}{\partial t_1} - 1\Big]\\ &\quad\times \exp\Big(-\frac{1}{2}\lambda t_1\Big)(\lambda\sqrt{t_2})^{(n-3)/2} K_{(n-3)/2}(\lambda\sqrt{t_2})\\ &= \Big[-2\lambda^2\frac{\partial}{\partial\lambda} + (n-6)\lambda - 1\Big]\\ &\quad\times \exp\Big(-\frac{1}{2}\lambda t_1\Big)(\lambda\sqrt{t_2})^{(n-3)/2} K_{(n-3)/2}(\lambda\sqrt{t_2})\,,\end{aligned} \tag{5.20}$$

proved in Appendix D. By integrating by parts and assuming that

$$\chi(0) = \chi(+\infty) = 0\,, \tag{5.21}$$

Eq. (5.18b) can be converted into a differential equation for $\chi(\lambda)$,

$$\Big[2\lambda^2\frac{d}{d\lambda} + (n-2)\lambda - 1\Big]\chi(\lambda) = 0\,. \tag{5.22}$$

Its solution is given by

$$\chi(\lambda) = A\lambda^{-(n-2)/2}\exp[-(2\lambda)^{-1}]\,, \tag{5.23}$$

where A is a normalization constant, and it satisfies Eq. (5.21) as it should be. We have therefore proved that the system of partial differential equations (5.18) has a solution satisfying the appropriate asymptotic conditions and that this solution admits the following integral representation:

$$\begin{aligned}f(t_1,t_2) = A&\int_0^\infty d\lambda\,\lambda^{-(n-2)/2}\\ &\times\exp\Big[-\frac{1}{2}(\lambda^{-1} + \lambda t_1)\Big](\lambda\sqrt{t_2})^{(n-3)/2}\\ &\times K_{(n-3)/2}(\lambda\sqrt{t_2})\,,\end{aligned} \tag{5.24}$$

where A is determined by Eq. (5.5).

It now remains to perform the integration in Eq. (5.24) to get an explicit expression for $f(t_1,t_2)$. This can be done by using the following expansion:

$$\begin{aligned}K_{(n-3)/2}(\lambda\sqrt{t_2}) &= \pi^{1/2}(2\lambda\sqrt{t_2})^{-1/2}\exp(-\lambda\sqrt{t_2})\\ &\times\sum_{k=0}^{m}\Gamma\Big[\frac{(n-2)}{2} + k\Big]\\ &\times\Big\{k!\,\Gamma\Big[\frac{(n-2)}{2} - k\Big]\Big\}^{-1}(2\lambda\sqrt{t_2})^{-k}\,,\end{aligned} \tag{5.25}$$

which is a consequence of Eqs. 8.451.6 and 8.468 of Ref. 15. For even n values, the right-hand side of Eq. (5.25) contains a finite sum, m being given by

$$m = (n-4)/2\,, \tag{5.26}$$

whereas for odd n values, it contains an asymptotic expansion valid for large values of the argument $\lambda\sqrt{t_2}$. When Eq. (5.25) is introduced into Eq. (5.24), the latter becomes

$$\begin{aligned}f(t_1,t_2) = A\,(2\pi)^{1/2}t_2^{(n-4)/4}&\sum_{k=0}^{m}\Gamma\Big[\frac{(n-2)}{2} + k\Big]\\ &\times\Big\{k!\,\Gamma\Big[\frac{(n-2)}{2} - k\Big]\Big\}^{-1}\\ &\times[(t_1 + 2\sqrt{t_2})(4t_2)^{-1}]^{k/2}K_k([t_1 + 2\sqrt{t_2}]^{1/2})\,,\end{aligned} \tag{5.27}$$

by successively applying Eq. 3.471.9 of Ref. 15,

$$\int_0^\infty x^{\nu-1}\exp\Big(-\frac{\beta}{x} - \gamma x\Big)dx = 2\Big(\frac{\beta}{\gamma}\Big)^{\nu/2}K_\nu(2\sqrt{\beta\gamma})\,,$$
$$\operatorname{Re}\beta > 0\,, \quad \operatorname{Re}\gamma > 0\,, \tag{5.28}$$

with $\nu = -k$, $\beta = \frac{1}{2}$, and $\gamma = \frac{1}{2}t_1 + \sqrt{t_2}$, and Eq. 8.486.16 of the same reference,

$$K_{-\nu}(z) = K_\nu(z)\,. \tag{5.29}$$

We have therefore obtained a representation of the weight function in the form of a finite sum for even n values, and of an asymptotic expansion valid for large values of t_2 for odd n values. In the latter case, it is also possible to get a representation valid for small values of t_2. From Eq. (5.24), we note for instance that for $t_2 = 0$, $f(t_1,t_2)$ reduces to

$$f(t_1,0) = A\,2^{(n-5)/2}\Gamma\left[\frac{(n-3)}{2}\right]\int_0^\infty d\lambda\lambda^{-(n-2)/2} \times \exp\left[-\frac{1}{2}(\lambda^{-1}+\lambda t_1)\right]. \tag{5.30}$$

By using Eqs. (5.28) and (5.29) with $\nu = -(n-4)/2, \beta = \frac{1}{2}$, and $\gamma = t_1/2$, Eq. (5.30) is transformed into the following expansion:

$$f(t_1,0) = A\,2^{(n-3)/2}\Gamma[(n-3)/2]t_1^{(n-4)/4}K_{(n-4)/2}(\sqrt{t_1}), \tag{5.31}$$

valid for any n value. The same procedure would enable us to obtain an approximate form of the weight function valid for small values of t_2 by keeping the first few terms in the expansion of $K_{(n-3)/2}(\lambda\sqrt{t_2})$ into powers of $\lambda\sqrt{t_2}$.

6. HOLSTEIN–PRIMAKOFF REPRESENTATION IN TERMS OF A THIRD CLASS OF COHERENT STATES

In the present section, we wish to introduce a third class of coherent states, which can be put into one-to-one correspondence with the standard coherent states,[16] and lead to HP representation.

In Secs. 2 and 3, we considered two discrete, nonorthogonal basis, namely the oscillator basis and its dual. Instead of them, we could use as well an intermediate orthogonal basis, obtained from the oscillator basis by the standard orthonormalization procedure. Its elements (for which we utilize a curly bracket instead of an angular one) are defined by

$$|\mathbf{N}\} = \sum_{\mathbf{N}'}|\mathbf{N}'\rangle(\mathbf{M}^{-1/2})_{\mathbf{N}',\mathbf{N}} = \sum_{\mathbf{N}'}|\mathbf{N}')(\mathbf{M}^{1/2})_{\mathbf{N}',\mathbf{N}}, \tag{6.1}$$

where $\mathbf{M}^{1/2}$ denotes the square root of the Hermitian, positive definite matrix $\mathbf{M}$, and $\mathbf{M}^{-1/2}$ the inverse of $\mathbf{M}^{1/2}$. The states $|\mathbf{N}\}$ satisfy the two following relations:

$$\{\mathbf{N}'|\mathbf{N}\} = \delta_{\mathbf{N}',\mathbf{N}}, \tag{6.2}$$

and

$$\sum_{\mathbf{N}}|\mathbf{N}\}\{\mathbf{N}| = I_c. \tag{6.3}$$

By analogy with Eqs. (2.10) and (3.28), let us now introduce the states $|\mathbf{v}\}$, depending upon some complex parameters $v_{ij} = v_{ji}$, $i,j = 1,...,d$, and defined by

$$|\mathbf{v}\} = \sum_{\mathbf{N}} F_{\mathbf{N}}(\mathbf{v}^*)|\mathbf{N}\}, \tag{6.4}$$

where $F_{\mathbf{N}}(\mathbf{v}^*)$ is given by Eq. (2.11) with $\mathbf{u}$ replaced by $\mathbf{v}$. As a direct consequence of Eqs. (6.2) and (2.11), the overlap of two such states, characterized by some parameters $\mathbf{v}$ and $\mathbf{v}'$, respectively, is given by

$$\{\mathbf{v}'|\mathbf{v}\} = \sum_{\mathbf{N}} F_{\mathbf{N}}(\mathbf{v}')F_{\mathbf{N}}(\mathbf{v}^*) = \exp\left\{\sum_{i\leqslant j}[(1+\delta_{ij})^{-1/2}v'_{ij}][(1+\delta_{ij})^{-1/2}v^*_{ij}]\right\}. \tag{6.5}$$

It therefore reduces to the reproducing kernel of a Bargmann–Hilbert space[13] of analytic functions in the ν variables $(1+\delta_{ij})^{-1/2}v_{ij}$, $1\leqslant i\leqslant j\leqslant d$. In matrix notation, Eq. (6.5) can be rewritten as

$$\{\mathbf{v}'|\mathbf{v}\} = \exp(\tfrac{1}{2}\,\mathrm{tr}\,\mathbf{v}'\mathbf{v}^*). \tag{6.6}$$

From Eq. (3.8), we also note that

$$\{\mathbf{v}'|\mathbf{v}\} = \langle\mathbf{v}'|\mathbf{v}). \tag{6.7}$$

It is now trivial to show that when each of the parameters $(1+\delta_{ij})^{-1/2}v_{ij}$, $i\leqslant j$, runs over the whole complex plane, the set of states $|\mathbf{v}\}$ form a continuous basis of the collective space. For such purpose, we shall prove that they give rise to the following unity resolution relation:

$$\int d\mu(\mathbf{v})|\mathbf{v}\}\{\mathbf{v}| = I_c, \tag{6.8}$$

where

$$d\mu(\mathbf{v}) = \prod_{i\leqslant j} d\mu[(1+\delta_{ij})^{-1/2}v_{ij}], \tag{6.9}$$

and $d\mu[(1+\delta_{ij})^{-1/2}v_{ij}]$ is the standard Bargmann measure defined in Eq. (3.15). By taking the matrix element of both sides of Eq. (6.8) between the bra $\{\mathbf{N}'|$ and the ket $|\mathbf{N}\}$, we indeed obtain the relation

$$\int d\mu(\mathbf{v})F_{\mathbf{N}'}(\mathbf{v}^*)F_{\mathbf{N}}(\mathbf{v}) = \delta_{\mathbf{N}',\mathbf{N}}. \tag{6.10}$$

Equation (6.10) is identically satisfied for any $\mathbf{N}$ and $\mathbf{N}'$ since $F_{\mathbf{N}}(\mathbf{v})$ is just the Bargmann representation of the boson state

$$\prod_{i\leqslant j}(N_{ij}!)^{-1/2}(a^\dagger_{ij})^{N_{ij}}|0) = F_{\mathbf{N}}(\bar{\mathbf{a}}^\dagger)|0), \tag{6.11}$$

built from ν independent boson creation operators $a^\dagger_{ij} = a^\dagger_{ji}$, $i,j = 1,...,d$, or the associated non-normalized operators $\bar{a}^\dagger_{ij} = \bar{a}^\dagger_{ji} = (1+\delta_{ij})^{1/2}a^\dagger_{ij}$, $i,j = 1,...,d$, defined in Eq. (3.13) of II.

We conclude that the third class of CS $|\mathbf{v}\}$ introduced above defines an alternative representation of $\mathscr{Sp}_c(2d,R)$, that we shall distinguish from the remaining ones by an inverted caret. In such a representation, the oscillator basis states, for instance, are represented by

$$\check{\bar{\phi}}(\mathbf{v}) = \{\mathbf{v}|\mathbf{N}\rangle = \sum_{\mathbf{N}'}F_{\mathbf{N}'}(\mathbf{v})(\mathbf{M}^{1/2})_{\mathbf{N}',\mathbf{N}}. \tag{6.12}$$

We now wish to establish a link between this representation and HP one. For such purpose, we note that the orthonormal states $|\mathbf{N}\}$ can be put into one-to-one correspondence with the boson states defined in Eq. (6.11). From Eqs. (6.4), (6.5), and (6.6), it is clear that in such a mapping, the CS $|\mathbf{v}\}$ are mapped onto the standard CS

$$\sum_{\mathbf{N}}F_{\mathbf{N}}(\mathbf{v}^*)F_{\mathbf{N}}(\bar{\mathbf{a}}^\dagger)|0) = \exp\left(\frac{1}{2}\,\mathrm{tr}\,\mathbf{v}^*\bar{\mathbf{a}}^\dagger\right)|0), \tag{6.13}$$

in whose representation $\bar{a}^\dagger_{ij}$ and $\bar{a}_{ij}$ are, respectively, represented by v_{ij} and $\Delta v_{ij} = (1+\delta_{ij})\partial/\partial v_{ij}$. Hence, by replacing v_{ij} and Δv_{ij} by $\bar{a}^\dagger_{ij}$ and $\bar{a}_{ij}$ in $\check{\mathscr{D}}^{c\dagger}$, $\check{\mathscr{D}}^c$, and $\check{\mathscr{E}}^c$, we obtain a boson representation of $\mathscr{Sp}_c(2d,R)$, which is nothing else than HP representation, since the mapping between $|\mathbf{N}\}$ and the states defined in Eq. (6.11) preserves scalar products.

We shall not try to implement this procedure for getting the HP representation of $\mathscr{Sp}_c(2d,R)$ since the latter was already found in II. We shall instead take advantage of the results previously proved to directly write the representation of the $\mathscr{Sp}_c(2d,R)$ generators in the basis $|\mathbf{v}\}$ as follows:

$$\check{\mathscr{D}}^{c\dagger} = [\check{\mathscr{C}}^c + (n-d-1)\mathbf{I}]^{1/2}\mathbf{v}, \tag{6.14a}$$

$$\check{\mathscr{D}}^{c} = \Delta_{\mathbf{v}}[\check{\mathscr{C}}^c + (n-d-1)\mathbf{I}]^{1/2}, \tag{6.14b}$$

$$\check{\mathscr{E}}^{c} = \check{\mathscr{C}}^c + (n/2)\mathbf{I}, \quad \check{\mathscr{C}}^c = \mathbf{v}\Delta_{\mathbf{v}}. \tag{6.14c}$$

As explained in II, the square root on the right-hand side of Eqs. (6.14a) and (6.14b) must be understood as a compact form for a finite expansion into powers of $\check{\mathscr{C}}^c$.

Before concluding this section, one more remark is in order. We can also establish a mapping between the oscillator basis states $|\mathbf{N}\rangle$ or the dual basis ones $|\mathbf{N})$ and the boson states defined in Eq. (6.11). The PCS or BGCS are then mapped onto the standard coherent states. Hence, by substituting $\bar{a}_{ij}^\dagger$ and $\bar{a}_{ij}$ for u_{ij} and Δu_{ij} in $\hat{\mathscr{D}}^{c\dagger}$, $\hat{\mathscr{D}}^c$, and $\hat{\mathscr{E}}^c$, or for w_{ij} and Δw_{ij} in $\mathscr{D}^{c\dagger}$, $\mathscr{D}^c$, and $\mathscr{E}^c$, we obtain two additional boson representations of $\mathscr{S}\!\mathscr{p}_c(2d,R)$. Both of them are generalized Dyson representations[11] since the Hermiticity properties of the generators are not preserved under the corresponding nonunitary mappings. This procedure was actually implemented in II for the BGCS representation.

7. CONCLUSION

In the present paper, we considered three different CS representations of the dynamical group $\mathscr{S}\!\mathscr{p}_c(2d,R)$ of microscopic collective states, when the latter are assumed to be scalar under O(n). They, respectively, correspond to the PCS already studied in Ref. 6, to a generalization of some CS proposed by Barut and Girardello for Sp(2,R), and to a third, new class of CS. The major merit of their joint study is that we now have at our disposal a unified treatment of both the BG and boson representations considered in I and II and of the work of Ref. 6.

Various questions may be raised regarding these three classes of CS. First we may ask what is their respective usefulness for practical purposes. This will of course depend upon the kind of application one has in mind. From the study carried out in the present paper, it is clear that each of the three classes shares one interesting property with the standard Glauber CS: PCS are obtained by acting with some group element upon a fixed vector in the vector space of an irreducible unitary representation, BGCS are eigenvectors of some lowering generators, and the third class CS lead to a unity resolution which makes use of Bargmann measure. Moreover, each one of them may also be considered as a generating function for some basis states: those of the oscillator basis, of its dual or of an intermediate orthogonal basis.

We may therefore answer the question in the following way. PCS may be useful in the calculation of some operator matrix elements between oscillator basis states, both owing to their generating function property and to the simple form assumed by some operators, such as the $\mathscr{S}\!\mathscr{p}_c(2d,R)$ generators, in PCS representation. Since the oscillator basis states have a simple form in the BGCS representation, the latter may be well suited to the construction of collective states satisfying some definite properties. An example of such a construction was actually given in II for the collective states classified according to the group chain $\mathscr{S}\!\mathscr{p}_c(2d,R) \supset \mathscr{U}_c(d)$. Finally, as it was mentioned in Sec. 6, the third class CS may provide a convenient starting point to derive the HP representation.

Second, we may ask if the three classes of CS considered in the present paper, and applicable with slight modifications to closed-shell nuclei, have a counterpart for open-shell nuclei, whose collective states transform under an IR $(\lambda_1\lambda_2\lambda_3)$ of O(n) for which $\lambda_1,\lambda_2,\lambda_3$ are not all equal. In a forthcoming paper, we plan to show that such an extension does actually exist.

APPENDIX A: PERELOMOV COHERENT STATE REPRESENTATION OF THE $\mathscr{S}\!\mathscr{p}_c(2d,R)$ GENERATORS

In this appendix, we wish to prove Eq. (2.16). Starting with the representation of $\mathscr{D}_{ij}$, we successively obtain

$$\begin{aligned}\hat{\mathscr{D}}_{ij}^{c}\langle\mathbf{u}|\psi\rangle &= \langle 0|\exp(\tfrac{1}{2}\operatorname{tr}\mathbf{u}\mathscr{D})\mathscr{D}_{ij}|\psi\rangle \\ &= \Delta_{u_{ij}}\langle 0|\exp(\tfrac{1}{2}\operatorname{tr}\mathbf{u}\mathscr{D})|\psi\rangle = \Delta_{u_{ij}}\langle\mathbf{u}|\psi\rangle,\end{aligned} \tag{A1}$$

showing that Eq. (2.16b) directly follows from the definition of PCS given in Eq. (2.2).

To prove Eqs. (2.16a) and (2.16c), we first note that for any operator X the following identity is satisfied

$$\begin{aligned}\hat{X}^{c}\langle\mathbf{u}|\psi\rangle &= \langle 0|[\exp(\tfrac{1}{2}\operatorname{tr}\mathbf{u}\mathscr{D})X \\ &\quad\times\exp(-\tfrac{1}{2}\operatorname{tr}\mathbf{u}\mathscr{D})]\exp(\tfrac{1}{2}\operatorname{tr}\mathbf{u}\mathscr{D})|\psi\rangle.\end{aligned} \tag{A2}$$

We then apply Baker–Campbell–Hausdorff formula

$$\begin{aligned}&\exp(Y)X\exp(-Y) \\ &= X + \sum_{m=1}^{\infty}(m!)^{-1}[Y,[Y,..,[Y,X]...]]_m\end{aligned} \tag{A3}$$

to the right-hand side of Eq. (A2) with $Y = \frac{1}{2}\operatorname{tr}\mathbf{u}\mathscr{D}$.

For the representation of $\mathscr{E}_{ij}$, we get in this way the relation

$$\hat{\mathscr{E}}_{ij}^{c}\langle\mathbf{u}|\psi\rangle = \langle 0|[\mathscr{E}_{ij} + (\mathbf{u}\mathscr{D})_{ij}]\exp(\tfrac{1}{2}\operatorname{tr}\mathbf{u}\mathscr{D})|\psi\rangle, \tag{A4}$$

which leads to Eq. (2.16c) by replacing $\mathscr{D}$ by its PCS representation, and by noting that $\mathscr{E}_{ij}$ reduces to $(n/2)\delta_{ij}$ when acting upon the bra $\langle 0|$.

Similarly, for the representation of $\mathscr{D}_{ij}^\dagger$, we obtain

$$\begin{aligned}\hat{\mathscr{D}}_{ij}^{c\dagger}\langle\mathbf{u}|\psi\rangle &= \langle 0|[\mathscr{D}_{ij}^\dagger + (\mathscr{E}\mathbf{u})_{ij} + (\mathscr{E}\mathbf{u})_{ji} \\ &\quad + (\mathbf{u}\mathscr{D}\mathbf{u})_{ij}]\exp(\tfrac{1}{2}\operatorname{tr}\mathbf{u}\mathscr{D})|\psi\rangle.\end{aligned} \tag{A5}$$

When acting upon the bra $\langle 0|$, $\mathscr{D}_{ij}^\dagger$ gives zero, and $\mathscr{E}$ reduces to $(n/2)\mathbf{I}$. The last term $(\mathbf{u}\mathscr{D}\mathbf{u})_{ij}$ between the brackets in Eq. (A5) can be replaced by

$$\begin{aligned}\sum_{kl} u_{ik}u_{lj}\Delta_{u_{kl}} &= \sum_{kl} u_{ik}\Delta_{u_{kl}}u_{lj} - (d+1)u_{ij} \\ &= [\hat{\mathscr{C}}^c\mathbf{u} - (d+1)\mathbf{u}]_{ij},\end{aligned} \tag{A6}$$

thus completing the proof of Eq. (2.16a).

APPENDIX B: SOLUTION OF EQUATION (4.6a)

The purpose of this appendix is to point out some details overlooked in the solution of Eq. (4.6a), as given in Sec. 4.

To start with, let us check that the variables W_{ij} and X_i, respectively, defined in Eqs. (4.8) and (4.9), are functionally independent. For such purpose, we have to prove that their Jacobian with respect to the old variables w'_{ij}, w^*_{ij}, $i \leqslant j$, is different from zero. Since the term of the Jacobian, whose degree in each of the variables $w^*_{11}, w^*_{22},...,w^*_{dd}, w'_{11},$

$w'_{22},\ldots,w'_{d-1,d-1}$ is the highest one compatible with the degrees of the variables preceding it in the list, is equal to

$$\left[\prod_{i=1}^{d}(w^*_{ii})^{d-i+1}\right]\left[\prod_{j=1}^{d-1}(w'_{jj})^{d-j}\right], \tag{B1}$$

and is clearly different from zero, this also holds true for the Jacobian itself.

We next turn to the solution of Eq. (4.11). The latter can be considered as a system of d linear equations in the d unknowns $\partial K(\mathbf{W},\mathbf{X})/\partial X_i$, $i = 1,\ldots,d$. In the determinant of its coefficients, the term of highest degree in the w'_{ij} variables is equal to $\det|(\mathbf{W}^{i-1}\mathbf{w}')_{jj}|$, where i and j are the row and column indices, respectively. This term is clearly different from zero since for the w'_{ij} and w^*_{ij} particular values

$$w'_{ij} = w'\delta_{ij},$$
$$w^*_{ij} = w^*_{ii}\delta_{ij}, \quad w^*_{11} \neq w^*_{22} \neq \cdots \neq w^*_{dd}, \tag{B2}$$

it is straightforward to show that

$$|\det(\mathbf{W}^{i-1}\mathbf{w}')_{jj}| = (w')^{d(d+1)/2}\prod_{i>j}(w^*_{ii} - w^*_{jj}) \neq 0. \tag{B3}$$

Therefore the determinant of the system is different from zero; hence its solution is given by Eq. (4.12), as stated in Sec. 4.

We now consider the change of variables defined in Eq. (4.14), and prove that its Jacobian is different from zero. The latter can indeed be shown to be equal to the following expression:

$$\frac{\partial(T_{11},\ldots,T_{dd})}{\partial(W_{11},\ldots,W_{dd})} = \det|i(\mathbf{W}^{i-1})_{jj}|, \tag{B4}$$

which reduces to

$$\frac{\partial(T_{11},\ldots,T_{dd})}{\partial(W_{11},\ldots,W_{dd})} = d!\prod_{i>j}(W_{ii} - W_{jj}) \neq 0, \tag{B5}$$

when

$$W_{ij} = W_{ii}\delta_{ij}, \quad W_{11} \neq W_{22} \neq \cdots \neq W_{dd}. \tag{B6}$$

As a final point, we prove that the solution of Eq. (4.15) is given by Eq. (4.16). We note that the derivatives of $K(\mathbf{T})$ with respect to T_{kk}, $k = 1,\ldots,d$, do not appear in Eq. (4.15). The latter therefore, represents a system of $d(d-1)$ linear equations in the $d(d-1)$ unknowns $\partial K(\mathbf{T})/\partial T_{kl}$, $k \neq l$. In the determinant of its coefficients, the term of highest degree in $W_{11},\ldots,W_{dd}$ is given by $\Pi_{i\neq j}(W_{ii} - W_{jj})$, and is clearly different from zero. Hence this also holds true for the determinant itself, thus completing the proof of Eq. (4.16).

APPENDIX C: REPRODUCING KERNEL OF BARUT–GIRARDELLO COHERENT STATES IN THREE DIMENSIONS

The purpose of the present appendix is to prove the validity of Eq. (4.35) for the reproducing kernel $K(t_1,t_2,t_3)$.

By taking Eqs. (4.23) and (4.24) into account, Eq. (4.17) assumes the following form for $d = 3$:

$$\begin{aligned}
\sum_{m_1m_2m_3} a_{m_1m_2m_3}\{&\delta_{ij}\left[2m_3(2m_3+n-4)t_1^{m_1}t_2^{m_2}\right]\\
&+ W_{ij}\left[2m_1(4m_3+n)t_1^{m_1-1}t_2^{m_2}\right.\\
&\left.+ 2m_2(4m_3+n-1)t_1^{m_1+1}t_2^{m_2-1}\right]\\
&+ (\mathbf{W}^2)_{ij}\left[4m_1(m_1-1)t_1^{m_1-2}t_2^{m_2}\right.\\
&+ 2m_2(4m_1-4m_3-n+1)t_1^{m_1}t_2^{m_2-1}\\
&\left.+ 4m_2(m_2-1)t_1^{m_1+2}t_2^{m_2-2}\right]\\
&+ (\mathbf{W}^3)_{ij}\left[-8m_1m_2t_1^{m_1-1}t_2^{m_2-1}\right.\\
&\left.- 8m_2(m_2-1)t_1^{m_1+1}t_2^{m_2-2}\right]\\
&+ (\mathbf{W}^4)_{ij}\left[4m_2(m_2-1)t_1^{m_1}t_2^{m_2-2}\right]\}t_3^{m_3}\\
= W_{ij}&\sum_{m_1m_2m_3} a_{m_1m_2m_3}t_1^{m_1}t_2^{m_2}t_3^{m_3}.
\end{aligned} \tag{C1}$$

In this equation, $\mathbf{W}^3$ and $\mathbf{W}^4$ can be reexpressed in terms of $\mathbf{I}$, $\mathbf{W}$, and $\mathbf{W}^2$. From Eq. (4.19) and the fact that any matrix satisfies its characteristic equation, we indeed obtain the relation

$$\mathbf{W}^3 = \mathbf{I}t_3 - \mathbf{W}t_2 + \mathbf{W}^2t_1, \tag{C2}$$

from which it follows that

$$\mathbf{W}^4 = \mathbf{I}t_1t_3 + \mathbf{W}(t_3 - t_1t_2) + \mathbf{W}^2(t_1^2 - t_2). \tag{C3}$$

When we introduce Eqs. (C2) and (C3) into Eq. (C1), we get a relation between the independent matrices $\mathbf{I}$, $\mathbf{W}$, and $\mathbf{W}^2$. By equating their coefficients to zero, we obtain the following set of three recursion relations for $a_{m_1m_2m_3}$:

$$\begin{aligned}
&m_3(2m_3+n-4)a_{m_1m_2m_3}\\
&\quad - 4(m_1+1)(m_2+1)a_{m_1+1,m_2+1,m_3-1}\\
&\quad - 2(m_2+1)(m_2+2)a_{m_1-1,m_2+2,m_3-1} = 0,
\end{aligned} \tag{C4a}$$

$$\begin{aligned}
&2(m_1+1)(4m_2+4m_3+n)a_{m_1+1,m_2,m_3}\\
&\quad + 2(m_2+1)(2m_2+4m_3+n-1)a_{m_1-1,m_2+1,m_3}\\
&\quad + 4(m_2+1)(m_2+2)a_{m_1,m_2+2,m_3-1} = a_{m_1m_2m_3},
\end{aligned} \tag{C4b}$$

$$\begin{aligned}
&2(m_1+1)(m_1+2)a_{m_1+2,m_2,m_3}\\
&\quad - (m_2+1)(2m_2+4m_3+n-1)a_{m_1,m_2+1,m_3} = 0.
\end{aligned} \tag{C4c}$$

From Eq. (C4c), we can express $a_{m_1m_2m_3}$ in terms of $a_{m_1+2m_2,0,m_3}$ as follows:

$$\begin{aligned}
a_{m_1m_2m_3} &= (m_1+2m_2)!([n-1]/2)_{2m_3}\\
&\quad \times\{m_1!m_2!([n-1]/2)_{m_2+2m_3}\}^{-1}a_{m_1+2m_2,0,m_3}.
\end{aligned} \tag{C5}$$

When we introduce Eq. (C5) into Eq. (C4a), the latter is transformed into

$$\begin{aligned}
&m_3(2m_3+n-4)(4m_3+n-3)(4m_3+n-5)a_{m_1+2m_2,0,m_3}\\
&\quad = 8(m_1+2m_2+1)(m_1+2m_2+2)(m_1+2m_2+3)\\
&\quad\quad \times(m_1+2m_2+4m_3+n-3)a_{m_1+2m_2+3,0,m_3-1},
\end{aligned} \tag{C6}$$

and its solution is given by

$$a_{m_1+2m_2,0,m_3}$$
$$=(m_1+2m_2+3m_3)!(m_1+2m_2+3m_3+n-3)_{m_3}$$
$$\times\{(m_1+2m_2)!m_3!([n-2]/2)_{m_3}$$
$$\times([n-1]/2)_{2m_3}\}^{-1}a_{m_1+2m_2+3m_3,0,0}\,. \tag{C7}$$

We next combine Eq. (C5) with Eq. (C7) to obtain the following result:

$$a_{m_1m_2m_3}$$
$$=(m_1+2m_2+3m_3)!(m_1+2m_2+3m_3+n-3)_{m_3}$$
$$\times\{m_1!m_2!m_3!([n-2]/2)_{m_3}$$
$$\times([n-1]/2)_{m_2+2m_3}\}^{-1}a_{m_1+2m_2+3m_3,0,0}\,, \tag{C8}$$

and we introduce the latter into Eq. (C4b). We then get the following equation:

$$2(m_1+2m_2+3m_3+1)(2m_1+4m_2+6m_3+n)$$
$$\times a_{m_1+2m_2+3m_3+1,0,0}=a_{m_1+2m_2+3m_3,0,0}\,, \tag{C9}$$

whose solution is given by

$$a_{m_1+2m_2+3m_3,0,0}$$
$$=\{2^{2m_1+4m_2+6m_3}(m_1+2m_2+3m_3)!$$
$$\times(n/2)_{m_1+2m_2+3m_3}\}^{-1}a_{000}\,. \tag{C10}$$

To obtain the sought for result, Eq. (4.35), it only remains to combine Eq. (C8) with Eq. (C10) and to take Eq. (4.22) into account.

APPENDIX D: PROOF OF EQUATIONS (5.19) and (5.20)

In the first part of this appendix, we shall show that the function defined in Eq. (5.19) is a solution of Eq. (5.18a), integrable in the domain (5.15).

Let us look for a solution of Eq. (5.18a) in the product form

$$f(t_1,t_2)=g(t_1)h(t_2)\,. \tag{D1}$$

The equation splits into the two following ordinary differential equations:

$$\left(2\frac{d^2}{dt_1^2}-\Lambda\right)g(t_1)=0\,, \tag{D2a}$$

and

$$\left[2t_2\frac{d^2}{dt_2^2}-(n-5)\frac{d}{dt_2}-\Lambda\right]h(t_2)=0\,, \tag{D2b}$$

where Λ is the separation constant.

In order that $f(t_1,t_2)$ be integrable in the domain (5.15), $g(t_1)$ must go to zero when $t_1\to\infty$. Equation (D2a) has a solution satisfying this condition,

$$g(t_1)=\exp[-(\Lambda/2)^{1/2}t_1]\,, \tag{D3}$$

for any positive value of Λ. From now on, we shall therefore restrict ourselves to such values.

By setting

$$x=\lambda\sqrt{t_2}\ \text{ and }\ h(x)=x^{\alpha}H(x)\,, \tag{D4}$$

where λ and α are two yet undetermined positive constants, Eq. (D2b) is transformed into

$$\left\{x^2\frac{d^2}{dx^2}+(2\alpha-n+4)x\frac{d}{dx}+\left[\alpha(\alpha-n+3)-\frac{2\Lambda}{\lambda^2}x^2\right]\right\}H(x)=0\,. \tag{D5}$$

For $\alpha=(n-3)/2$, and $\lambda=(2\Lambda)^{1/2}$, Eq. (D5) reduces to the differential equation

$$\left\{x^2\frac{d^2}{dx^2}+x\frac{d}{dx}-\left[x^2+\left(\frac{n-3}{2}\right)^2\right]\right\}H(x)=0\,, \tag{D6}$$

whose solutions are the modified Bessel functions $I_{\pm(n-3)/2}(x)$ and $K_{(n-3)/2}(x)$. In order that $f(t_1,t_2)$ goes to zero when t_1 and t_2 both tend to infinity, it is necessary to choose $K_{(n-3)/2}(x)$ for $H(x)$ since $I_{\pm(n-3)/2}(x)$ would lead to a divergence when $t_2=t_1^2/4\to\infty$. Hence

$$h(t_2)=(\lambda\sqrt{t_2})^{(n-3)/2}K_{(n-3)/2}(\lambda\sqrt{t_2})\,. \tag{D7}$$

We have therefore found particular solutions of Eq. (5.18a),

$$g_\lambda(t_1)h_\lambda(t_2)=\exp(-\tfrac{1}{2}\lambda t_1)(\lambda\sqrt{t_2})^{(n-3)/2}K_{(n-3)/2}(\lambda\sqrt{t_2})\,, \tag{D8}$$

having appropriate asymptotic properties and depending upon the separation constant λ, where $0<\lambda<+\infty$. By linearly combining them with arbitrary coefficients $\chi(\lambda)$, only subject to the condition that the resulting integral exists, we obtain the function defined in Eq. (5.19), which is therefore a well-behaved solution of Eq. (5.18a).

In the second part of this appendix, we shall prove Eq. (5.20). For such purpose, our starting point is the following equation:

$$\left[4t_1\frac{\partial^2}{\partial t_1^2}+8t_2\frac{\partial^2}{\partial t_1\partial t_2}-2(n-6)\frac{\partial}{\partial t_1}-1\right]$$
$$\times\exp\left(-\frac{1}{2}\lambda t_1\right)x^{(n-3)/2}K_{(n-3)/2}(x)$$
$$=\left\{-\lambda\left[2t_1\frac{\partial}{\partial t_1}+4t_2\frac{\partial}{\partial t_2}-(n-6)\right]-1\right\}$$
$$\times\exp\left(-\frac{1}{2}\lambda t_1\right)x^{(n-3)/2}K_{(n-3)/2}(x)\,, \tag{D9}$$

where x is defined in Eq. (D4). It is now straightforward to see that

$$t_1\frac{\partial}{\partial t_1}\exp\left(-\frac{1}{2}\lambda t_1\right)=-\frac{1}{2}\lambda t_1\exp\left(-\frac{1}{2}\lambda t_1\right)$$
$$=\lambda\frac{\partial}{\partial\lambda}\exp\left(-\frac{1}{2}\lambda t_1\right), \tag{D10}$$

and

$$2t_2\frac{\partial}{\partial t_2}x^{(n-3)/2}K_{(n-3)/2}(x)=x\frac{\partial}{\partial x}x^{(n-3)/2}K_{(n-3)/2}(x)$$
$$=\lambda\frac{\partial}{\partial\lambda}x^{(n-3)/2}K_{(n-3)/2}(x)\,. \tag{D11}$$

Hence

$$\left(t_1\frac{\partial}{\partial t_1}+2t_2\frac{\partial}{\partial t_2}\right)\exp\left(-\frac{1}{2}\lambda t_1\right)x^{(n-3)/2}K_{(n-3)/2}(x)$$
$$=\lambda\frac{\partial}{\partial\lambda}\exp\left(-\frac{1}{2}\lambda t_1\right)x^{(n-3)/2}K_{(n-3)/2}(x)\,. \tag{D12}$$

By introducing Eq. (D12) into the right-hand side of Eq. (D9), the latter is transformed into Eq. (5.20), whose proof is thus completed.

[1]V. Vanagas, "The Microscopic Nuclear Theory Within the Framework of the Restricted Dynamics," in *Lecture Notes in Physics* (University of Toronto, Toronto, 1977); V. Vanagas, "The Microscopic Theory of the Collective Motion in Nuclei," in *Group Theory and Its Applications in Physics-1980*, edited by T. H. Seligman, AIP Conference Proceedings No. 71 (American Institute of Physics, New York, 1981).

[2]G. F. Filippov, V. I. Ovcharenko, and Yu. F. Smirnov, *Microscopic Theory of Collective Excitations in Nuclei* (in Russian) (Naukova Dumka, Kiev, 1981).

[3]O. Castaños, A. Frank, E. Chacón, P. Hess, and M. Moshinsky, J. Math. Phys. **23**, 2537 (1982).

[4]J. Deenen and C. Quesne, J. Phys. A: Math. Gen. **16**, 2095 (1983).

[5]O. Castaños, A. Frank, E. Chacón, P. O. Hess, and M. Moshinsky, Phys. Rev. C **25**, 1611 (1982).

[6]P. Kramer, Ann. Phys. (N.Y.) **141**, 254, 269 (1982).

[7]J. Deenen and C. Quesne, J. Math. Phys. **23**, 878 (1982).

[8]J. Deenen and C. Quesne, J. Math. Phys. **23**, 2004 (1982).

[9]A. O. Barut and L. Girardello, Commun. Math. Phys. **21**, 41 (1971).

[10]T. Holstein and H. Primakoff, Phys. Rev. **58**, 1098 (1940).

[11]F. J. Dyson, Phys. Rev. **102**, 1217 (1956).

[12]A. M. Perelomov, Commun. Math. Phys. **26**, 222 (1972); Usp. Fiz. Nauk **123**, 23 (1977) [Sov. Phys. Usp. **20**, 703 (1977)].

[13]V. Bargmann, Commun. Pure Appl. Math. **14**, 187 (1961).

[14]A. C. Aitken, *Determinants and Matrices* (Oliver and Boyd, Edinburgh, 1956).

[15]I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1980).

[16]R. J. Glauber, Phys. Rev. **130**, 2529 (1963); **131**, 2766 (1963).

Mixed supertableaux of the superunitary groups. I. SU($n|1$)

F. Delduc
LPTHE,[a] *75230 Paris, France*

M. Gourdin
LPTHE,[a] *75230 Paris, France and CERN, Geneva, Switzerland*

(Received 17 June 1983; accepted for publication 30 September 1983)

Young supertableaux of the graded unitary groups SU($n|1$) are classified and interpreted in terms of representations of the group. A distinction between typical and atypical representations arises naturally, and we propose a new type of generalized atypical supertableaux for non-fully-reducible representations.

PACS numbers: 02.20. + b

1. INTRODUCTION

Young supertableaux have been introduced by Balantekin and Bars[1,2] for a graphical description of tensorial representations of Lie superalgebras as a generalization of ordinary Young tableaux for ordinary Lie algebras. The simplest case appears to be that of superunitary groups SU($n|m$) defined as the set of complex linear transformations of the special graded linear group SPL($n|m$) leaving invariant an even regular sesquilinear form G,[3]

$$\mathrm{SU}(n|m) = \{M \in \mathrm{SPL}(n|m) | M^+GM = G\},$$
$$\mathrm{su}(n|m) = \{X \in \mathrm{spl}(n|m) | X^+G + GX = 0\}.$$

The relation between supertableaux and representations of basic simple supergroups has been initiated by Bars[4] and Morel and Ruegg.[5] For special superunitary groups SU($n|m$) the legal purely covariant and legal purely contravariant supertableaux describe one irreducible representation of SU($n|m$), typical or atypical, but the correspondence is not one-to-one and the concept of equivalence between supertableaux of these types has to be introduced. Unfortunately, there exist irreducible representations of SU($n|m$) and also not fully reducible representations which cannot be visualized in that way, and mixed supertableaux containing both covariant and contravariant boxes must be introduced. Their discussion is considerably more complicated than for covariant and contravariant supertableaux, and the aim of this paper is a classification and an interpretation of the mixed supertableaux of SU($n|1$). The general case for SU($n|m$) will be considered in a subsequent publication.

The fundamental representation of SU($n|m$) acts in a graded ($n+m$)-dimensional space $V = V_0 \oplus V_1$ where V_0 is bosonic and V_1 fermionic. We have two possible graduations of V and, by tensor product, two classes of representations of SU($n|m$)

$$\dim V_0 = n, \quad \dim V_1 = m, \quad \text{Class I},$$
$$\dim V_0 = m, \quad \dim V_1 = n, \quad \text{Class II}.$$

Because of the isomorphism of the two supergroups SU($n|m$) and SU($m|n$) it is obvious that class I (II) representation of SU($n|m$) are also class II (I) representations of SU($m|n$). Equivalently a class II supertableau is simply obtained from a class I supertableau by the transposition operation exchanging the roles of rows and columns. In the rest of this paper we shall limit ourselves to class I representations of SU($n|1$). For a measure of the number of covariant and contravariant boxes of the rows and columns of a supertableau of SU($n|m$) we use the notations of Ref. 6:

$$\left|\begin{matrix} b_j \\ \bar{b}_j \end{matrix}\right| \text{ counts the } \left|\begin{matrix} \text{covariant} \\ \text{contravariant} \end{matrix}\right| \text{ boxes of the row } j,$$

$$\left|\begin{matrix} c_k \\ \bar{c}_k \end{matrix}\right| \text{ counts the } \left|\begin{matrix} \text{covariant} \\ \text{contravariant} \end{matrix}\right| \text{ boxes of the column } k,$$

with the usual constraints

$$b_1 \geqslant b_2 \geqslant \cdots \geqslant b_j \geqslant \cdots \geqslant 0, \qquad \bar{b}_1 \geqslant \bar{b}_2 \geqslant \cdots \geqslant \bar{b}_j \geqslant \cdots \geqslant 0,$$
$$c_1 \geqslant c_2 \geqslant \cdots \geqslant c_k \geqslant \cdots \geqslant 0, \qquad \bar{c}_1 \geqslant \bar{c}_2 \geqslant \cdots \geqslant \bar{c}_k \geqslant \cdots \geqslant 0.$$

A class I supertableau of SU($n|m$) is legal if and only if the numbers of covariant and contravariant boxes of its rows satisfy *one* of the inequalities

$$b_{n+1-j} + \bar{b}_{j+1} \leqslant m, \quad j = 0,1,\ldots,n, \tag{1}$$

or, equivalently if the numbers of covariant and contravariant boxes of its columns satisfy *one* of the inequalities

$$c_{k+1} + \bar{c}_{m+1-k} \leqslant n, \quad k = 0,1,\ldots,m. \tag{2}$$

These two sets of constraints are clearly equivalent, and for a legal class I supertableau of SU($n|m$) at least one of the inequalities (1) and one of the inequalities (2) is fulfilled.

For purely contravariant and purely covariant supertableaux the conditions given in Ref. 6

(i) $b_{n+1} \leqslant m$; $c_{m+1} \leqslant n$ in the purely covariant case,

(ii) $\bar{b}_{n+1} \leqslant m$; $\bar{c}_{m+1} \leqslant n$ in the purely contravariant case

are obviously contained in Eqs. (1) or (2).

The proof of the inequalities (1) can be obtained by using two complementary techniques:

(i) the reduction of a legal supertableau with respect to the subgroup SU(n) ⊗ SU(m) ⊗ U(1) associated with the bosonic part of the superalgebra must exist; such a criterion has been partly discussed in Ref. 6;

(ii) the computation of the dimension of the supertableau with a determinant, as proposed in Ref. 1 must give, for a legal typical supertableau, a positive result.

---
[a] LPTHE, Tour 16 El, Université Pierre et Marie Curie, 4 Place Jussieu, 75230 Paris, Cedex 05.

We shall not give more details in this paper and we postpone the general proof to a subsequent publication.

2. KAC–DYNKIN PARAMETERS OF SU($n|1$)

(1) The graded Lie algebra of the superunitary group SU($n|1$) has a Bose sector which is the Lie algebra of the direct product SU(n) $\otimes$ U(1). We have n Cartan operators $H_1,H_2,...,H_n$ the $(n-1)$ first ones belonging to SU(n) and the U(1) generator Q is defined by

$$Q=\frac{1}{n}\sum_1^n jH_j.$$

The Kac–Dynkin parameters $\{a_1,a_2,...,a_{n-1};a_n\}$ define an irreducible representation of SU($n|1$).[6–8] They are the eigenvalues of the Cartan generators for the highest weight of the representation chosen by Kac as corresponding to the lowest eigenvalue of Q.

The $(n-1)$ first Kac–Dynkin parameters $(a_1,a_2,...,a_{n-1})$ describe an irreducible representation of SU(n), precisely that associated with the lowest eigenvalue of Q. As a consequence, they are nonnegative integers. The last parameter a_n may be any real number. In the supertableau approach it is an algebraic integer. In what follows we restrict ourselves for a_n to algebraic integer values, advocating a remark on this point made in Ref. 6, even if the question does not seem to us entirely clear.

(2) We have two types of irreducible representations of SU($n|1$), the typical representations and the atypical representations.[6–9] Fixing $a_1,a_2,...,a_{n-1}$ and defining

$$A_j=\sum_{n-j}^{n-1}(1+a_s),\quad j=1,2,...,n-1, \tag{3}$$

we have the following result. An irreducible representation of SU($n|1$) is atypical if and only if a_n takes one of the n values

$$a_n=0,-A_1,-A_2,...,-A_{n-1}.$$

In particular all the representations with $a_n>0$ and $a_n<-A_{n-1}$ are typical. For typical representations of SU($n|1$) the dimension is even—with an equal number of bosonic and fermionic weights—and it turns out to be independent of a_n and given by [6–9]

$$\dim\{a_1,a_2,...,a_{n-1};a_n\}=2^n\dim(a_1,a_2,...,a_{n-1}),$$

where the last factor is the dimension of the SU(n) representation given by the well-known formula

$$\dim(a_1,a_2,...,a_{n-1})=\prod_{1\leqslant i\leqslant j\leqslant n-1}\left[1+\frac{\Sigma_i^j a_s}{j-i+1}\right].$$

For atypical representations the dimension is always lower.

(3) Finally, let us give as an illustration, three important examples:

(i) the fundamental representation $\{1,0,...,0;0\}$ has the dimension $n+1$; it is atypical and associated with the supertableau with one covariant box.

(ii) the fundamental representation $\{0,...,0;-1\}$ is the contragradient of the previous one with the same dimension $n+1$; it is atypical and associated with the supertableau with one contravariant box

(iii) the adjoint representation $\{1,0,....,0;-1\}$ has the dimension $(n+1)^2-1$ and is self-contragradient; it is atypical when $n\geqslant 3$ and associated with the supertableau with one covariant box and one contravariant box. For SU($2|1$) the adjoint representation is typical, and it has the dimension 8.

3. LEGAL SUPERTABLEAUX OF SU($n/1$)

(1) In SU($n|1$) the two sets of constraints (1) and (2) are written as

$$b_{n+1-j}+\bar{b}_{j+1}\leqslant 1,\quad j=0,1,2,...,n, \tag{4}$$

$$c_{k+1}+\bar{c}_{2-k}\leqslant n,\quad k=0,1. \tag{5}$$

We then can distinguish between two categories of supertableaux, depending on the value of the sum of covariant and contravariant boxes of the first columns $c_1+\bar{c}_1$:

(α) When $c_1+\bar{c}_1<n$, the two inequalities (5) are simultaneously satisfied and we can describe these supertableaux with c_1 parameters of type b_j and $\bar{c}_1$ parameters of type $\bar{b}_j$ fixing entirely the number of boxes of each row.

(β) When $c_1+\bar{c}_1\geqslant n$, we solve Eqs. (4) and (5), and we easily find $2(n+1)$ solutions characterizing $2(n+1)$ types of supertableaux. In this case only $n+1$ free parameters determine the supertableau, the other ones being fixed.

(2) The construction and the discussion of the various types of legal supertableaux will be made by the following method. With a legal supertableau of SU($n|1$) we associate a point P in a $(n+1)$-dimensional space. The coordinates of P are chosen in such a way that they produce the same Kac–Dynkin parameters for the supertableau as those given by the SU(n) $\otimes$ U(1) analysis of the supertableau.[6] We then introduce $n+1$ coordinates

$$x_1\geqslant x_2\geqslant\cdots\geqslant x_n \quad\text{and}\quad y_1.$$

The Kac–Dynkin parameters are then computed with the formulae

$$a_j=x_j-x_{j+1}\geqslant 0,\quad j=1,2,...,n-1,$$

$$a_n=x_n-y_1.$$

Inverting these relations, we are able to express the parameters b_l and $\bar{b}_l$ as linear combinations of the Kac–Dynkin parameters a_j and of y_1. This last quantity has a range of variation defined by the positivity constraints on the parameters $b_l,\bar{b}_l,c_1,\bar{c}_1$, and it is interpreted as describing the *equivalence* between supertableaux of different topology but having the same Kac–Dynkin parameters.

4. ATYPICAL SUPERTABLEAUX $c_1+\bar{c}_1<n$

(1) Let us call $N=c_1$ the number of rows with covariant boxes and $\bar{N}=\bar{c}_1$ the number of rows with contravariant boxes, and let us study the case where $N+\bar{N}<n$. The corresponding supertableaux are defined by $N+\bar{N}$ positive integers $b_j,\bar{b}_j$, with the constraints

$$b_1\geqslant b_2\geqslant\cdots\geqslant b_N\geqslant 1,$$

$$\bar{b}_1\geqslant\bar{b}_2\geqslant\cdots\geqslant\bar{b}_{\bar{N}}\geqslant 1,$$

and they are represented in Fig. 1.

The Kac–Dynkin parameters of this supertableau are defined by the eigenvalues of the Cartan subalgebra generators for the highest weight of the representation. Such an analysis is made in Appendix A, and it is equivalent to the

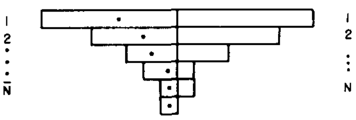


FIG. 1. Atypical supertableau $c_1 + \bar{c}_1 < n$.

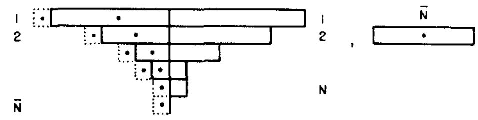


FIG. 2. $\mathrm{SU}(n) \otimes \mathrm{U}(1)$ component of Fig. 1 containing the highest weight.

following choice of coordinates for the point P of the $(n+1)$-dimensional space, associated with the supertableau

$$\begin{aligned} x_s &= b_s, & s &= 1,2,\ldots,N, \\ x_s &= 0, & s &= N+1,\ldots,n-\bar{N}, \\ x_s &= 1 - \bar{b}_{n+1-s}, & s &= n-\bar{N}+1,\ldots,n, \end{aligned} \tag{6}$$

and

$$y_1 = \bar{c}_1 = \bar{N}.$$

We then compute the Kac–Dynkin parameters with the usual formulae, and we get

$$\begin{aligned} a_s &= b_s - b_{s+1}, & s &= 1,2,\ldots,N-1, \\ a_n &= b_n, \\ a_s &= 0, & s &= N+1,\ldots,n-\bar{N}-1, \\ a_{n-\bar{N}} &= \bar{b}_{\bar{N}} - 1, \\ a_s &= \bar{b}_{n-s} - \bar{b}_{n-s+1}, & s &= n-\bar{N}+1,\ldots,n-1, \end{aligned} \tag{7}$$

and

$$a_n = 1 - \bar{b}_1 - \bar{N}.$$

The formulae (7) can be uniquely inverted, and we immediately obtain the length of the covariant and contravariant rows in terms of the $(n-1)$ Kac–Dynkin parameters of the $\mathrm{SU}(n)$ part

$$\begin{aligned} b_j &= \sum_j^N a_s, & j &= 1,2,\ldots,N, \\ \bar{b}_j &= 1 + \sum_j^{\bar{N}} a_{n-s}, & j &= 1,2,\ldots,\bar{N} \end{aligned} \tag{8}$$

The last Kac–Dynkin parameter a_n turns out to be a linear combination of the same a_j's given by

$$a_n = -\bar{N} - \sum_1^{\bar{N}} a_{n-s} = -A_{\bar{N}}.$$

It follows that a_n takes the atypical value $-A_{\bar{N}}$ previously defined in Eq. (2).

The supertableaux with $c_1 + \bar{c}_1 < n$ describe the atypical irreducible representations of $\mathrm{SU}(n|1)$ and the correspondence between the supertableaux and the representation is one to one without equivalence. As an example, the N row covariant supertableaux with $N < n$ correspond to the atypical representations $a_n = 0$.

(2) The $\mathrm{SU}(n) \otimes \mathrm{U}(1)$ decomposition of the atypical irreducible representations is easily made by using the supertableaux. The component of the lowest eigenvalue of Q represented in Fig. 2 contains the highest weight, and by definition of the Kac–Dynkin parameters we simply have

$$q_{\min} = \frac{1}{n} \sum_1^n j a_j. \tag{9}$$

The other levels in q are obtained by transferring covariant and contravariant boxes. For one box the spacing is $(1 - 1/n)$ and the total number of possible levels is $N + \bar{N} + 1$. It follows that the maximal value of q is given by

$$q_{\max} = q_{\min} + (N + \bar{N})(1 - 1/n).$$

When all possible $\mathrm{SU}(n) \otimes \mathrm{U}(1)$ components are present, simple combinatorics gives the total number of these components to be $2^{N+\bar{N}}$.

5. TYPICAL SUPERTABLEAUX $c_1 + \bar{c}_1 \geqslant n$

(1) We now consider the case where $c_1 + \bar{c}_1 \geqslant n$, and we introduce the constraints (4) and (5). We classify the supertableaux $c_1 + \bar{c}_1 \geqslant n$ with one constraint (4) and one constraint (5)

$$I_k^j \Rightarrow \begin{cases} b_{n+1-j} + \bar{b}_{j+1} \leqslant 1 & (10) \\ c_{k+1} + \bar{c}_{2-k} \leqslant n. & (11) \end{cases}$$

The parameter j can take $n+1$ values $j = 0,1,\ldots,n$ and the parameter k takes two values only for $m = 1$, $k = 0,1$. The total number of types of supertableaux $c_1 + \bar{c}_1 \geqslant n$ is then $2(n+1)$.

The I_k^j type supertableaux are first defined by a set of n positive integers describing the nontrivial rows

$$\begin{aligned} & b_1 \geqslant b_2 \geqslant \cdots \geqslant b_{n-j} \geqslant 1, \\ & \bar{b}_1 \geqslant \bar{b}_2 \geqslant \cdots \geqslant \bar{b}_j \geqslant 1. \end{aligned} \tag{12}$$

The inequality (10) means that one of the two quantities b_{n-j+1} or $\bar{b}_{j+1}$ must vanish, the other one being bounded by one. The choice between these two possibilities is made by the inequality (11), and we get

$$\bar{b}_{j+1} = 0, \quad \bar{c}_1 = j, \quad c_1 \geqslant n - j \quad \text{for } I_1^j,$$

$$b_{n-j+1} = 0, \quad c_1 = n - j, \quad \bar{c}_1 \geqslant j \quad \text{for } I_0^j.$$

(2) Let us first discuss the n parameters b and $\bar{b}$. The x coordinates of the associated point P of the $(n+1)$-dimensional space S are given, as previously, in Eq. (6)

$$\begin{aligned} x_s &= b_s, \quad s = 1,2,\ldots,n-j, \\ x_s &= 1 - \bar{b}_{n+1-s}, \quad s = n-j+1,\ldots,n, \end{aligned} \tag{13}$$

and we compute the $(n-1)$ Kac–Dynkin parameters of the $\mathrm{SU}(n)$ part

$$a_s = b_s - b_{s+1}, \quad s = 1,2,\ldots,n-j-1, \tag{14}$$

$$a_{n-j} = b_{n-j} + \bar{b}_j - 1, \tag{15}$$

$$a_s = \bar{b}_{n-s} - \bar{b}_{n-s+1}, \quad s = n-j+1,\ldots,n-1. \tag{16}$$

(3) Let us first study the case $j = 0$. Equations (14) only are present, and they are inverted as follows:

$$b_l = \sum_l^{n-1} a_s + b_n, \quad l = 1,2,\ldots,n-1. \tag{17}$$

The positivity constraints on the b_l parameters imply a trivial ordering of the x_l coordinates

$$x_1 \geqslant x_2 \geqslant \cdots \geqslant x_n \geqslant 1.$$

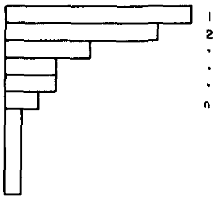


FIG. 3. Typical covariant supertableau $c_1 + \bar{c}_1 \geqslant n$.

the set of supertableaux of type I^0 is divided in two subsets

$I^0_1 \quad \bar{c}_1 = 0, \quad c_1 \geqslant n,$ purely covariant supertableaux,

$I^0_0 \quad c_1 = n, \quad \bar{c}_1 \geqslant 1,$ mixed supertableaux.

The purely covariant supertableaux I^0_1 are represented in Fig. 3. The y_1 coordinate of the associated point P is chosen so that a_n agrees with the value obtained using the $SU(n) \otimes U(1)$ reduction (see Appendix B). With

$$y_1 = n - c_1$$

we compute the last Kac–Dynkin parameter a_n

$$a_n = x_n - y_1 = b_n + c_1 - n. \tag{18}$$

The positivity constraints $b_n \geqslant 1$ and $c_1 \geqslant n$ give a double inequality for y_1

$$1 - a_n \leqslant y_1 \leqslant 0. \tag{19}$$

The compatibility of these inequalities restricts a_n to positive values only

$$a_n \geqslant 1.$$

As a consequence, purely covariant supertableaux of type I^0_1 are associated to typical irreducible representations of $SU(n|1)$ with $a_n > 0$. Giving now a typical irreducible representation of $SU(n|1)$ by its Kac–Dynkin parameters with $a_n > 0$, the purely covariant supertableaux associated with this representation are constructed as follows:

$$b_l = \sum_{l}^{n-1} a_s + a_n + y_1, \qquad c_1 = n - y_1, \tag{20}$$

with the parameter y_1 varying in the range (19). We then obtain a_n equivalent typical covariant supertableaux I^0_1.

The mixed supertableaux of type I^0_0 are represented in Fig. 5. The y_1 coordinate is chosen as previously (see Appendix B) and with

$$y_1 = \bar{c}_1$$

we determine the last Kac–Dynkin parameter a_n

$$a_n = x_n - y_1 = b_n - \bar{c}_1. \tag{21}$$

The positivity constraint $\bar{c}_1 \geqslant 1$ implies trivially

$$y_1 \geqslant 1,$$

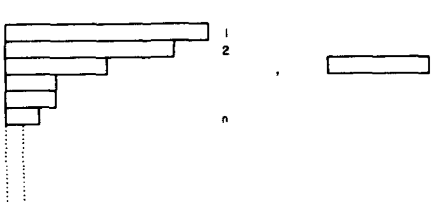


FIG. 4. $SU(n) \otimes U(1)$ component of Fig. 3 containing the highest weight.

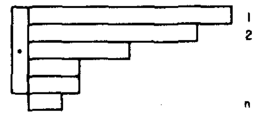


FIG. 5. Mixed supertableau of type I^0_1.

and we have no restriction on a_n in this case. When a_n has a typical value, we obtain an infinite set of mixed supertableaux of type I^0_0 associated with a typical irreducible representation of $SU(n|1)$. These supertableaux are parametrized with the coordinate $y_1 \geqslant 1$

$$b_l = \sum_{l}^{n-1} a_s + a_n + y_1, \qquad \bar{c}_1 = y_1. \tag{22}$$

The case where a_n has an atypical value will be discussed in Sec. 7.

(4) We are now interested in the case $j = n$ which is the contragradient[10] of the previous one. Only Eqs. (16) are present, and they are inverted as follows:

$$\bar{b}_l = -\sum_{n+1-l}^{n-1} a_s + \bar{b}_1, \quad l = 2,3,\ldots,n. \tag{23}$$

The positivity constraints on the $\bar{b}_l$ parameters imply a trivial ordering of the x_l coordinates

$$0 \geqslant x_1 \geqslant x_2 \geqslant \cdots \geqslant x_n.$$

The set of supertableaux of type I^n is divided into two subsets:

$I^n_1, \quad \bar{c}_1 = n, \quad c_1 \geqslant 1,$ mixed supertableaux,

$I^n_0, \quad \bar{c}_1 \geqslant n, \quad c_1 = 0,$ purely contravariant supertableaux.

The purely contravariant supertableaux I^n_0 are represented in Fig. 7. The y_1 coordinate is chosen on the $SU(n) \otimes U(1)$ basis as previously (see Appendix B), and we have

$$y_1 = \bar{c}_1.$$

The last Kac–Dynkin parameter a_n has the value

$$a_n = x_n - y_1 = 1 - \bar{b}_1 - \bar{c}_1. \tag{24}$$

The positivity constraints $\bar{b}_n \geqslant 1$, $\bar{c}_1 \geqslant n$ give a double inequality for y_1

$$n \leqslant y_1 \leqslant -a_n - \sum_{1}^{n-1} a_s. \tag{25}$$

The compatibility of these inequalities restricts a_n to typical values

$$a_n \leqslant -n - \sum_{1}^{n-1} a_s = -1 - A_{n-1}. \tag{26}$$

As a consequence, purely contravariant supertableaux I^n_0 are associated to typical irreducible representations of $SU(n|1)$ with $a_n < -A_{n-1}$.

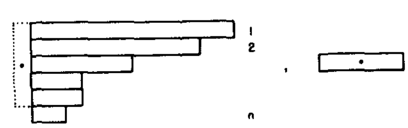


FIG. 6. $SU(n) \otimes U(1)$ component of Fig. 5 containing the highest weight.

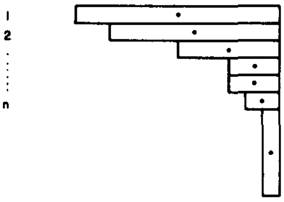


FIG. 7. Typical contravariant supertableau $c_1 + \bar{c}_1 > n$.

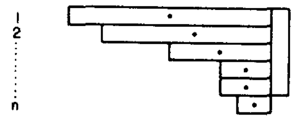


FIG. 9. Mixed supertableau of type I_0^n.

Giving now a typical irreducible representation of SU($n|1$) by its Kac–Dynkin parameters with $a_n < -A_{n-1}$, the purely contravariant supertableaux associated with this representation are constructed as follows:

$$\begin{aligned} \bar{b}_l &= -\sum_{n+1-l}^{n-1} a_s + 1 - a_n - y_1, \\ \bar{c}_1 &= y_1, \end{aligned} \tag{27}$$

with the parameter y_1 varying in the range (25). We then obtain $-a_n - A_{n-1}$ equivalent typical contravariant supertableaux I_0^n.

The mixed supertableaux of type I_{1}^{n} *—* are represented in Fig. 9. The y_1 coordinate is chosen as previously (see Appendix B), and with

$$y_1 = n - c_1$$

we determine the Kac–Dynkin parameter a_n

$$a_n = x_n - y_1 = 1 - n - \bar{b}_1 + c_1. \tag{28}$$

The positivity constraint $c_1 \geqslant 1$ implies trivially

$$y_1 \leqslant n - 1,$$

and we have no restriction on a_n. When a_n has a typical value, we obtain an infinite set of mixed supertableaux of type I_1^n associated to a typical irreducible representation of SU($n|1$). These supertableaux are parametrized with the coordinate y_1

$$\begin{aligned} \bar{b}_l &= -\sum_{n+1-l}^{n-1} a_s + 1 - a_n - y_1, \\ c_1 &= n - y_1 \end{aligned} \tag{29}$$

with now $y_1 \leqslant n - 1$.

(5) For the intermediate situations $0 < j < n$ the relation (15) is present and when combined with the positivity constraints $b_{n-j} \geqslant 1$ and $\bar{b}_j \geqslant 1$, it gives

$$a_{n-j} \geqslant 1.$$

This inequality is a necessary and sufficient condition for the existence of mixed supertableaux of type I_k^j for $0 < j < n$.

Assuming now this condition to be fulfilled, we invert the relations (14), (15), and (16) using $\bar{b}_1$ as a parameter,

$$\begin{aligned} b_l &= \sum_{l}^{n-1} a_s + 1 - \bar{b}_1, \quad l = 1,2,\ldots,n-j, \\ \bar{b}_l &= -\sum_{n+1-l}^{n-1} a_s + \bar{b}_1, \quad l = 1,2,\ldots,j\ . \end{aligned} \tag{30}$$

The positivity constraints (12) imply bounds for the x coordinates of the point P. In particular for $x_n = 1 - \bar{b}_1$ we get

$$1 - \sum_{n-j}^{n-1} a_s \leqslant x_n \leqslant -\sum_{n-j+1}^{n-1} a_s, \tag{31}$$

and for the ordering of the x coordinates

$$\sum_{1}^{n-1} a_s \geqslant x_1 \geqslant x_2 \geqslant \cdots \geqslant x_{n-j} > 0 \geqslant x_{n-j+1} \geqslant \cdots \geqslant x_n \geqslant 1 - \sum_{n-j}^{n-1} a_s.$$

The set of mixed supertableaux of type I^j is divided in two subsets

I_1^j, $\bar{c}_1 = j$, $c_1 \geqslant n - j$, mixed supertableaux,

I_0^j, $c_1 = n - j$, $\bar{c}_1 \geqslant j$, mixed supertableaux.

The mixed supertableaux of type I_0^j are represented in Fig. 11. The y_1 coordinate of the point P is given by (see Appendix B)

$$y_1 = \bar{c}_1,$$

and the Kac–Dynkin parameter a_n has the value

$$a_n = x_n - y_1 = 1 - \bar{b}_1 - \bar{c}_1. \tag{32}$$

The positivity constraints $\bar{b}_j \geqslant 1$, $\bar{c}_1 \geqslant j$ give a double inequality for y_1,

$$j \leqslant y_1 \leqslant -\sum_{n+1-j}^{n-1} a_s - a_n. \tag{33}$$

The compatibility of these inequalities gives an upper bound for the values of a_n associated with supertableaux of type I_0^j

$$a_n \leqslant -j - \sum_{n-j+1}^{n-1} a_s = -1 - A_{j-1}. \tag{34}$$

On the other hand, the third positivity constraint $b_{n-j} \geqslant 1$ implies a second lower limit for y_1

$$1 - \sum_{n-j}^{n-1} a_s - a_n \leqslant y_1. \tag{35}$$

Giving now an irreducible typical representation of SU($n|1$) by its Kac–Dynkin parameters with $a_{n-j} \geqslant 1$ and a_n typical satisfying the inequality (34), we associate with this represen-

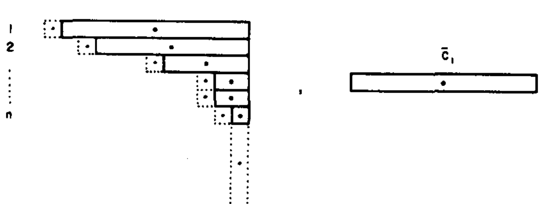


FIG. 8. SU(n) ⊗ U(1) component of Fig. 7 containing the highest weight.

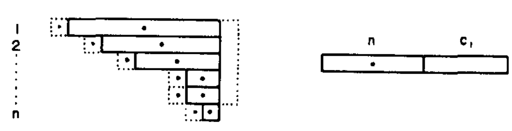


FIG. 10. SU(n) ⊗ U(1) component of Fig. 9 containing the highest weight.

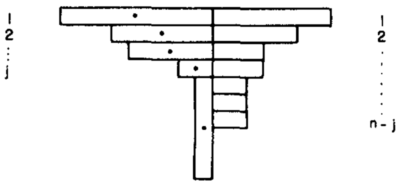


FIG. 11. Mixed supertableau of type I_0^j.

tation mixed supertableaux of type I_0^j described by

$$
\begin{aligned}
b_l &= \sum_{l}^{n-1} a_s + a_n + y_1, && l = 1,2,\ldots,n-j,\\
\bar{b}_l &= -\sum_{n-l+1}^{n-1} a_s + 1 - a_n - y_1, && l = 1,2,\ldots,j, \qquad (36)\\
\bar{c}_1 &= y_1,
\end{aligned}
$$

with y_1 varying in the range intersection of the ranges (33) and (35). The number of equivalent supertableaux of type I_0^j describing the same typical irreducible representation of $SU(n|1)$ depends on a_n. The result is

(i) $1 - A_j \leqslant a_n \leqslant -1 - A_{j-1}$,
$1 - A_{j-1} - a_n$ supertableaux,

(ii) $a_n < -A_j$,
a_{n-j} supertableaux.

The mixed supertableaux of type I_1^j are represented in Fig. 13. The y_1 coordinate of the point P is given by (see Appendix B)

$$y_1 = n - c_1,$$

and the Kac–Dynkin parameter a_n has the value

$$a_n = x_n - y_1 = 1 - n - \bar{b}_1 + c_1. \qquad (37)$$

The positivity constraints $b_{n-j} \geqslant 1$, $c_1 \geqslant n - j$ give a double inequality for y_1

$$1 - \sum_{n-j}^{n-1} a_s - a_n \leqslant y_1 \leqslant j. \qquad (38)$$

The compatibility of these inequalities implies a lower bound for the values of a_n associated with supertableaux of type I_1^j

$$a_n \geqslant 1 - j - \sum_{n-j}^{n-1} a_s = 1 - A_j. \qquad (39)$$

On the other hand, the third positivity constraint $\bar{b}_j \geqslant 1$ implies a second upper limit for y_1

$$y_1 \leqslant -\sum_{n-j+1}^{n-1} a_s - a_n. \qquad (40)$$

Giving now an irreducible representation of $SU(n|1)$ by its Kac–Dynkin parameters with $a_{n-j} \geqslant 1$ and a_n typical satis-

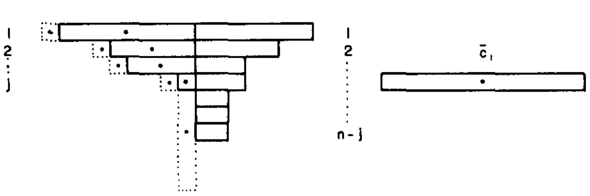


FIG. 12. $SU(n) \otimes U(1)$ component of Fig. 11 containing the highest weight.

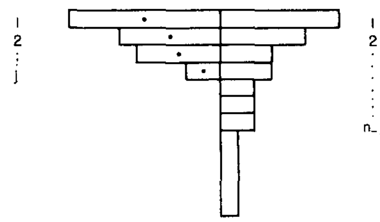


FIG. 13. Mixed supertableau of type I_1^j.

fying the inequality (39), we associate with this representation mixed supertableaux of type I_1^j described by

$$
\begin{aligned}
b_l &= \sum_{l}^{n-1} a_s + a_n + y_1, && l = 1,2,\ldots,n-j,\\
\bar{b}_l &= -\sum_{n-l+1}^{n-1} a_s + 1 - a_n - y_1, && l = 1,2,\ldots,j, \qquad (41)\\
c_1 &= n - y_1,
\end{aligned}
$$

with y_1 varying in the range intersection of the ranges (38) and (40). The number of equivalent supertableaux describing the same typical irreducible representation depends on a_n. The result is

(i) $1 - A_j \leqslant a_n \leqslant -1 - A_{j-1}$,
$-1 + A_j + a_n$ supertableaux,

(ii) $a_n > -A_{j-1}$,
a_{n-j} supertableaux.

Finally, the intersection of the two sets I_0^j and I_1^j corresponds to mixed supertableaux with only $(n-j)$ covariant rows and j contravariant rows. They have the shape drawn in Fig. 1 with now $N = n - j$ and $\bar{N} = j$. The length of the covariant and contravariant rows is given by Eqs. (36) and (41) with $y_1 = j$

$$
\begin{aligned}
b_l &= \sum_{l}^{n-1} a_s + a_n + j, && l = 1,2,\ldots,n-j,\\
\bar{b}_l &= -\sum_{n-l+1}^{n-1} a_s + 1 - a_n - j, && l = 1,2,\ldots,j,
\end{aligned}
$$

and we have $c_1 = n - j$ and $\bar{c}_1 = j$ fixed. In this case where $c_1 + \bar{c}_1 = n$ we have no equivalence.

The subset of mixed supertableaux $I_1^j \cap I_0^j$ describes the typical irreducible representations of $SU(n|1)$ with a_n in the range intersection of the ranges (34) and (39)

$$1 - A_j \leqslant a_n \leqslant -1 - A_{j-1}.$$

This finite subset has the dimension a_{n-j}.

For $j = 0$ and $j = n$ the analogs of this subset of typical

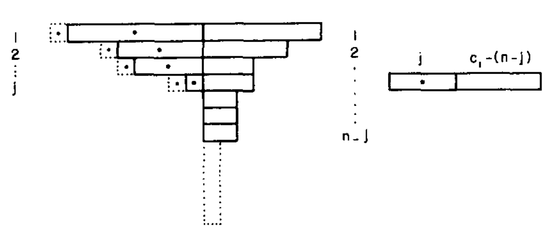


FIG. 14. $SU(n) \otimes U(1)$ component of Fig. 13 containing the highest weight.

supertableaux are, respectively, the n covariant row supertableaux of I_1^0 and the n contravariant row supertableaux of I_0^n.

In conclusion the full set of supertableaux $c_1 + \bar{c}_1 = n$ gives a unique description of all typical representations of SU$(n|1)$.

(6) For the typical supertableaux $c_1 + \bar{c}_1 \geqslant n$ the relations of contragradience are simply expressed. It is straightforward to check that, for all supertableaux I_k^j, the exchanges $b_j \Leftrightarrow \bar{b}_j$, $c_1 \Leftrightarrow \bar{c}_1$ imply the following relations between the coordinates of two contragradient points P and $P^{\rm CT}$ in the $(n+1)$-dimensional space S

$$x_j^{\rm CT} = 1 - x_{n+1-j}, \quad j = 1,2,\ldots,n,$$

$$y_1^{\rm CT} = n - y_1.$$

As a consequence, we obtain the relations between the Kac–Dynkin parameters of two typical irreducible representations of SU$(n|1)$

$$a_j^{\rm CT} = a_{n-j}, \quad j = 1,2,\ldots,n-1,$$

$$a_n^{\rm CT} = 1 - n - \sum_1^{n-1} a_s - a_n.$$

(7) The SU$(N)\otimes$U(1) decomposition of the typical irreducible representations considered in this section is easily made using the supertableaux. The components with the lowest eigenvalue of the U(1) operator Q are represented for each type of typical supertableaux in Figs. 4, 6, 8, 10, 12, and 14. We simply have

$$q_{\min} = \frac{1}{n}\sum_1^n j a_j, \tag{42}$$

and the other levels in q are obtained by transferring the covariant boxes from the SU(n) part to the U(1) part, and the contravariant boxes from the U(1) part to the SU(n) part, in all the ways compatible with the symmetries of the boxes. The spacing for one box is $(1 - 1/n)$ and the number of possible levels is $n+1$. The maximal value of q is then given

$$q_{\max} = q_{\min} + n(1 - 1/n) = q_{\min} + n - 1.$$

The level $q_j = q_{\min} + j(1 - 1/n)$ contains, at most, C_n^j components SU(n) and the total number of these components is simply

$$\sum_{j=0}^{j=n} C_n^j = 2^n.$$

When some of the Kac–Dynkin parameters $a_1,a_2,\ldots,a_{n-1}$ of a typical representation vanish the number of SU$(n)\otimes$U(1) representations is, by accident, less than 2^n but the number of levels in q stays at $n+1$.

6. TENSOR PRODUCTS INVOLVING THE FUNDAMENTAL REPRESENTATIONS

(1) A straightforward way to construct representations of SU$(n|1)$ is to perform the tensor product of a particular representation by one of the two fundamental representations. The reduction of the tensor product of these two representations is simply obtained by considering the tensor product within SU(n) of their SU$(n)\otimes$U(1) components and by rearranging the result into representations of SU$(n|1)$.

We now consider the tensor product of an irreducible representation of SU(n) given by its $n-1$ Dynkin parameters $(a_1,a_2,\ldots,a_{n-1})$ by one of the two fundamental representations of SU(n), $(1,0,\ldots,0)$, or $(0\ldots,0,1)$. The result is

$$\begin{aligned}(a_1,a_2,\ldots,a_{n-1}) &\otimes (1,0,\ldots,0)\\ &= (a_1+1,a_2,\ldots,a_{n-1}) \oplus (a_1,\ldots,a_{n-2}, a_{n-1}-1)\\ &\quad \bigoplus_{i=1}^{i=n-2} (a_1,\ldots,a_i - 1, a_{i+1}+1,\ldots,a_{n-1}),\end{aligned} \tag{43}$$

$$\begin{aligned}(a_1,a_2,\ldots,a_{n-1}) &\otimes (0,\ldots,0,1)\\ &= (a_1-1,a_2,\ldots,a_{n-1}) \oplus (a_1,\ldots,a_{n-2}, a_{n-1}+1)\\ &\quad \bigoplus_{i=1}^{i=n-2} (a_1,\ldots,a_i + 1, a_{i+1}-1,\ldots,a_{n-1}).\end{aligned} \tag{44}$$

For the first two terms of the sums (43) and (44) only one Dynkin parameter a_1 or a_{n-1} is modified; for the $n-2$ other terms two Dynkin parameters are modified as indicated. When all the a_j's are nonzero, the sums (43) and (44) have n terms. When some a_j's are vanishing, the corresponding terms involving $a_j - 1$ are not present in the sums.

(2) As an immediate consequence, we get the following formulae for the tensor product of an irreducible representation of SU$(n|1)$ given by its n Kac–Dynkin parameters $\{a_1,a_2,\ldots,a_{n-1};a_n\}$ by one of the two fundamental representations $\{1,0,\ldots,0;0\}$ or $\{0,\ldots,0;-1\}$:

$$\begin{aligned}\{a_1,a_2,\ldots,a_{n-1};a_n\} &\otimes \{1,0,\ldots,0;0\}\\ &= \{a_1+1,a_2,\ldots;a_n\} \oplus \{a_1,\ldots,a_{n-1};a_n+1\}\\ &\quad \bigoplus_{i=1}^{i=n-1} \{a_1,\ldots,a_i - 1, a_{i+1}+1,\ldots;a_n\},\end{aligned} \tag{45}$$

$$\begin{aligned}\{a_1,a_2,\ldots,a_{n-1};a_n\} &\otimes \{0,0,\ldots,0;-1\}\\ &= \{a_1-1,a_2,\ldots,a_n\} \oplus \{a_1,\ldots,a_{n-1};a_n-1\}\\ &\quad \bigoplus_{i=1}^{i=n-1} \{a_1,\ldots,a_i + 1, a_{i+1}-1,\ldots;a_n\}.\end{aligned} \tag{46}$$

When all the a_j's, $j = 1,2,\ldots,n-1$, are nonzero, the sums (45) and (46) have $n+1$ terms. When some of these a_j's are vanishing, the corresponding terms involving $a_j - 1$ disappear from the sum. Of course, this last remark does not apply to a_n, which can be any algebraic integer, and the minimal number of terms of the sums (45) and (46) is two.

Equations (45) and (46) have a translation in the language of supertableaux. A detailed discussion will be given later, but it is straighforward to check that there are, at most, $n+1$ independent ways to add (suppress) a covariant (contravariant) box to a legal supertableaux of SU$(n|1)$ to obtain a new legal supertableau.

(3) Let us first consider the case where the original representation is an atypical irreducible representation as studied in Sec. 4. From Eqs. (7) and (9) we get the Kac–Dynkin parameters of such a representation

$$a_N \geqslant 1, \quad a_s = 0, \quad \text{for } s = N+1,\ldots,n-\bar{N}-1,$$

$$a_n = -A_{\bar{N}},$$

with $N + \bar{N} < n$, and we have seen that such a representation is uniquely associated with a supertableau with N covariant rows and $\bar{N}$ contravariant rows.

For $N+\bar{N}<n-1$ we obtain, at most, $N+\bar{N}+1$ terms in Eqs. (45) and (46), each term being associated with an atypical irreducible representation. The construction of these representations is trivial in the supertableau approach.

When $N+\bar{N}=n-1$ we obtain, at most, $N+\bar{N}=n-1$ terms associated with an atypical irreducible representation, and in all cases one term associated with a typical irreducible representation. In the language of supertableaux this term has $c_1+\bar{c}_1=n$, and this subset of typical supertableaux has been studied in Sec. 5.

(4) We now suppose that the original representation is a typical irreducible representation as studied in Sec. 5. The maximal number of terms in Eqs. (45) and (46) is $n+1$, and it corresponds to the maximal number of independent ways we have to add (suppress) a covariant (contravariant) or a contravariant (covariant) box to a legal typical supertableau, the result being again a legal supertableau.

In Eq. (45) either all terms are typical or two and only two terms become atypical, the other ones remaining typical. This last situation occurs when the original a_n parameter takes one of the values $-A_k-1$ assumed to be typical. It is then straightforward to check that the second term and one of the other ones are atypical. For Eq. (46) the situation is entirely analogous and the degeneracy arises when the original a_n takes one of the values $-A_k+1$ assumed to be typical. This property will be referred, in what follows, as the mechanism of pair production of atypical representations and atypical supertableaux.

7. ATYPICAL MIXED SUPERTABLEAUX $c_1+\bar{c}_1>n$

(1) We consider the mixed supertableaux studied in Sec. 5 for which the Kac–Dynkin parameter a_n takes one of the n possible atypical values $-A_k$ with $k=0,1,\ldots,n-1$. As a consequence, the two parameters c_1 and $\bar{c}_1$ are now determined when the length of the rows $b_j,\bar{b}_j$ is known.

For the mixed supertableaux of type I_0^j with $j=0,1,\ldots,n-1$, we have $c_1=n-j$, and from Eqs. (21) and (32) we get

$$\bar{c}_1=j+b_{n-j}+A_k-A_j.$$

Because of the positivity constraint, $c_1+\bar{c}_1>n$, we immediately see that k takes only $(n-j)$ values, $k=j,j+1,\ldots,n-1$.

For the mixed supertableaux of type I_1^j with $j=1,2,\ldots,n$, we have $\bar{c}_1=j$, and from Eqs. (28) and (37) we have

$$c_1=n-j+\bar{b}_j+A_{j-1}-A_k.$$

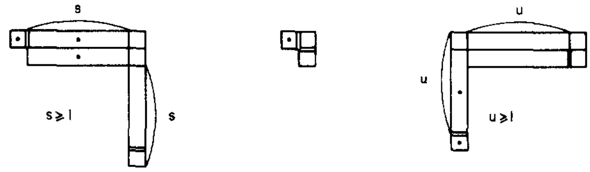


FIG. 16. Generalized atypical supertableaux of SU(2|1) for [1;0].

Again the positivity constraint $c_1+\bar{c}_1>n$ restricts k to j values $k=0,1,\ldots,j-1$.

(2) We now come back to Eqs. (45) and (46) in the degenerate case. Two terms, $\mathscr{T}_1$ and $\mathscr{T}_2$, become simultaneously atypical, and it is straightforward to study their $SU(n)\otimes U(1)$ content and to deduce, from that content, their dimension. This $SU(n)\otimes U(1)$ analysis exhibits, for $\mathscr{T}_1$ and for $\mathscr{T}_2$, two atypical components and one of these components is the same for $\mathscr{T}_1$ and $\mathscr{T}_2$:

$$\mathscr{T}_1=\mathscr{A}_1+\mathscr{A}_2,\quad \mathscr{T}_2=\mathscr{A}_2+\mathscr{A}_3.$$

This result is identical to that found by Marcu[11] in the SU(2|1) case, and his conclusion was that the four atypical components are the basis of a non-fully-reducible representation of SU(2|1) with the structure

$$\begin{tikzcd} & \mathscr{A}_1 \arrow[dr] & \\ \mathscr{A}_2 \arrow[ur] \arrow[dr] & & \mathscr{A}_2 \\ & \mathscr{A}_3 \arrow[ur] & \end{tikzcd}$$

We conjecture that the same result holds also for $SU(n|1)$.

(3) In the language of supertableaux the appearance of two atypical terms in Eqs. (45) and (46) correspond to the appearance of two atypical supertableaux T_1 and T_2. We have no consistent way to relate individually T_1 or T_2 to the terms $\mathscr{T}_1$ or $\mathscr{T}_2$ or, equivalently, to find a definite $SU(n)\otimes U(1)$ content of T_1 and T_2 separately.

However, it is formally possible to compute the dimension of T_1 and T_2 by using the determinant technique,[1] and we make the following observations:

(i) the results dim T_1 and dim T_2 are not always positive definite for a legal atypical mixed supertableau $c_1+\bar{c}_1>n$;

(ii) the sum dim T_1 + dim T_2 is positive definite;

(iii) the simple relation holds

$$\dim T_1+\dim T_2=\dim\mathscr{T}_1+\dim\mathscr{T}_2.$$

These features are clearly in favor of a non-fully-reducible representation with four atypical components, and we introduce the notation $[\![a_2]\!]$ for such a representation. In fact, they express the impossibility in this tensor representation to separate the components T_1 and T_2 of the full tensor, one of them being associated with a trace.

An atypical supertableau is only a part of a non-fully-

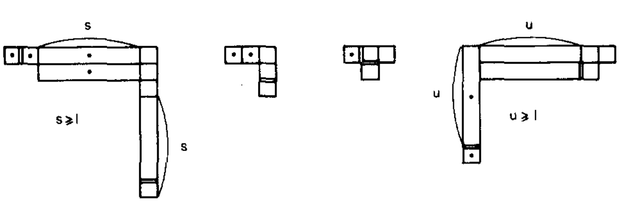


FIG. 15. Generalized atypical supertableaux of SU(2|1) for [2;0].

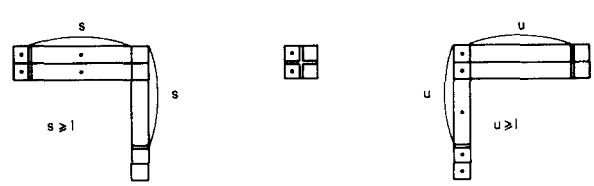


FIG. 17. Generalized atypical supertableaux of SU(2|1) for [0;0].

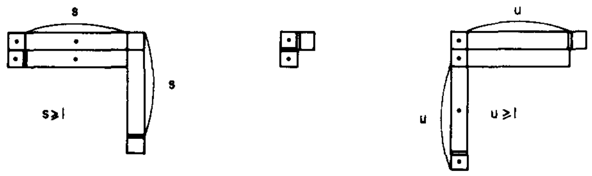


FIG. 18. Generalized atypical supertableaux of SU(2|1) for [0; − 1].

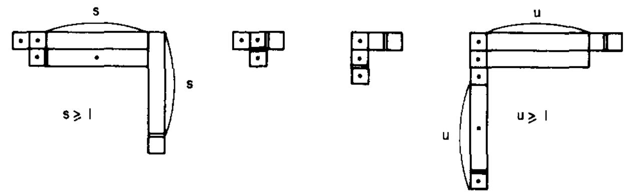


FIG. 19. Generalized atypical supertableaux of SU(2|1) for [1; − 2].

reducible representation of SU(n|1), and only the consideration of pairs of neighboring atypical supertableaux in the space S makes sense. We then introduce new generalized atypical mixed supertableaux incorporating the two atypical components whose sum is associated with a non-fully-reducible representation of SU(n|1). In a generalized mixed supertableau, double bars will indicate which ordinary atypical supertableaux are contained in the generalized one.

(4) Let us be a little more precise for the SU(2|1) case for which the non-full-reducibility has been proved. We only have two Kac–Dynkin parameters, $a_1 \geqslant 0$ and a_2, and the atypical values of a_2 are 0 and $-(1+a_1)$. We then have non-fully-reducible representations $[\![a_1;0]\!]$ and $[\![a_1;-1-a_1]\!]$. The four atypical components of these representations are

$$[\![a_1,0]\!] = \begin{tikzcd}[column sep=small] & \{a_1+1;0\} \arrow[dr] & \\ \{a_1;0\} \arrow[ur] \arrow[dr] & & \{a_1,0\} \\ & \{a_1-1;0\} \arrow[ur] & \end{tikzcd} \quad \text{for} \quad a_1 > 1,$$

$$[\![0;0]\!] = \begin{tikzcd}[column sep=small] & \{1;0\} \arrow[dr] & \\ \{0;0\} \arrow[ur] \arrow[dr] & & \{0;0\} \\ & \{0;-1\} \arrow[ur] & \end{tikzcd} \quad \text{self-contragradient dimension 8},$$

$$[\![a_1;-1-a_1]\!] = \begin{tikzcd}[column sep=small] & \{a_1+1;-2-a_1\} \arrow[dr] & \\ \{a_1;-1-a_1\} \arrow[ur] \arrow[dr] & & \{a_1;-1-a_1\} \\ & \{a_1-1;-a_1\} \arrow[ur] & \end{tikzcd} \quad \text{for} \quad a_1 \rangle 1,$$

$$[\![0;-1]\!] = \begin{tikzcd}[column sep=small] & \{1;-2\} \arrow[dr] & \\ \{0;-1\} \arrow[ur] \arrow[dr] & & \{0;-1\} \\ & \{0;0\} \arrow[ur] & \end{tikzcd} \quad \text{dimension 12}.$$

We have illustrated in Figs. 15–19 the five particular cases of generalized atypical supertableaux [2;0], [1;0], [0;0], [0; − 1], and [1; − 2], giving the full set of equivalent generalized supertableaux.

(5) For the general case of SU(n|1) thing are analogous but more complicated. The non-fully-reducible representation is defined by its four atypical components

$$\begin{tikzcd}[column sep=small] & \mathscr{A}_1 \arrow[dr] & \\ \mathscr{A}_2 \arrow[ur] \arrow[dr] & & \mathscr{A}_2. \\ & \mathscr{A}_3 \arrow[ur] & \end{tikzcd}$$

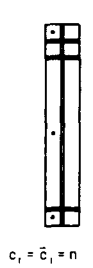


FIG. 20. Generalized atypical supertableau of SU(n|1) for [0,...,0;0].

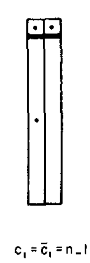


FIG. 21. Generalized atypical supertableau of SU(n|1) for [1,0,...,0; − 1].

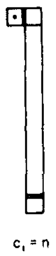


FIG. 22. Generalized atypical supertableau of SU(n|1) for [0,...,0,1;0].

Giving now an atypical component $\mathscr{A}_2$ by its Kac–Dynkin parameters $\mathscr{A}_2 = \{a_1, a_2, ..., a_{n-1}; -A_k\}$ with $k = 0,1,...,n-1$, we determine its two partners $\mathscr{A}_1$ and $\mathscr{A}_3$. The general result is the following:

$$\mathscr{A}_1 = \{a_1,...,a_{n-1-k}+1, a_{n-k}-1,...,a_{n-1}; -A_k+1\},$$

$$\mathscr{A}_3 = \{a_1,...,a_{n-1-k}-1, a_{n-k}+1,...,a_{n-1}; -A_k-1\}.$$

When $a_{n-k} = 0$, the term $\mathscr{A}_1$ has to be modified as follows: define $0 \leqslant r \leqslant k$ by $a_{n-k} = a_{n-k+1} = \cdots = a_{n-r-1} = 0$, $a_{n-r} \geqslant 1$. Then the new term $\mathscr{A}_1$ has $a_{n-r} \Rightarrow a_{n-r} - 1$, and the last parameter is now $-A_r + 1$.

When $a_{n-1-k} = 0$, the term $\mathscr{A}_3$ has to be modified. Defining now $k \leqslant s \leqslant n-1$ by $a_{n-1-k} = a_{n-2-k} = \cdots = a_{n-s} = 0$, $a_{n-1-s} \geqslant 1$. Then the new term $\mathscr{A}_3$ has $a_{n-1-s} \Rightarrow a_{n-1-s} - 1$, and the last parameter becomes $-A_s - 1$.

These results hold for any value of $k = 0,1,...,n-1$. Let us give now a few simple examples:

(i) $\mathscr{A}_2 = \{0,...,0;0\}$: we obtain

$$\mathscr{A}_1 = \{0,...,0;1-n\} \quad \text{and} \quad \mathscr{A}_3 = \{0,...,0,1;0\}.$$

$\mathscr{A}_1$ and $\mathscr{A}_3$ are contragradient atypical representations and $\mathscr{A}_2$ is a self-contragradient one. It follows that [0,...,0;0] is also self-contragradient and its dimension is 2^{n+1}.

(ii) $\mathscr{A}_2 = \{1,0,...,0;-1\}$: we obtain

$$\mathscr{A}_1 = \{0,...,0,1;1-n\} \quad \text{and} \quad \mathscr{A}_3 = \{1,0,...,0,1,0;0\}.$$

Again $\mathscr{A}_1$ and $\mathscr{A}_3$ are contragradient atypical representations, and $\mathscr{A}_2$ is a self-contragradient one. It follows that [1,0,...,0; -1] is also self-contragradient and its dimension is $n2^{n+1}$.

(iii) $\mathscr{A}_2 = \{0,...,0,1;0\}$: we obtain

$$\mathscr{A}_1 = \{0,...,0,2;0\} \quad \text{and} \quad \mathscr{A}_3 = \{0,...,0;0\}.$$

The dimension of [0,...,0,1;0] is $(n+1)2^n$.

(iv) $\mathscr{A}_2 = \{1,0,...,0,1;0\}$: we obtain

$$\mathscr{A}_1 = \{1,0,...,0,2;0\} \quad \text{and} \quad \mathscr{A}_3 = \{1,0,...,0;0\}.$$

The dimension of [1,0,...,0,1;0] is $(n^2+n-1)2^n$. We have illustrated these results in Figs. 20–23 by representing one generalized atypical supertableau of SU(n|1) for each example considered.

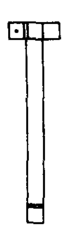


FIG. 23. Generalized atypical supertableau of SU(n|1) for [1,0,...,0,1;0].

8. CONCLUDING REMARKS

We have constructed the supertableaux of class I of SU(n|1), and they can be classified as follows:

(i) $c_1 + \bar{c}_1 < n$: we have the atypical covariant, contravariant, and mixed supertableaux uniquely associated with an atypical irreducible representation of SU(n|1);

(ii) $c_1 + \bar{c}_1 = n$: we have typical covariant, contravariant and mixed supertableaux uniquely associated with a typical irreducible representation of SU(n|1);

(iii) $c_1 + \bar{c}_1 > n$: we have two possibilites:

(α) The covariant, contravariant and mixed typical supertableaux are associated with typical irreducible representations of SU(n|1), and we have equivalences.

(β) The mixed atypical supertableaux are bounded by pairs into generalized mixed supertableaux associated to non-fully-reducible representations of SU(n|1) with four atypical components. Again we have equivalences.

The reduction of a tensor product of representations of SU(n|1) considered here can be obtained by using the supertableaux, and the ordinary rules of Young tableaux can be applied provided we take care of the case of mixed atypical supertableaux appearing in pairs and reinterpreted as forming generalized mixed atypical supertableaux.

The problem of equivalence of supertableaux has been discussed, and it appears only in the case $c_1 + \bar{c}_1 > n$. It happens that only the atypical irreducible representations are described by one supertableau. All the typical representations and the non-fully-reducible representations with four atypical components are described by an infinity of equivalent supertableaux. Incidentally, we have not found supertableaux corresponding to non-fully-reducible representations of SU(n|1) containing only two or three atypical components or representations where the Cartan operator Q is not diagonalized.[11]

The extension of these considerations to the superunitary groups SU(n|m) is in progress, but the classification of supertableaux is entirely analogous to that found here starting from the constraints of legality (1) and (2).

ACKNOWLEDGMENTS

One of the authros (M. G.) wishes to express his thanks to H. Ruegg and B. Morel for an interesting discussion and also to the CERN Theory Division for the hospitality extended to him.

APPENDIX A: SU(n) ⊗ U(1) DECOMPOSITION OF ATYPICAL SUPERTABLEAUX

We consider an atypical supertableau with N rows of b_j covariant boxes and $\bar{N}$ rows of contravariant boxes, $N + \bar{N} < n$, as represented in Fig. 1. The highest weight of this representation belongs to the SU(n) ⊗ U(1) component having the lowest eigenvalue of the U(1) generator Q. Such a component is obtained, from the original diagram, by transferring the first contravariant column of $\bar{N}$ boxes to the U(1) part as a row of $\bar{N}$ contravariant boxes. The result is shown in

Fig. 2. It is straightforward to determine the $(n-1)$ Kac–Dynkin parameters $a_1,a_2,...,a_{n-1}$ on the SU(n) part of this component, and we immediately get the relation (7) of the text. The computation of the lowest eigenvalue of Q is straightforward:

$$q_{\min} = \frac{1}{n}\left\{\sum_1^N b_j - \sum_1^{\bar N}(\bar b_j - 1)\right\} - \bar N.$$

Using now the Eqs. (8), we compute the two sums

$$\sum_1^N b_s = \sum_1^N s a_s$$

$$\sum_1^{\bar N} \bar b_s = \sum_1^{\bar N} s a_{n-s} = n\sum_1^{\bar N} a_{n-s} - \sum_1^{\bar N}(n-s)a_{n-s},$$

and we get

$$q_{\min} = \frac{1}{n}\sum_1^{n-1} s a_s - \bar N - \sum_1^{\bar N} a_{n-s} = \frac{1}{n}\sum_1^{n} s a_s.$$

The value obtained for a_n is atypical

$$a_n = -\bar N - \sum_1^{\bar N} a_{n-j} = -A_{\bar N}.$$

APPENDIX B: SU(n) ⊗ U(1) DECOMPOSITION OF TYPICAL SUPERTABLEAUX

(1) For supertableaux of type I^0 of the SU(n) ⊗ U(1) component containing the highest weight is shown in Figs. 4 and 6 (see also Ref. 6). We now compute for this component the eigenvalue of the U(1) operator Q. The result is simply

$$q_{\min} = \frac{1}{n}\sum_1^n b_j + \Delta, \tag{B1}$$

with

$\Delta = c_1 - n$ for covariant supertableaux I_1^0,

$\Delta = -\bar c_1$ for mixed supertableaux I_0^0.

The set of parameters $a_1,a_2,...,a_{n-1}$ is then trivially obtained as in the text, and, using Eq. (17), we get

$$\sum_1^n b_j = \sum_1^{n-1} s a_s + n b_n. \tag{B2}$$

The Kac–Dynkin parameter a_n is given by

$$a_n = q_{\min} - \frac{1}{n}\sum_1^{n-1} s a_s.$$

Using Eqs. (B1) and (B2), we get

$$a_n = b_n + \Delta.$$

The coordinate y_1 is given by Δ.

(2) For supertableaux of type I^n the SU(n) ⊗ U(1) component containing the highest weight is shown in Figs. 8 and 10 [see also Ref. 6]. We now compute for this component the eigenvalue of Q. The result is

$$q_{\min} = -\frac{1}{n}\sum_1^n(\bar b_j - 1) + \Delta \tag{B3}$$

with

$\Delta = -\bar c_1$ for contravariant supertableaux I_0^n,

$\Delta = c_1 - n$ for mixed supertableaux I_1^n.

The set of parameters $a_1,a_2,...,a_{n-1}$ is obtained as in the text, and, using Eq. (23), we get

$$\sum_1^n(\bar b_j - 1) = n(\bar b_1 - 1) - \sum_1^{n-1} s a_s. \tag{B4}$$

For the Kac–Dynkin parameter a_n, we obtain from Eqs. (B3) and (B4) the value

$$a_n = 1 - \bar b_1 + \Delta,$$

leading again to

$$y_1 = -\Delta.$$

(3) Now for mixed supertableaux of type I^j with $0<j<n$ the SU(n) ⊗ U(1) component containing the highest weight is shown in Figs. 12 and 14. We compute for this component the eigenvalue of Q

$$q_{\min} = \frac{1}{n}\left[\sum_1^{n-j} b_s - \sum_1^{j}(\bar b_s - 1)\right] + \Delta, \tag{B5}$$

with Δ given by

$\Delta = -\bar c_1$ for mixed supertableaux I_0^j,

$\Delta = c_1 - (n-j) - j = c_1 - n$ for mixed supertableaux I_1^j.

Again we immediately obtain the set of parameters $a_1,a_2,...,a_{n-1}$ as in the text. We now use the Eqs. (30) to compute the sums

$$\sum_1^{n-j} a_s = \sum_1^{n-j} s a_s + (n-j)\sum_{n-j+1}^{n-1} a_s + (n-j)(1-\bar b_1),$$

$$\sum_1^{j}(\bar b_s - 1) = -j(1-\bar b_1) + (n-j)\sum_{n-j+1}^{n-1} a_s - \sum_{n-j+1}^{n-1} s a_s.$$

Combining both results,

$$\sum_1^{n-j} b_s - \sum_1^{j}(\bar b_s - 1) = \sum_1^{n-1} s a_s + n(1-\bar b_1). \tag{B6}$$

Using Eqs. (B5) and (B6), we get

$$a_n = (1-\bar b_1) + \Delta,$$

corresponding again to

$$y_1 = -\Delta.$$

[1] A. B. Balantekin and I. Bars, J. Math. Phys. **22**, 1149,1810; (1981) **23**, 1239 (1982).

[2] A. B. Balantekin, J. Math. Phys. **23**, 486 (1982).

[3] Small letters indicate the superalgebra of the supergroup.

[4] I. Bars, *Proceedings of the School of Supersymmetry, Mexico 1981* (Plenum, New York, 1982).

[5] B. Morel and H. Ruegg, unpublished.

[6] I. Bars, B. Morel, and H. Ruegg, J. Math. Phys. **24**, 2253 (1983).

[7] V. G. Kac, Adv. Math. **26** (1), 8 (1977); V. G. Kac, Commun. Math. Phys. **53**, 31 (1977); V. G. Kac, *Lecture Notes in Mathematics*, Vol. 676 (Springer-Verlag, New York, 1978), p. 597.

[8] J. P. Hurni and B. Morel, J. Math. Phys. **24**, 157 (1983).

[9] M. Scheunert, W. Nahm, and V. Rittenberg, J. Math. Phys. **18**, 155 (1977).

[10] The supertableaux are contragradient if one is deduced from the second by replacing covariant (contravariant) boxes by contravariant (covariant) boxes.

[11] M. Marcu, J. Math. Phys. **21**, 1277 (1980).

Differentiable vectors and sharp momentum states of helicity representations of the Poincaré group

Bengt Ek and Bengt Nagel
Department of Theoretical Physics, Royal Institute of Technology, S-100 44 Stockholm, Sweden

(Received 8 November 1983; accepted for publication 6 January 1984)

This paper discusses some mathematical difficulties in handling sharp momentum eigenvectors for a massless helicity representation of the Poincaré group, related to the non-nuclearity of the space of differentiable vectors, and to the existence of singularities in the Lorentz group generators. A simple characterization of the nuclear space of differentiable vectors of the extension of the representation to a representation of the conformal group is given in terms of functions on the space $\mathbb{R}^4 - \{0\}$. Using a fibration of this space over the forward light cone (in momentum space), the singularity in the generators is shown to be related to the fact that the standard presentation of the helicity representations should be reformulated in terms of nontrivial sections over the light cone. The problem is partly identical with the one encountered in the study of an electron in a magnetic monopole field, the generator singularity taking the place of the Dirac string.

PACS numbers: 02.20. + b, 02.40. + m, 11.30. − j

I. INTRODUCTION

This paper is concerned with the helicity representations, i.e., the mass zero, "discrete spin" representations of (the covering of) the Poincaré group, and with their generalized sharp momentum state vectors.

Most physicists, of course, use sharp momentum states of photons and other massless particles without hesitation. A recent example is given by the paper of Weinberg and Witten,[1] where possible restrictions on the possibility of constructing quantum field theories with massless fields are discussed. We shall be concerned with two (seemingly unrelated) difficulties in the mathematical description of such sharp momentum states.

The standard way[2,3] to introduce generalized states in a group representation is via a "Gelfand triplet" $\mathscr{D} \subset \mathscr{H} \subset \mathscr{D}'$. Here $\mathscr{H}$ is the Hilbert space of our representation, and $\mathscr{D}$ is a dense invariant domain of definition for all the infinitesimal generators, equipped with a topology making it a nuclear topological space. The infinitesimal generators should be essentially self-adjoint on $\mathscr{D}$ and continuous as operators in $\mathscr{D}$; $\mathscr{D}'$, the topological dual of $\mathscr{D}$, is then a possible space of generalized state vectors. In particular it contains complete sets of generalized eigenvectors for all the generators.

The first difficulty with the helicity representations concerns the choice of the nuclear space $\mathscr{D}$. For many representations the natural choice is the *maximal* common invariant domain of the infinitesimal generators, i.e., the space of differentiable vectors of the representation, with a certain natural topology. For the helicity representations, however, this space of differentiable vectors is not nuclear.[3] Since a helicity representation can be uniquely extended to a representation of (the covering of) the conformal group SO(4,2),[4–6] we choose the nuclear space of differentiable vectors of this extended representation as our $\mathscr{D}$. The fact that the helicity representations can be extended to a larger semisimple group seems to be closely related to the non-nuclearity of the space of differentiable vectors.

The second difficulty is related to the singularity which appears for nonzero helicity in the expressions for the infinitesimal generators.[7–9] It turns out to be natural to consider our representation space as a Hilbert space of sections in a certain complex line bundle over the forward light cone in momentum space, rather than as a space of functions on the cone. As L^2-spaces these are equivalent, but the differentiable vectors are most naturally expressed as sections.

These line bundles are mathematically the same as the ones used in the description of "magnetic monopoles without strings.[10,11]" The state of a charged particle in the field of a magnetic monopole is described by a "wave section" rather than a wave function.

The line bundles involved can all be described using a certain projection $\dot{\mathbb{R}}^4 \to \dot{\mathbb{R}}^3$ ($\dot{\mathbb{R}}^k = \mathbb{R}^k - \{0\}$), which restricted to the unit sphere is the Hopf map $S^3 \to S^2$. The fiber over each point of $\dot{\mathbb{R}}^3$ is a circle. Our analysis is simplified by describing the sections of the representation space as functions on $\dot{\mathbb{R}}^4$ with a certain equivariance condition along the fibers (see Ref. 5).

The nuclear space of differentiable vectors of the conformal group then turns out to be the set of equivariant functions in $\mathscr{S}(\dot{\mathbb{R}}^4)$, the space of rapidly decreasing test functions. The generalized states can then be identified with equivariant elements of $\mathscr{S}'(\mathbb{R}^4)$, the space of tempered distributions.

Here two remarks are in place.

First, even for helicity zero, where the state vectors can be considered as ordinary functions of $\mathbf{p} \in \dot{\mathbb{R}}^3$ (the forward light cone in p-space), not all the differentiable vectors will correspond to functions which are differentiable at $\mathbf{p} = 0$. This is because our projection $\dot{\mathbb{R}}^4 \to \dot{\mathbb{R}}^3$ is quadratic.

Second, since the differentiable vectors are not necessarily zero at the origin, we will have generalized state vectors with support at the origin describing non-wave-like states of the "field."

The structure of the set of sharp momentum states is then also clear. A δ-function in a point of $\dot{\mathbb{R}}^4$ describes a state with the corresponding momentum, and δ-functions in different points on the same fiber over $\dot{\mathbb{R}}^3$ correspond to the same state, up to a phase. This means that the **p**-dependence of the momentum states should not be considered as a function, but rather as a section in a certain bundle.

More explicitly, if we try to fix the phase at each **p** and let $|\mathbf{p}\rangle$ denote the corresponding generalized state vector of momentum **p**, we shall find that the mapping $\dot{\mathbb{R}}^3 \ni \mathbf{p} \to |\mathbf{p}\rangle \in \mathscr{D}'$ cannot be continuous for nonzero helicity. This is not at all surprising, it generalizes the fact that we cannot choose a continuously varying plane of polarization for photons in all points of the sphere (there is no continuous field of nonzero tangent vectors on a sphere).

This article is organized as follows. In Sec. II we deduce the form of the helicity representations using the method of induced representations. There the projection $\dot{\mathbb{R}}^4 \to \dot{\mathbb{R}}^3$ will also be described. In Sec. III we give the local (infinitesimal) form of the representations, exhibit the extension to the conformal group, and show how our form of the representations corresponds to that of Refs. 7–9. The action of the infinitesimal generators on functions expanded in "monopole harmonics" is also given and the analogy with magnetic monopoles is presented. In Sec. IV we deal with the concepts of differentiable vectors and nuclearity, and find the spaces of differentiable vectors of our representations of the conformal group. Generalized eigenstates of the momentum operators are presented in Sec. V and a brief discussion of the relation between states of zero momentum and space-time independent fields is given. The results are summarized and discussed in Sec. VI.

II. THE REPRESENTATIONS

We shall study the $(0,\lambda)$, i.e., zero mass, finite helicity λ, representations of the double covering group $\overline{\mathscr{P}}$ of the proper Poincaré group, where

$$\overline{\mathscr{P}} = \mathbb{R}^4 \times \mathrm{SL}(2,\mathbb{C}) = \{(a,A);\ a\in\mathbb{R}^4,\ A\in\mathrm{SL}(2,\mathbb{C})\}. \quad (2.1)$$

The covering homomorphism $A \to \Lambda(A)$ of SL(2,$\mathbb{C}$) onto the proper Lorentz group $\mathscr{L}'_+$ is defined as usual by

$$H(\Lambda(A)p) = AH(p)A^+, \quad (2.2)$$

where

$$H(p) = p^0\sigma_0 + \mathbf{p}\cdot\boldsymbol{\sigma} \quad (2.3)$$

is a parametrization of Minkowski p-space $\{p=(p^0,\mathbf{p})\}$ by 2×2 Hermitian matrices, σ_0 is the 2×2 unit matrix, and $\boldsymbol{\sigma}$ the Pauli matrices. The scalar product in Minkowski space is $ap = a^0p^0 - \mathbf{a}\cdot\mathbf{p}$.

The theory of induced representations is presented in many books and articles. We mention only one rather recent reference,[12] which also gives a treatment of induced representations of Lie groups in terms of sections of fiber bundles, appropriate for our purpose. For the special case of induced representations of the Poincaré group, which is also treated in many places, we can mention, apart from the original article by Wigner,[13] the articles by Guillot and Petit,[8] which treat explicitly the form of the representations in terms of functions on the group SL(2,$\mathbb{C}$).

In the induced representation framework the zero mass representations correspond to a stability point in p-space situated on the light cone, say $\mathring{p} = (1,0,0,1)$. This gives

$$H(\mathring{p}) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix},$$

and the corresponding stability group—the little group L in this case—is

$$\overline{E}(2) = \left\{ \begin{pmatrix} e^{i\phi} & z \\ 0 & e^{-i\phi} \end{pmatrix};\ 0 \leqslant \phi < 2\pi,\ z\in\mathbb{C} \right\}. \quad (2.4)$$

Given an inducing representation $U_V(L)$ of the little group L in a vector space V, the corresponding representation of $\overline{\mathscr{P}}$ can be realized on the space of equivariant functions from SL(2,$\mathbb{C}$) to V by the formula

$$[U((a,A))F](B) = \exp[ia\cdot\Lambda(B)\mathring{p}]F(A^{-1}B). \quad (2.5)$$

Equivariance means that for $B\in\mathrm{SL}(2,\mathbb{C})$, $C\in L$,

$$F(BC) = U_V(C^{-1})F(B). \quad (2.6)$$

It should be noted that this form of the induced representation is uniquely fixed by the choice of stability point $\mathring{p}$ and inducing representation $U_V(L)$ of the little group corresponding to $\mathring{p}$.

Equation (2.6) means that the function F is determined by giving its value for one element of each of the left cosets of SL(2,$\mathbb{C}$) with respect to the subgroup L, or expressed in fiber bundle terminology, F is a section of the fiber bundle SL(2,$\mathbb{C}$) $\times_L V$ associated to the principal bundle [SL(2,$\mathbb{C}$), SL(2,$\mathbb{C}$)/L,L]. The base space of both bundles, the homogeneous space $X = \mathrm{SL}(2,\mathbb{C})/L$, can be identified with the orbit $\{\Lambda(A)\mathring{p}; A\in\mathrm{SL}(2,\mathbb{C})\}$ of SL(2,$\mathbb{C}$) in Minkowski space through the point $\mathring{p}$.

In the case of a massive representation this orbit is the mass hyperboloid $\{p; p^2 = m^2, p^0 > 0\}$ and is the same as—more precisely, is diffeomorphic to—three-space $\mathbb{R}^3$. Since this space is contractible, every fiber bundle with $\mathbb{R}^3$ as a base is trivial, i.e., of the form $\mathbb{R}^3 \times \mathscr{F}$, where $\mathscr{F}$ is the fiber. This corresponds to the possibility of choosing the set of A's parametrizing the orbit $\{\Lambda(A)\mathring{p}\}$ in a smooth way. In other words, there exists a global smooth section in the principal fiber bundle [SL(2,$\mathbb{C}$), SL(2,$\mathbb{C}$)/L,L] (not the same L and $\mathring{p}$ as before!). Let us call the parametrizing function $A(p)$; here p varies over the orbit, and we have $\Lambda(A(p))\mathring{p} = p$. This global cross section defines an isomorphism relating a function $f: X = \mathrm{SL}(2,\mathbb{C})/L \to V$ to an equivariant function $F: \mathrm{SL}(2,\mathbb{C}) \to V$ through the relation $f(p) = F(A(p))$. The representation (2.5) on equivariant functions, when transferred to functions on the homogeneous space, then takes the familiar form

$$[U((a,A))f](p) = e^{ia\cdot p}U_V[A(p)^{-1}AA(\Lambda(A)^{-1}p)]f(\Lambda(A)^{-1}p). \quad (2.7)$$

The representation space of the corresponding unitary representation is the space $L^2(\mathbb{R}^3, d\mathbf{p}/p^0, V)$ of square integrable V-valued functions on the mass hyperboloid, equipped with the invariant measure. The explicit form of the representation depends, besides on $\mathring{p}$ and $U_V(L)$, on the choice of cross section $A(p)$. Different choices give unitary equivalent forms of the representation as do different choices of $\mathring{p}$ on the mass hyperboloid, provided the associated representations $U_V(L)$ and $U_V(L')$ of the conjugated little groups L and L' are equivalent in an obvious way.

The set of differentiable vectors turns out to be the set of C^{∞}, rapidly decreasing V-valued functions on the mass hyperboloid. Since V is finite dimensional, this set of functions can easily be given a topology making it a nuclear space, and the generalized eigenvectors of the group generators are then contained in the set of tempered V-valued distributions on the mass hyperboloid.

We now return to the mass zero representations. In this case the homogeneous space $SL(2,\mathbb{C})/\overline{E}(2)$ is diffeomorphic to the pointed cone $\{p;p^2=0,p^0>0\}$, or, equivalently, to three-space with the origin removed, $\dot{\mathbb{R}}^3$. The principal bundle is nontrivial in this case, which means that it is not possible to choose a (smooth or even continuous) cross section $A(p)$ covering the whole base space. There must be discontinuities along a curve. In one standard choice this singularity curve is taken to be the line $\{p=(\kappa,0,0,-\kappa), \kappa>0\}$ through the "south pole"; on this line (the formula corresponding to) (2.7) does not make sense. As far as the definition of the unitary representation is concerned, this lower-dimensional singularity does not matter, but when one is interested in the proper definition of the infinitesimal generators and the differentiable vectors it is necessary to face the fact that (for $\lambda\neq 0$) one really has to do, not with ordinary V-valued functions, but with nontrivial sections in the associated vector bundle $SL(2,\mathbb{C})\times_L V$, which sections have to be defined on (at least) two coordinate patches over $\dot{\mathbb{R}}^3$. In the study of the infinitesimal generators a domain of these operators is sometimes given as the set of rapidly vanishing C^{∞} functions on $\mathbb{R}^3$ which vanish also in a small cylinder containing the negative z axis, along which the generators are singular. Although this gives a subset of differentiable vectors invariant under the action of the generators, this subset is neither dense in representation space nor invariant under finite group transformations, and is not very useful in defining generalized eigenvectors of the generators.

Since for the helicity representations the inducing representation of the little group $\overline{E}(2)$ is one-dimensional—i.e., V is one-dimensional—the space of the induced representation is a space of sections of a complex line bundle over the base space $\dot{\mathbb{R}}^3$. Since $\dot{\mathbb{R}}^3$ can be written $S^2\times\mathbb{R}_+$, where $\mathbb{R}_+$ is the open positive half-axis, the nontrivial part of the bundle is a complex line bundle over S^2. These are well-known, and enter, e.g., in the geometric quantization of a system consisting of a particle moving on a sphere. The corresponding problem on $\dot{\mathbb{R}}^3$ instead of S^2 appears—apart from in the case studied here—when one treats the quantum mechanics of a charged particle, e.g., an electron, in the field of a point magnetic monopole situated at the origin.[10,11] The singularity of the generators alluded to above and exhibited in Sec. III is the exact equivalent of the "Dirac string" containing the singularity of the vector potential representing the magnetic field of the monopole. The different possible complex line bundles are indexed by the integers; these integers appear as the monopole charges or as the possible values of 2λ in our case.

In the study of the differentiable vectors of a $(0,\lambda)$ representation it is convenient to use a form of the representation intermediate between (2.5) and (2.7), namely a form realized on a set of equivariant functions on $\dot{\mathbb{R}}^4$. This form is related to the so-called KAN factorization (or global Iwasawa decomposition, see, e.g., Ref. 12 or 14) of the group $SL(2,\mathbb{C})$, which shows that this group is, as a manifold, diffeomorphic to the product $S^3\times\mathbb{R}_+\times\mathbb{R}^2$, where the last factor is actually the translation part of the little group $\overline{E}(2)$, given by z in (2.4). This part is represented trivially in the inducing representation in the helicity case (but not in the "continuous spin" representations).

The KAN factorization expresses an element of $SL(2,\mathbb{C})$ uniquely in the form

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \beta & -\bar{\gamma} \\ \gamma & \bar{\beta} \end{pmatrix}\begin{pmatrix} \mu & 0 \\ 0 & 1/\mu \end{pmatrix}\begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & b' \\ c & d' \end{pmatrix}\begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix}, \tag{2.8}$$

where

$$\mu=\sqrt{|a|^2+|c|^2},\quad \beta=a/\mu,\quad \gamma=c/\mu,$$
$$z=(\bar{a}b+\bar{c}d)/\mu^2. \tag{2.9}$$

Here $\mu\in\mathbb{R}_+$, $z\in\mathbb{C}$, and β and γ fulfill $|\beta|^2+|\gamma|^2=1$.

Since the first factor on the first rhs of (2.8) is an element of $SU(2)\approx S^3$, the relation shows the diffeomorphism $SL(2,\mathbb{C})\approx S^3\times\mathbb{R}_+\times\mathbb{R}^2\approx\dot{\mathbb{R}}^4\times\mathbb{R}^2$. The last equality in (2.8) shows that the $\dot{\mathbb{R}}^4$ part of $SL(2,\mathbb{C})$ is parametrized by the first column (a,c) of the $SL(2,\mathbb{C})$ matrix.

For the representation $(0,\lambda)$—where 2λ is an integer—the inducing $\overline{E}(2)$ representation is

$$U_V\left[\begin{pmatrix} e^{i\alpha/2} & z \\ 0 & e^{-i\alpha/2} \end{pmatrix}\right]=e^{i\lambda\alpha}. \tag{2.10}$$

From (2.6) we conclude that if

$$C=\begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix},$$

then

$$F(BC)=F(B), \tag{2.11}$$

i.e., F is constant on every fiber $\mathbb{R}^2$ over $\dot{\mathbb{R}}^4$.

Changing notation and defining $\dot{\mathbb{R}}^4=\dot{\mathbb{C}}^2$ as the space of complex nonzero two-spinors $u=(u_1,u_2)$ $[=(a,c)]$, we define equivariant functions on $\dot{\mathbb{C}}^2$ from (2.6) by

$$g(e^{i\alpha/2}u)=e^{-i\lambda\alpha}g(u). \tag{2.12}$$

The invariant volume element on $SL(2,\mathbb{C})$ underlying the (unitary) representation (2.5) corresponds to the Euclidean volume element on $\dot{\mathbb{R}}^4$, and (2.5) takes the form

$$[U((a,A))g](u)=e^{ia\cdot p(u)}g(A^{-1}u). \tag{2.13}$$

Here $p(u)$ is defined by $H(p(u))=2uu^+$, or in component form,

$$p(u)=(|u_1|^2+|u_2|^2,u_1\bar{u}_2+\bar{u}_1u_2,$$
$$i(u_1\bar{u}_2-\bar{u}_1u_2),|u_1|^2-|u_2|^2). \tag{2.14}$$

Equation (2.14) gives an explicit expression for the projection map from $u\in\dot{\mathbb{R}}^4$ to the point $\mathbf{p}\in\dot{\mathbb{R}}^3$; all points on the circle $\{e^{i\alpha/2}u;\ 0\leqslant\alpha<4\pi\}$ project on the same $\mathbf{p}$. The radius $|\mathbf{p}|$ corresponds to $|u_1|^2+|u_2|^2$.

The essential part of the fibration of $\dot{\mathbb{R}}^4\approx S^3\times\mathbb{R}_+$ over $\dot{\mathbb{R}}^3\approx S^2\times\mathbb{R}_+$ is the so-called Hopf fibration (S^3,S^2,S^1), or, in group language, the principal fiber bundle [SU(2), SU(2)/U(1), U(1)] of SU(2) over the homogeneous space of left cosets of SU(2) with respect to a U(1) subgroup. The projection map $S^3\to S^2$—the Hopf map—of this fibration can be realized by combining the map $S^3\to\mathbb{C}$ given by $(u_1,u_2)\to u_2/u_1$ with a stereographic map from the complex plane $\mathbb{C}$ to the unit sphere S^2.

III. INFINITESIMAL GENERATORS AND EXTENSIONS OF THE REPRESENTATIONS

A. The infinitesimal generators on equivariant functions

To find explicit expressions for the infinitesimal generators of the representations presented in Sec. II we introduce real Euclidean coordinates $(\eta_1,\eta_2,\eta_3,\eta_4)$ and "spherical" coordinates (η,θ,ϕ,ψ) on $\dot{\mathbb{C}}^2=\dot{\mathbb{R}}^4$ as follows:

$$\begin{aligned}
&u_1=\eta_4-i\eta_3=\eta\cos\tfrac{1}{2}\theta e^{-i(\phi+\psi)/2},\\
&u_2=\eta_2-i\eta_1=\eta\sin\tfrac{1}{2}\theta e^{i(\phi-\psi)/2},\\
&\eta\in\mathbb{R}_+,\quad \theta\in[0,\pi],\quad \phi\in[0,2\pi),\quad \psi\in[0,4\pi).
\end{aligned}\tag{3.1}$$

Then $\eta^2=\eta_1^2+\eta_2^2+\eta_3^2+\eta_4^2$ and $\mathbf{p}(u)$, the image of u under the projection $\dot{\mathbb{C}}^2\to\dot{\mathbb{R}}^3$ defined in (2.14) has spherical coordinates (η^2,θ,ϕ).

A normalized u (i.e., $\eta=1$) corresponds to an element of SU(2) [choose $z=0,\mu=1$ in (2.8)], and (ϕ,θ,ψ) are then the Euler angles defined to give the following element of SU(2):

$$e^{-i(1/2)\phi\sigma_z}e^{-i(1/2)\theta\sigma_y}e^{-i(1/2)\psi\sigma_z}.\tag{3.2}$$

The Euclidean volume element is

$$d^4\eta=d\eta_1\,d\eta_2\,d\eta_3\,d\eta_4=\frac{1}{16}\,d\psi\,\frac{d^3\mathbf{p}}{p},\tag{3.3}$$

where $\mathbf{p}=\mathbf{p}(u)$, $p=|\mathbf{p}|$. (N.B. This p is different from the four-vector p used in Sec. II!)

The condition (2.12) for equivariance in the $(0,\lambda)$ representation becomes

$$g(\eta,\theta,\phi,\psi-\alpha)=e^{-i\lambda\alpha}g(\eta,\theta,\phi,\psi),\tag{3.4}$$

i.e., our representation space is the closed subspace of $L^2(\mathbb{R}^4)$ consisting of functions for which this condition is fulfilled almost everywhere.

If we define the infinitesimal generators P, $\mathbf{J}$, $\mathbf{N}$ of the representation (2.13) by

$$\begin{aligned}
&U((a,I))=e^{ia\cdot P}=e^{i(a^0P^0-\mathbf{a}\cdot\mathbf{P})},\quad a\in\mathbb{R}^4,\\
&\left.\begin{aligned}U((0,e^{-i(1/2)\boldsymbol{\alpha}\cdot\boldsymbol{\sigma}}))=e^{-i\boldsymbol{\alpha}\cdot\mathbf{J}},\\ U((0,e^{(1/2)\boldsymbol{\beta}\cdot\boldsymbol{\sigma}}))=e^{-i\boldsymbol{\beta}\cdot\mathbf{N}},\end{aligned}\right\}\quad \boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^3,
\end{aligned}\tag{3.5}$$

we find, with the notation $\partial_i=\partial/\partial\eta_i$,

$$\begin{aligned}
P^0&=p^0=\eta^2,\\
\mathbf{P}&=\mathbf{p}=(2(\eta_1\eta_3+\eta_2\eta_4),2(-\eta_1\eta_4+\eta_2\eta_3),\\
&\quad -\eta_1^2-\eta_2^2+\eta_3^2+\eta_4^2),\\
\mathbf{J}&=(i/2)(-\eta_4\partial_1+\eta_3\partial_2-\eta_2\partial_3+\eta_1\partial_4,\\
&\quad -\eta_3\partial_1-\eta_4\partial_2+\eta_1\partial_3+\eta_2\partial_4,\\
&\quad \eta_2\partial_1-\eta_1\partial_2-\eta_4\partial_3+\eta_3\partial_4),\\
\mathbf{N}&=(i/2)(-\eta_3\partial_1-\eta_4\partial_2-\eta_1\partial_3-\eta_2\partial_4,\\
&\quad \eta_4\partial_1-\eta_3\partial_2-\eta_2\partial_3+\eta_1\partial_4,\\
&\quad \eta_1\partial_1+\eta_2\partial_2-\eta_3\partial_3-\eta_4\partial_4).
\end{aligned}\tag{3.6}$$

B. Extensions of the representations

The extensions of these representations to the conformal algebra so(4,2)[4,5] then have the following form for the generators D and K:

$$\begin{aligned}
D&=\frac{i}{2}\eta_i\,\partial_i+i=\frac{i}{2}\eta\frac{\partial}{\partial\eta}+i,\\
K_0&=-\tfrac{1}{4}\partial_i\,\partial_i=-\tfrac{1}{4}\partial^2,\\
\mathbf{K}&=\tfrac{1}{4}(2(\partial_1\partial_3+\partial_2\partial_4),\ 2(-\partial_1\partial_4+\partial_2\partial_3),\\
&\quad -\partial_1^2-\partial_2^2+\partial_3^2+\partial_4^2).
\end{aligned}\tag{3.7}$$

They have been chosen to satisfy the following commutation relations,[9] with

$$\begin{aligned}
&M^{jk}=J_i\epsilon_{ijk},\quad M^{0i}=N_i,\quad M^{\mu\nu}=-M^{\nu\mu},\\
&K^\mu=(K^0,\mathbf{K}),\quad P^\mu=(P^0,\mathbf{P});
\end{aligned}\tag{3.8}$$

$$\begin{aligned}
&[M^{\mu\nu},M^{\lambda\sigma}]=i(g^{\nu\lambda}M^{\mu\sigma}-g^{\nu\sigma}M^{\mu\lambda}-g^{\mu\lambda}M^{\nu\sigma}+g^{\mu\sigma}M^{\nu\lambda}),\\
&[M^{\mu\nu},P^\lambda]=i(g^{\nu\lambda}P^\mu-g^{\mu\lambda}P^\nu),\quad [P^\mu,P^\nu]=0,\\
&[M^{\mu\nu},K^\lambda]=i(g^{\nu\lambda}K^\mu-g^{\mu\lambda}K^\nu),\quad [K^\mu,K^\nu]=0,\\
&[K^\mu,P^\nu]=2i(g^{\mu\nu}D-M^{\mu\nu}),\quad [M^{\mu\nu},D]=0,\\
&[D,P^\mu]=iP^\mu,\quad [D,K^\mu]=-iK^\mu.
\end{aligned}\tag{3.9}$$

To see that this extension of the $(0,\lambda)$ representation is not only on the level of algebras, but can be integrated to a representation of the covering group of the conformal group SO(4,2), we study a certain representation of the symplectic algebra sp(8,$\mathbb{R}$) on all of $L^2(\mathbb{R}^4)$.[15,16]

The 36-dimensional algebra sp(8,$\mathbb{R}$) consists of 8×8 real matrices of the form

$$\begin{pmatrix}P & Q\\ R & S\end{pmatrix},$$

with P, Q, R, and S 4×4 matrices such that $S=-P^t$, $R=R^t$, $Q=Q^t$. Here A^t denotes the transpose of A.

It is not difficult to see that then the symmetric operators

$$\eta^tR\eta+(i/2)(\eta^tS\,\partial-\partial^tP\,\eta)+\tfrac{1}{4}\,\partial^tQ\partial,\tag{3.10}$$

where η and ∂ are interpreted as column vectors, define a representation of i sp(8,$\mathbb{R}$) on $L^2(\mathbb{R}^4)$. It contains a subalgebra isomorphic to i so(4,2), and the restriction of this subalgebra to the subspace of λ-equivariant functions is exactly the $(0,\lambda)$ representation given above.

To prove that our sp(8,$\mathbb{R}$) representation—and hence also the so(4,2) representation—can be integrated, we apply the well-known criterion due to Nelson.[17]

Suppose $\{X_1, \ldots, X_d\}$ is a set of symmetric operators on a Hilbert space $\mathscr{H}$, with a common invariant dense domain $\mathscr{D}$ and such that $\{iX_1, \ldots, iX_d\}$ define a representation of a real Lie algebra with corresponding simply connected Lie group G.

If then the Nelson operator $A = \Sigma X_n^2$ is essentially self-adjoint on $\mathscr{D}$, there is a unique unitary representation of G on $\mathscr{H}$ with generators $X_1, \ldots, X_d$.

Apart from repetitions, the following operators form a basis for our sp(8,ℝ) representation:

$$\eta_i \eta_j, i(\eta_i \partial_j + \partial_j \eta_i)/2\sqrt{2},$$
$$\partial_i \partial_j/4; \quad i,j = 1, \ldots, 4. \tag{3.11}$$

We then get the Nelson operator as

$$A = (-\partial^2/4 + \eta^2)^2 + \tfrac{3}{2}. \tag{3.12}$$

If we choose for $\mathscr{D}$ the set $\mathscr{S}(\mathbb{R}^4)$ of C^∞, rapidly decreasing (with all derivatives) functions on $\mathbb{R}^4$, this set fulfills the conditions of Nelson's criterion. That it is a common invariant domain of all the generators is obvious, and that A, being essentially the square of the isotropic harmonic oscillator Hamiltonian in four dimensions, is essentially self-adjoint on $\mathscr{D}$ is well known; it follows most easily from the fact that the eigenfunctions of A, being Hermite functions, all belong to $\mathscr{D}$.

This proves the integrability of our sp(8,ℝ) representation, and hence *a fortiori* of the $(0,\lambda)$ representation of so(4,2).

Now we introduce the operators **I**, which are generators for our sp(8,ℝ) representation but not for so(4,2):

$$\begin{aligned}\mathbf{I} = (i/2)(&\eta_4 \partial_1 + \eta_3 \partial_2 - \eta_2 \partial_3 - \eta_1 \partial_4,\\ &-\eta_3 \partial_1 + \eta_4 \partial_2 + \eta_1 \partial_3 - \eta_2 \partial_4,\\ &\eta_2 \partial_1 - \eta_1 \partial_2 + \eta_4 \partial_3 - \eta_3 \partial_4)\end{aligned} \tag{3.13}$$

(i.e., like **J** except for a change of sign of η_4 and ∂_4).

Then

$$[I_i, I_j] = i\epsilon_{ijk} I_k, \quad [I_i, J_j] = 0, \quad \mathbf{I}^2 = \mathbf{J}^2, \tag{3.14}$$

i.e., the **I** generate a su(2) subalgebra of sp(8,ℝ) commuting with our su(2) subalgebra of so(4,2). Both **J** and **I** can be considered as acting on $L^2(S^3)$ and with our identification of S^3 with SU(2), **J** and **I** correspond to the left and right regular representations, respectively.

One finds

$$I_3 = i\frac{\partial}{\partial\psi}, \tag{3.15}$$

I_3 commutes with our so(4,2) subalgebra, and the space of λ-equivariant functions in $L^2(\mathbb{R}^4)$ is exactly the eigenspace of I_3 corresponding to eigenvalue $-\lambda$.

As for the commutation relations between **I** and the so(4,2) generators, one finds that the "**J**-scalars" P^0, D, K^0 are also "**I**-scalars," i.e., they commute with **I**, and that each component of the "J-vectors" **P**, **N**,**K** is the three-component of an "**I**-vector." The 36 (including **I** and **J**) operators so obtained form a basis for our sp(8,ℝ) representation.

C. The infinitesimal generators on "functions" on the cone

To arrive at the "usual" form of the infinitesimal generators expressed in **p**,[7,9] we first transform the expressions above to coordinates **p**, ψ or (p,θ,ϕ,ψ) in $\mathbb{R}^4$. The results are

$$\begin{aligned}
&P^0 = p = |\mathbf{p}|, \quad \mathbf{P} = \mathbf{p},\\
&\mathbf{J} = -i\mathbf{p}\times\nabla - \frac{i}{\sin\theta}\mathbf{e}_\rho\frac{\partial}{\partial\psi},\\
&\mathbf{N} = -ip\nabla + i\cot\theta\,\mathbf{e}_\phi\frac{\partial}{\partial\psi},\\
&D = i(\mathbf{p}\cdot\nabla + 1) = i\left(p\frac{\partial}{\partial p} + 1\right),\\
&K^0 = -p\nabla^2 - \frac{1}{p\sin\theta}\frac{\partial^2}{\partial\psi^2} + \frac{2\cos\theta}{p\sin^2\theta}\frac{\partial^2}{\partial\phi\,\partial\psi},\\
&\mathbf{K} = -\mathbf{p}\nabla^2 + 2(\mathbf{p}\cdot\nabla)\nabla + 2\nabla + \frac{1}{p\sin^2\theta}(\mathbf{e}_\mathbf{p} - 2\cos\theta\,\mathbf{e}_z)\\
&\qquad\times\frac{\partial^2}{\partial\psi^2} + \frac{2}{p\sin^2\theta}\left(-p\sin\theta\,\mathbf{e}_\phi\frac{\partial}{\partial p_z} + \mathbf{e}_z\frac{\partial}{\partial\phi}\right)\frac{\partial}{\partial\psi}.
\end{aligned} \tag{3.16}$$

Here $\nabla = (\partial/\partial p_x, \partial/\partial p_y, \partial/\partial p_z)$, $\mathbf{e}_\rho = (\cos\phi, \sin\phi, 0)$, $\mathbf{e}_\mathbf{p} = \mathbf{p}/p$, $\mathbf{e}_\phi = (-\sin\phi, \cos\phi, 0)$, and $\mathbf{e}_z = (0,0,1)$.

Also,

$$\begin{aligned}
&I_3 = i\frac{\partial}{\partial\psi},\\
&I_\pm = I_1 \pm iI_2 = -e^{\mp i\psi}\left(\pm\frac{\partial}{\partial\theta} + \frac{i}{\sin\theta}\frac{\partial}{\partial\phi}\right.\\
&\qquad\left. - i\cot\theta\frac{\partial}{\partial\psi}\right).
\end{aligned} \tag{3.17}$$

These operators still act on equivariant functions on $\mathbb{R}^4$. As described in Sec. II, we arrive at a description acting on functions on the forward light cone (or $\dot{\mathbb{R}}^3$) by choosing smooth local sections in the bundle $(\dot{\mathbb{R}}^4, \dot{\mathbb{R}}^3, S^1)$. The "standard" choices are $\psi = \mp\phi$, which give charts a and b, respectively, and which are defined and smooth except on the negative and positive p_z axis, respectively.

For the chart a, for any λ-equivariant $g \in L^2(\mathbb{R}^4)$, we write

$$g_a(\eta^2, \theta, \phi) = g(\eta, \theta, \phi, -\phi), \quad \theta \neq \pi, \tag{3.18}$$

and find

$$g(\eta, \theta, \phi, \psi) = g_a(\mathbf{p})e^{i\lambda(\phi+\psi)}, \quad \theta \neq \pi. \tag{3.19}$$

The generators' action on $g_a \in L^2(\mathbb{R}^3, d\mathbf{p}/p)$, i.e., in the chart a, is then

$$\begin{aligned}
&P_a^0 = p, \quad \mathbf{P}_a = \mathbf{p},\\
&\mathbf{J}_a = -i\mathbf{p}\times\nabla + \lambda\frac{\mathbf{p} + p\mathbf{e}_z}{p + p_z},\\
&\mathbf{N}_a = -ip\nabla + \lambda\frac{\mathbf{e}_z\times\mathbf{p}}{p + p_z},\\
&D_a = i\left(p\frac{\partial}{\partial p} + 1\right),\\
&K_a^0 = -p\nabla^2 - \frac{2i\lambda}{p + p_z}\frac{\partial}{\partial\phi} + \frac{2\lambda^2}{p + p_z},\\
&\mathbf{K}_a = -\mathbf{p}\nabla^2 + 2(\mathbf{p}\cdot\nabla)\nabla + 2\nabla + 2i\lambda\frac{\mathbf{p} + p\mathbf{e}_z}{p + p_z}\times\nabla\\
&\qquad - \frac{2\lambda^2}{p + p_z}\mathbf{e}_z.
\end{aligned} \tag{3.20}$$

These expressions are the same as those given in Ref. 9.

D. A basis in the representation space

To construct a suitable basis for the representations of the conformal algebra so(4,2), note that it contains two commuting subalgebras: a su(2) algebra spanned by $\mathbf{J}$ and a su(1,1) algebra spanned by P^0, K^0, D.

For the latter one finds with the basis

$$A_0 = \tfrac{1}{2}(P^0 + K^0), \quad A_1 = \tfrac{1}{2}(P^0 - K^0), \quad A_2 = D, \tag{3.21}$$

the Casimir operator

$$A_0^2 - A_1^2 - A_2^2 = \tfrac{1}{2}(P^0K^0 + K^0P^0) - D^2. \tag{3.22}$$

But in our representation

$$\mathbf{J}^2 + D^2 = \tfrac{1}{2}(P^0K^0 + K^0P^0), \tag{3.23}$$

so the Casimirs of su(2) and of su(1,1) coincide.

As we noted above, $\mathbf{J}$ generate the left regular representation of SU(2) on itself, and the Peter–Weyl theorem gives us a complete and orthogonal set of functions of the angles:

$$X^l_{\lambda m}(\theta,\phi,\psi) = D^l_{\lambda m}(-\psi,-\theta,-\phi), \quad l = 0,\tfrac{1}{2},1,\ldots, \quad \lambda,m = -l,-l+1,\ldots,l. \tag{3.24}$$

Here $D^l_{\lambda m}(\alpha,\beta,\gamma)$ is the matrix element (λ,m) of the unitary l-representation of SU(2) for the group element with Euler angles (α,β,γ), and $(-\psi,-\theta,-\phi)$ corresponds to the inverse of the group element given by (ϕ,θ,ψ). We use phase conventions as, e.g., in Ref. 18, thus

$$X^l_{\lambda m}(\theta,\phi,\psi) = \sqrt{(l+\lambda)!(l-\lambda)!(l+m)!(l-m)!}\,e^{i(\lambda\psi + m\phi)} \times \sum_t \frac{(-1)^{t-\lambda+m}}{(l+\lambda-t)!(l-m-t)!t!(t-\lambda+m)!} \times (\cos\tfrac{1}{2}\theta)^{2l+\lambda-m-2t}(\sin\tfrac{1}{2}\theta)^{2t-\lambda+m}. \tag{3.25}$$

Then, from the representation condition of the $D^l_{\lambda m}$ we get

$$\begin{aligned} \mathbf{J}^2X^l_{\lambda m} &= l(l+1)X^l_{\lambda m}, \\ J_3X^l_{\lambda m} &= mX^l_{\lambda m}, \\ J_\pm X^l_{\lambda m} &= \sqrt{l(l+1) - m(m\pm 1)}\,X^l_{\lambda m\pm 1}, \end{aligned} \tag{3.26}$$

and

$$\begin{aligned} I_3X^l_{\lambda m} &= -\lambda X^l_{\lambda m}, \\ I_\pm X^l_{\lambda m} &= -\sqrt{l(l+1) - \lambda(\lambda \mp 1)}X^l_{\lambda\mp 1 m}. \end{aligned} \tag{3.27}$$

So the index λ introduced in $X^l_{\lambda m}$ is indeed the helicity [cf. the discussion after (3.15)], and an arbitrary function in the space of the $(0,\lambda)$ representation can be written

$$f(\eta,\theta,\phi,\psi) = \sum_{l=|\lambda|,|\lambda|+1,\ldots}\ \sum_{m=-l}^{l} f_{lm}(\eta)X^l_{\lambda m}(\theta,\phi,\psi). \tag{3.28}$$

The equivariance condition is thus taken care of by the $X^l_{\lambda m}$, and the $f_{lm}(\eta)$ are ordinary functions. The su(1,1) generators act on the f_{lm} only, and for each l,m we have an irreducible representation. Since the value of the Casimir operator is $l(l+1)$, and the "compact" generator A_0 is evidently positive [see (3.21), (3.6), and (3.7)], this irreducible representation is the representation D^+_l from the discrete principal series (see, e.g., Ref. 19). As A_0 is essentially the four-dimensional isotropic harmonic oscillator Hamiltonian, a suitable basis for our representation could then be the set of eigenfunctions of this operator, with fixed λ and all l,m.

Below we give the expressions for the action on $f(p) \times X^l_{\lambda m}$ (with $p=\eta^2$) of the so(4,2) generators. Using the "I-vector" properties noted above, it is enough to find the action of P_3, N_3, K_3 on, e.g., X^l_{ll} and use the Wigner–Eckarts theorem both in $\mathbf{J}$ and in $\mathbf{I}$, and self-adjointness.

The results are given for the spherical components

$$P_{-1} = (P_1 - iP_2)/\sqrt{2}, \quad P_0 = P_3, \quad P_{+1} = -(P_1 + iP_2)/\sqrt{2}, \tag{3.29}$$

and similarly for $\mathbf{N}$ and $\mathbf{K}$:

$$\begin{aligned} P_\sigma f(p)X^l_{\lambda m} &= \langle l+1,m+\sigma|1\sigma,lm\rangle\,\langle l+1,\lambda\,|10,l\lambda\,\rangle pf(p)X^{l+1}_{\lambda,m+\sigma} \\ &\quad + \langle l,m+\sigma|1\sigma,lm\rangle\,\langle l\lambda\,|10,l\lambda\,\rangle pf(p)X^l_{\lambda,m+\sigma} \\ &\quad + \langle l-1,m+\sigma|1\sigma,lm\rangle\,\langle l-1,\lambda\,|10,l\lambda\,\rangle \\ &\quad \times pf(p)X^{l-1}_{\lambda,m+\sigma}, \end{aligned} \tag{3.30}$$

$$\begin{aligned} N_\sigma f(p)X^l_{\lambda m} &= -i\langle l+1,m+\sigma|1\sigma,lm\rangle\,\langle l+1,\lambda\,|10,l\lambda\,\rangle \\ &\quad \times p^{l+1}\frac{d}{dp}p^{-l}f(p)X^{l+1}_{\lambda,m+\sigma} \\ &\quad - i\langle l,m+\sigma|1\sigma,lm\rangle\,\langle l\lambda\,|10,l\lambda\,\rangle \\ &\quad \times \frac{d}{dp}pf(p)X^l_{\lambda,m+\sigma} \\ &\quad - i\langle l-1,m+\sigma|1\sigma,lm\rangle\,\langle l-1,\lambda\,|10,l\lambda\,\rangle \\ &\quad \times p^{-l}\frac{d}{dp}p^{l+1}f(p)X^{l-1}_{\lambda,m+\sigma}, \end{aligned} \tag{3.31}$$

$$\begin{aligned} K_\sigma f(p)X^l_{\lambda m} &= \langle l+1,m+\sigma|1\sigma,lm\rangle\,\langle l+1,\lambda\,|10,l\lambda\,\rangle \\ &\quad \times p^{l+1}\frac{d^2}{dp^2}p^{-l}f(p)X^{l+1}_{\lambda,m+\sigma} \\ &\quad + \langle l,m+\sigma|1\sigma,lm\rangle\,\langle l\lambda\,|10,l\lambda\,\rangle \\ &\quad \times p^{l}\frac{d}{dp}p^{-2l}\frac{d}{dp}p^{l+1}f(p)X^l_{\lambda,m+\sigma} \\ &\quad + \langle l-1,m+\sigma|1\sigma,lm\rangle\,\langle l-1,\lambda\,|10,l\lambda\,\rangle \\ &\quad \times p^{-l}\frac{d^2}{dp^2}p^{l+1}f(p)X^{l-1}_{\lambda,m+\sigma}, \end{aligned} \tag{3.32}$$

$$\sigma = 0, \pm 1,$$

and for the generators of our su(1,1) subalgebra

$$\begin{aligned} P^0f(p)X^l_{\lambda m} &= pf(p)X^l_{\lambda m}, \\ Df(p)X^l_{\lambda m} &= i\frac{d}{dp}pf(p)X^l_{\lambda m}, \\ K^0f(p)X^l_{\lambda m} &= -p^l\frac{d}{dp}p^{-2l}\frac{d}{dp}p^{l+1}f(p)X^l_{\lambda m}. \end{aligned} \tag{3.33}$$

E. Relations with magnetic monopoles

As has been pointed out earlier, the above description of the $(0,\lambda)$ representation parallels that of a charged particle in the field of a magnetic monopole, given, e.g., in Refs. 10 and 11. (The analogy, at least on a formal level, between this magnetic monopole problem and massless helicity representations has been noted by other authors.[20]) In the monopole case, the base space is the three-dimensional configuration space instead of the light cone, and the equivariance condition is the transformation rule for the wave function under gauge transformations. The value of λ for the monopole case is proportional to the product of the electric charge of the particle and the magnetic charge of the monopole, and the fact that 2λ must be an integer is the famous Dirac charge quantization condition.

Our choice of local sections giving charts a and b corresponds to the gauges a and b of Ref. 11 and the relation between our functions $X^l_{\lambda m}$ and the "monopole harmonics" in Ref. 11 is

$$Y_{q,lm_a}(\theta,\phi) = \sqrt{(2l+1)/4\pi}\, X^l_{-q,m_a}(\theta,\phi), \tag{3.34}$$

i.e., $Y_{q,lm}$ is $X^l_{-q,m}$ normalized on S^2.

For q integer, the operators $I_\pm$ give a convenient way to find $Y_{q,lm}$ from the ordinary spherical harmonics $Y_{lm} = Y_{0,lm}$.

IV. DIFFERENTIABLE VECTORS AND NUCLEARITY

In the following we collect first some standard results on differentiable vectors of group representations, nuclearity, etc. For details we refer, e.g., to Ref. 3 and references there.

The set of differentiable vectors of a unitary representation $U(G)$ of a Lie group G in a Hilbert space $\mathscr{H}$ is the (dense, linear) set $\mathscr{D}$ of all $x \in \mathscr{H}$ such that the map $G \to \mathscr{H}$ defined by $g \to U(g)x$ is a C^∞ function on G (differentiability in the $\mathscr{H}$ norm sense). Equivalently $\mathscr{D}$ can be defined as the largest common invariant domain of a set of generators $\{J_1, \dots, J_n\}$ of $U(G)$. This can also be expressed as follows: form the Nelson operator

$$A = J_1^2 + \dots + J_n^2, \tag{4.1}$$

associated to the choice $\{J_1, \dots, J_n\}$ of generators. Here A can be shown to be essentially self-adjoint (i.e., its self-adjoint extension is uniquely fixed) on a suitable domain of $U(G)$, e.g., the Gårding domain (which can actually be shown to coincide with $\mathscr{D}$, see Ref. 21). If we denote the self-adjoint closure of A by $\overline{A}$, $\mathscr{D}$ can be defined as the intersection of the domains of all powers of $\overline{A}$:

$$\mathscr{D} = \bigcap_n \mathscr{D}(\overline{A}^n). \tag{4.2}$$

We can give $\mathscr{D}$ a natural Fréchet space topology by the countable set of scalar products $(x,y)_n = (x,(\overline{A}+1)^{2n}y)$. We can then form the triplet

$$\mathscr{D} \subset \mathscr{H} \subset \mathscr{D}', \tag{4.3}$$

where $\mathscr{D}'$ can be identified (via an antilinear isomorphism) with the space of linear functionals on $\mathscr{D}$. Here $\mathscr{D}$ (and hence $\mathscr{H}$) is dense in $\mathscr{D}'$ in the topology of $\mathscr{D}'$ as a strong dual of $\mathscr{D}$. Here $\mathscr{D}$ is invariant not only under the action of all generators (and hence of the elements in their enveloping algebra), but also under $U(G)$; all these operators are also continuous in the topology of $\mathscr{D}$. By duality they can then be extended to continuous operators in $\mathscr{D}'$. We then have in $\mathscr{D}'$ a representation $U'(G)$ of G with everywhere defined and continuous generators $J_1', \dots, J_n'$.

The advantage with working in the extended domain $\mathscr{D}'$ instead of the restricted domain $\mathscr{D}$ is that under certain conditions we are assured to find in $\mathscr{D}'$ a complete set of (generalized) eigenvectors of the generators $J_1, \dots, J_n$ and other operators in the enveloping algebra of $U(G)$.

According to the nuclear spectral theorem this last assertion holds if $\mathscr{D}$, with the Fréchet topology defined above, is a nuclear space. The condition for $\mathscr{D}$ to be nuclear can be expressed in terms of the spectral properties of the Nelson operator $\overline{A}$ (which is evidently nonnegative).

It states that $\mathscr{D}$ is nuclear if and only if $\overline{A}$ has a purely discrete spectrum $\{\lambda_1, \lambda_2, \dots\}$ such that for some positive M the sum

$$\sum_i \frac{1}{(1+\lambda_i)^M} < \infty. \tag{4.4}$$

This condition is equivalent to requiring that the operator $(I+\overline{A})^{-M/2}$ should be of the Hilbert–Schmidt type.

Using this criterion for nuclearity it is easy to see that the set of differentiable vectors of a $(0,\lambda)$ representation of $\overline{\mathscr{P}}$ cannot be nuclear[3]: Suppose the Nelson operator $A = \mathbf{J}^2 + \mathbf{N}^2 + \mathbf{P}^2 + P^{02}$ had a smallest (obviously positive) eigenvalue λ_1, with corresponding normalized eigenfunction Φ. Forming the (still normalized) eigenfunction

$$\Phi_\kappa = e^{-i\kappa D}\Phi,$$

where D is as in (3.7) or (3.20), we easily deduce from the commutation relations (3.9) that

$$(\Phi_\kappa, \overline{A}\Phi_\kappa) = (\Phi, \overline{A}\Phi) + (e^{-2\kappa} - 1)\,(\Phi, (\mathbf{P}^2 + P^{02})\Phi).$$

Since the scalar product in the last term is positive, we obtain for $\kappa > 0$ the impossible result $(\Phi_\kappa, \overline{A}\Phi_\kappa) < \lambda_1$.

Now we shall show that the set of differentiable vectors of the representation of the covering group of SO(4,2) obtained by extension from our $(0,\lambda)$ representation of $\overline{\mathscr{P}}$ is a nuclear space. Using this space as our $\mathscr{D}$, we get what is called a Gelfand triplet

$$\mathscr{D} \subset \mathscr{H} \subset \mathscr{D}', \tag{4.5}$$

where $\mathscr{D}$ is a dense subspace of $\mathscr{H}$ with a nuclear Fréchet topology. Any self-adjoint operator in $\mathscr{H}$, which is essentially self-adjoint on $\mathscr{D}$ and is continuous as an operator in $\mathscr{D}$, then has a complete set of generalized eigenvectors in $\mathscr{D}'$, the dual of $\mathscr{D}$.

For the following it is convenient to observe that a covering of SO(4,2) is actually the group SU(2,2). Since this group [as well as SO(4,2), of course!] is semisimple with a finite center, it follows from a general theorem[3] that the set of differentiable vectors, with the natural topology, is a nuclear space. We will show this directly in our case, at the same time obtaining a concrete characterization of the set of differentiable vectors of SU(2,2).

For the proof we study the reducible representations of $\overline{\mathscr{P}}$, SU(2,2), and Sp(8,$\mathbb{R}$) in $L^2(\mathbb{R}^4)$ obtained by dropping the equivariance condition, i.e., allowing all possible half-integer values of λ. As follows from Sec. III, we have in this space, besides a unitary representation (UR, for short) of $\overline{\mathscr{P}}$, also UR's of the chain

$$\mathrm{Sp}(8,\mathbb{R}) \supset \mathrm{SU}(2,2) \supset \mathrm{SU}(1,1) \supset \mathrm{U}(1), \tag{4.6}$$

where U(1) is the group generated by $A_0 = \frac{1}{2}(P^0 + K^0)$. Denoting the corresponding sets of differentiable vectors—with their natural topologies—by $\mathscr{D}(G)$, we shall prove the following chain inclusion (set-theoretical and topological)

$$\mathscr{S}(\mathbb{R}^4) = \mathscr{D}(\mathrm{U}(1)) \supset \mathscr{D}(\mathrm{SU}(1,1)) \supset \mathscr{D}(\mathrm{SU}(2,2)) \supset \mathscr{D}(\mathrm{Sp}(8,\mathbb{R})) = \mathscr{S}(\mathbb{R}^4), \tag{4.7}$$

which implies equality everywhere, and in particular

$$\mathscr{D}(\mathrm{SU}(2,2)) = \mathscr{S}(\mathbb{R}^4). \tag{4.8}$$

Here $\mathscr{S}(\mathbb{R}^4)$ is the set of C^∞ functions on $\mathbb{R}^4$, rapidly decreasing (with all derivatives) at infinity, with the standard topology. This space is a nuclear Fréchet space.

Proof of (4.7): The only nontrivial parts of the chain are the equalities at the ends. These equalities both follow from the fact that the Nelson operators of the groups Sp(8,$\mathbb{R}$) and U(1) are both essentially the square of the (rescaled) harmonic oscillator Hamiltonian in four dimensions [cf. (3.12), and (3.21) combined with (3.6) and (3.7)]. It is well-known that the domain of all powers of this oscillator Hamiltonian is just $\mathscr{S}(\mathbb{R}^4)$. In fact, $\mathscr{S}(\mathbb{R}^4)$ can be topologized, in the way indicated under formula (4.2), by the powers of the operator $-\Delta + r^2$.

The spectrum of the generator A_0 of U(1) in $L^2(\mathbb{R}^4)$ can in fact be obtained explicitly from the discussion following formula (3.28). The result is as follows: For every $j = 0, \frac{1}{2}, 1, \ldots$, there is a sequence of eigenvalues $j, j+1, j+2, \ldots$, each with degeneracy $(2j+1)^2$. The nuclearity of the space $\mathscr{D}(\mathrm{U}(1))$ then follows from an application of the criterion (4.4).

Concluding: the set of differentiable vectors of the representation of SU(2,2) extending the $(0,\lambda)$ representation of $\overline{\mathscr{P}}$ is just the set of equivariant functions of the nuclear space $\mathscr{S}(\mathbb{R}^4)$. This set, which we shall denote $\mathscr{S}_\lambda(\mathbb{R}^4)$ or $\mathscr{S}_\lambda$, is also nuclear, as a (closed) linear subset of a nuclear space.

V. GENERALIZED MOMENTUM EIGENSTATES

Using the results of the last section, for the $(0,\lambda)$ representation of $\overline{\mathscr{P}}$ we use the Gelfand triplet $\mathscr{D} \subset \mathscr{H} \subset \mathscr{D}'$, with $\mathscr{D} = \mathscr{S}_\lambda(\mathbb{R}^4)$. The space of generalized state vectors $\mathscr{D}'$ is then $\mathscr{S}_\lambda(\mathbb{R}^4)'$.

Since any continuous linear functional on $\mathscr{S}_\lambda$ can be extended to all of $\mathscr{S}$, we can consider $\mathscr{D}'$ as a space of equivalence classes of $\mathscr{S}'$, two elements of $\mathscr{S}'$ being equivalent if their restrictions to $\mathscr{S}_\lambda$ coincide. Alternatively, one could fix the extension and identify $\mathscr{D}'$ with the set of elements of $\mathscr{S}'$ which are zero on $\mathscr{S}_{\lambda'}$ for all $\lambda' \neq \lambda$, i.e., simply take $\mathscr{D}'$ as the space of λ-equivariant elements of $\mathscr{S}'$, denoted $\mathscr{S}'_\lambda$. The latter description may be more natural, but we shall also find the former convenient when we now turn to a discussion of the eigenstates of energy-momentum.

The spectrum of P^μ is the closed forward light cone. The corresponding generalized eigenvectors are fundamentally different for nonzero eigenvalues on the one hand, and for eigenvalues at the origin on the other.

For any nonzero p_0^μ on the cone, it is not hard to see that the most general solution in $\mathscr{S}'(\mathbb{R}^4)$ of the equation $P^\mu F = p_0^\mu F$ is given by a distribution F_1 on S^1 by the expression $(F,f) = (F_1, f_0)$, where $f_0(\psi) = f(\sqrt{p_0}, \theta_0, \phi_0, \psi)$. Here (p_0, θ_0, ϕ_0) are the spherical coordinates of p_0. Due to the equivariance condition, for each nonzero momentum the corresponding space of sharp momentum states in $\mathscr{D}'$ is thus one-dimensional. Such a state vector can be represented by a multiple of the δ-function in some point on the fiber over p_0^μ. For $\lambda \neq 0$, δ-functions in different points on the fiber will correspond to different phases for the eigenvector. This means that the space of sharp momentum eigenvectors for nonzero momentum have a nontrivial line bundle structure over the forward light cone.

In $\mathscr{S}'_\lambda$ we can expand the eigenvectors as

$$\delta(\eta - \eta_0) \approx |\eta_0|^{-3}\delta(|\eta| - |\eta_0|) \times \sum_{l = |\lambda|, |\lambda|+1, \ldots} \frac{2\pi^2}{2l+1} \times \sum_{m=-l}^{l} \overline{X^l_{\lambda m}(\theta_0,\phi_0,\psi_0)}\, X^l_{\lambda m}(\theta,\phi,\psi). \tag{5.1}$$

Generalized eigenvectors of zero energy-momentum can be represented by elements of $\mathscr{S}'$ with support at the origin, i.e., by finite sums of derivatives of $\delta(\eta)$. To find all these eigenvectors we introduce the following derivation operators:

$$\begin{aligned} \pi_1 &= \tfrac{1}{2}(i\,\partial_4 - \partial_3), \quad \bar{\pi}_1 = \pi_1^* = \tfrac{1}{2}(i\,\partial_4 + \partial_3), \\ \pi_2 &= \tfrac{1}{2}(i\,\partial_2 - \partial_1), \quad \bar{\pi}_2 = \pi_2^* = \tfrac{1}{2}(i\partial_2 + \partial_1). \end{aligned} \tag{5.2}$$

Then the π's all commute with each other and

$$\begin{aligned} &[\pi_j, u_k] = i\delta_{jk}, \quad [\bar{\pi}_j, \bar{u}_k] = i\delta_{jk}, \\ &[\pi_j, \bar{u}_k] = [\bar{\pi}_j, u_k] = 0. \end{aligned} \tag{5.3}$$

A general λ-equivariant element F of $\mathscr{S}'$ with support at the origin can then be expressed as a unique finite linear combination of terms of the form

$$\pi_1^{n_1}\pi_2^{n_2}\bar{\pi}_1^{\bar{n}_1}\bar{\pi}_2^{\bar{n}_2}\delta, \tag{5.4}$$

with $n_i, \bar{n}_i = 0,1,\ldots$, and $n_1 + n_2 - \bar{n}_1 - \bar{n}_2 = 2\lambda$. Here $P^\mu F = 0$ is equivalent to $u_i\bar{u}_j F = 0$, all $i,j \in \{0,1\}$. But $u_i\bar{u}_j\pi_1^{n_1}\pi_2^{n_2}\bar{\pi}_1^{\bar{n}_1}\bar{\pi}_2^{\bar{n}_2}\delta = n_i\bar{n}_j\pi_i^{n_i-1}\bar{\pi}_j^{\bar{n}_j-1}\cdots\delta$, and by linear independence we must have $n_i\bar{n}_j = 0$ for all terms in F. This means that for $\lambda \geqslant 0$ the eigenspace corresponding to $p^\mu = 0$ is spanned by the linearly independent vectors

$$\pi_1^n\pi_2^{2\lambda-n}\delta, \quad n = 0,1,\ldots,2\lambda, \tag{5.5}$$

and for $\lambda \leqslant 0$ by

$$\bar{\pi}_1^n\bar{\pi}_2^{-2\lambda-n}\delta, \quad n = 0,1,\ldots,-2\lambda. \tag{5.6}$$

The space of generalized eigenvectors for zero eigenvalue is thus $(2|\lambda|+1)$-dimensional. In the case of the electromagnetic field, for instance, we have two representations, with $\lambda = \pm 1$, and the total dimension of the zero momentum eigenspace is 6. These six linearly independent generalized states correspond to the three components each, of constant electric and magnetic fields.

VI. SUMMARY

In this study of sharp momentum states for a helicity representation of the Poincaré group we have in some detail analyzed two problems which do not appear for the massive representations, and given their solutions.

Since the space of differentiable vectors of the representation is not nuclear, we have used the differentiable vectors of the extension of the representation to the conformal group.

The string-like singularities in the generators, i.e., the nontriviality of the line-bundles involved, were taken care of by utilizing equivariant functions on $\dot{\mathbb{C}}^2$ (or $\dot{\mathbb{R}}^4$) rather than functions on the forward light cone in p-space.

Let us summarize some of the advantages with working in $\dot{\mathbb{C}}^2$, rather than on the cone.

First, the representation of the Poincaré group takes a very simple form, both globally (2.13) and locally (3.6), and the extension to the conformal algebra (3.7) is easy to find. It is interesting that the different helicity representations all look formally the same, they only act on different subspaces of $L^2(\mathbb{C}^2)$. Taking them all together, the extension to $\mathrm{Sp}(8,\mathbb{R})$ used in Sec. III is very natural.

Second, the space of differentiable vectors of the conformal group is just a subspace of the well-known $\mathscr{S}(\mathbb{R}^4)$. This greatly simplified the analysis of the sharp momentum states in Sec. V.

Third, although we do not give the details in this paper, the decomposition of the $(0,\lambda)$ representation in irreducible representations of the Lorentz group is very simple to obtain. The result is a direct integral decomposition over the continuous index in the principal series of unitary irreducible representations of $\mathrm{SL}(2,\mathbb{C})$ (the covering group of the proper Lorentz group), whereas the discrete index is simply related to the helicity λ.

Fourth, functions on $\dot{\mathbb{C}}^2$ also give the natural representation space for the continuous spin representations.[22] We thus find a unified description of all the massless representations of the Poincaré group.

In our opinion, these advantages greatly outweigh the advantage of an intuitively more clear picture with functions on the light cone; a picture which is anyhow often misleading, as has been discussed in Ref. 23.

In Sec. V we pointed out that the sharp momentum states are, in their dependence on **p**, sections in a nontrivial bundle instead of ordinary functions. That this is a possible objection to the treatment in Ref. 1 was what first stimulated our interest in the problem. This and other questions related to the results of Ref. 1 are discussed in Ref. 23.

Finally, we shall indicate a rather direct generalization of the treatment in this paper, namely to a semidirect product $\mathbb{R}^{n^2}\times\mathrm{SL}(n,\mathbb{C})$, where $\mathrm{SL}(n,\mathbb{C})$ is taken to act on the space $\mathbb{R}^{n^2}$ of Hermitian $n\times n$ matrices in a way analogous to that given in (2.2). If we assume the rank of these Hermitian matrices to be 1, i.e., only one eigenvalue different from zero, and use a one-dimensional inducing representation for the corresponding isotropy group, it can be shown that the representation of the original group can be realized on a space of equivariant functions on $\dot{\mathbb{C}}^n$. The space of differentiable vectors will not be nuclear in this case. It seems reasonable to assume that one can use the differentiable vectors of an extension of the given representation to a (degenerate) representation of a semisimple group containing the original group. This space should consist of the equivariant functions in $\mathscr{S}(\mathbb{R}^{2n})$.

ACKNOWLEDGMENTS

We thank our friends Professor M. Flato and Professor D. Sternheimer for drawing our attention to the problem studied in this paper, and for many lively and informative discussions.

[1] S. Weinberg and E. Witten, Phys. Lett. B **96**, 59 (1980).
[2] I. M. Gel'fand and N. Y. Vilenkin, *Generalized Functions* (Academic, New York, 1964), Vol. 4.
[3] B. Nagel in *Studies in Mathematical Physics* (Reidel, Dordrecht, 1971).
[4] D. Sternheimer, J. Math. Pures Appl. **47**, 289 (1968).
[5] G. Mack and I. Todorov, J. Math. Phys. **10**, 2078 (1969).
[6] E. Angelopoulos and M. Flato, Lett. Math. Phys. **2**, 405 (1978).
[7] J. S. Lomont and H. E. Moses, J. Math. Phys. **3**, 405 (1962).
[8] J. C. Guillot and J. L. Petit, Helv. Phys. Acta **39**, 281 and 300 (1966).
[9] S. K. Bose and R. Parker, J. Math. Phys. **10**, 812 (1969).
[10] W. Greub and H-L. Petry, J. Math. Phys. **16**, 1347 (1975).
[11] T. T. Wu and C. N. Yang, Nucl. Phys. B **107**, 365 (1976).
[12] A. Kirillov, *Eléments de la théorie des représentations* (Mir, Moscow, 1974) [English translation, *Elements of the Theory of Representations* (Springer-Verlag, Berlin, 1976)].
[13] E. Wigner, Ann. Math. **40**, 149 (1939).
[14] S. Helgason, *Differential Geometry, Lie Groups and Symmetric Spaces* (Academic, New York, 1978), p. 270.
[15] C. Itzykson, Comm. Math. Phys. **4**, 92 (1967).
[16] V. Bargmann, Indiana Symposium Lectures, 1968.
[17] E. Nelson, Ann. Math. **70**, 572 (1959).
[18] A. Messiah, *Quantum Mechanics* (North Holland, Amsterdam, 1962), Vol. II, p. 1072.
[19] V. Bargmann, Ann. Math. **48**, 568 (1947).
[20] H. Bacry, J. Phys. A **14**, L73 (1981); A. O. Barut, J. Phys. A **14**, L267 (1981), and references there to earlier work by Barut and co-workers; A. P. Balachandran, G. Marmo, B-S. Skagerstam, and A. Stern, "Gauge Symmetries and Fibre Bundles," in *Lecture Notes in Physics* 188 (Springer-Verlag, Berlin, 1983).
[21] J. Dixmier and P. Malliavin, Bull. Sci. Math. **102**, 307 (1978).
[22] B. Ek, Lett. Math. Phys. **7**, 339 (1983).
[23] M. Flato, C. Fronsdal, and D. Sternheimer, Commun. Math. Phys. **90**, 563 (1983).

Gradient property of bifurcation equation for systems with rotational symmetry

Giuseppe Gaeta and Pietro Rossi
Department of Physics, New York University, New York, New York 10003

(Received 20 December 1983; accepted for publication 20 January 1984)

We show that any polynomial equation covariant with respect to any representation of SO(2) is a gradient equation. The same holds for the fundamental representation of SO(3). For the other representations of SO(3) we give a simple necessary and sufficient condition an equation has to satisfy in order to be a gradient one. As a side result, we obtain a formula for the decomposition of the symmetrized power of an irreducible representation of SO(3) in a sum of irreducible representations. We apply our results to symmetric bifurcation theory.

PACS numbers: 02.20. + b

We want to investigate the gradient property of the bifurcation equation for systems with symmetry, in this case with rotational symmetry, i.e., we would like to make some precise statement about the possibility of expressing the bifurcation equation for such systems as a gradient equation. Since a well-known theorem[1] ensures that the bifurcation equation has all the symmetry of the original equation, what we are really investigating is simply the gradient property for covariant operators homogeneous in the basis function of a representation; i.e., for monomials of these functions which still transform according to the same representation.

Given a representation Γ of a group $\mathscr{G}$, which acts on the space V spanned by the vectors $\{\mathbf{v}_1,...,\mathbf{v}_n\}$, we call vectors (scalars) of order K the objects which follow the representation Γ (the identity representation) in the symmetrized product $(\Gamma\otimes\Gamma\otimes\cdots\otimes\Gamma)_s=(\Gamma^{\otimes K})_s$. So what we want to know is under which conditions any vector of order K can be expressed as the gradient of a scalar of order $(K+1)$. Here we deal with the cases $\mathscr{G}=\mathrm{SO}(2)$ and $\mathscr{G}=\mathrm{SO}(3)$.

Following Sattinger[1] we start with the formula

$$\sum_n z^n\chi^n(g)=\det[1-zT(g)]^{-1}, \tag{1}$$

where $\chi^n(g)$ is the character of the symmetrized n-tensor product $[T(g)^{\otimes n}]_s$ and z is a complex variable $|z|<1$. Multiplying both sides by $\overline{\chi^{(\mu)}(g)}$ and integrating with the invariant Haar measure we have

$$\int d\nu(g)\det[1-zT(g)]^{-1}\,\overline{\chi^{(\mu)}(g)}=\sum_n z^n\int\chi^n(g)\,\overline{\chi^{(\mu)}(g)}\,d\nu(g),$$

but

$$\int\chi^n(g)\,\overline{\chi^{(\mu)}(g)}\,d\nu(g)=c_n^\mu$$

is the number of times the representation T^μ is contained in $(T^{\mu\otimes n})_s$. We note that, since $\det(A)=\det(UAU^+)$ and $\mathrm{tr}(A)=\mathrm{tr}(UAU^+)$, where U is a unitary matrix, the function to be integrated is a class function, and we can consider, instead of the representation $T^{(\mu)}(g)$, any other unitarily equivalent representation, so that we have

$$I^{(\mu)}=\sum_{k=0}^\infty c_k z^k=\int\det[1-zT(l)]^{-1}\,\overline{\chi^{(\mu)}(l)}\,d\nu(l), \tag{2}$$

where the integration is over the set of conjugate classes, and $d\nu(l)$ is the corresponding class measure.

Now we came to the case $\mathscr{G}=\mathrm{SO}(2)$. First of all we treat the fundamental representation, which is given by

$$T^{(1)}(\theta)=\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix},$$

or, after a unitary transformation,

$$T^{(1)}(\theta)=\begin{pmatrix}e^{-i\theta} & 0\\ 0 & e^{i\theta}\end{pmatrix}.$$

Now formula (2) yields

$$I^{(\mu)}=\frac{1}{2\pi}\int_0^{2\pi}[(1-ze^{-i\theta})(1-ze^{i\theta})]^{-1}\,\overline{\chi^{(\mu)}(\theta)}\,d\theta.$$

Namely, considering vectors and scalars (we always indicate with $T^{(0)}$ the identity representation)

$$I^{(0)}=\frac{1}{2\pi}\int_0^{2\pi}[(1-ze^{-i\theta})(1-ze^{i\theta})]^{-1}\,d\theta, \tag{3a}$$

$$I^{(1)}=\frac{1}{2\pi}\int_0^{2\pi}[(1-ze^{-i\theta})(1-ze^{i\theta})]^{-1}(e^{i\theta}+e^{-i\theta})\,d\theta. \tag{3b}$$

These integrals are, of course, transformed into integrals over the unit circle of the complex plane through the change of variable $\omega=e^{i\theta}$, that is,

$$I^{(0)}=\frac{1}{2\pi i}\int[(1-z\omega)(\omega-z)]^{-1}\,d\omega, \tag{4a}$$

$$I^{(1)}=\frac{1}{2\pi i}\int[\omega(1-z\omega)(\omega-z)]^{-1}(\omega^2+1)d\omega. \tag{4b}$$

Evaluation of these with the residue formula gives

$$I^{(0)}=\sum_{n=0}^\infty z^{2n}, \tag{5a}$$

$$I^{(1)}=\sum_{n=0}^\infty z^{(2n+1)}, \tag{5b}$$

so that we have only one vector at each odd order, and one scalar at even order. Since the gradient of a scalar is obviously a vector, we have that gradient property holds for the fundamental representation of SO(2).

For what concerns the general case, namely any representation $T^{(m)}(g)$,

 0022-2488/84/061671-03$02.50

$$T^{(m)}(\theta) = T^{(1)}(m\theta) = \begin{pmatrix} \cos(m\theta) & \sin(m\theta) \\ -\sin(m\theta) & \cos(m\theta) \end{pmatrix},$$

performing an unitary transformation we have

$$T^{(m)}(\theta) = \begin{pmatrix} e^{-im\theta} & 0 \\ 0 & e^{im\theta} \end{pmatrix}.$$

Inserting this in formula (2), and performing the change of variable $\phi = m\theta$, we have, using the 2π periodicity of $e^{i\phi}$, formulas (3a) and (3b), with ϕ appearing for θ, so that what we did for the fundamental representation holds for any representation of SO(2).

We have therefore proved, for any representation of SO(2), that any vector of any order is the gradient of a scalar.

This implies, of course, that for any system with SO(2) symmetry, the bifurcation equation is a gradient equation.

Now we pass to the case $\mathscr{G} = \mathrm{SO}(3)$. The representations of SO(3) will be indicated by D^l, their dimension being $(2l+1)$, and D^0 being the identity representation.

We consider at first the fundamental representation, i.e., the three-dimensional representation D^1. Formula (2) gives now

$$I^{(0)} = \frac{1}{2\pi}\int_0^{2\pi}(1-\cos\theta) \times [(1-ze^{-i\theta})(1-z)(1-ze^{i\theta})]^{-1}\,d\theta, \qquad (6a)$$

$$I^{(1)} = \frac{1}{2\pi}\int_0^{2\pi}(1-\cos\theta)[(1-ze^{-i\theta})(1-z)(1-ze^{i\theta})]^{-1} \times (e^{-i\theta}+1+e^{i\theta})\,d\theta, \qquad (6b)$$

after transformation to complex integrals on the unit circle we get

$$I^{(0)} = -\frac{1}{4\pi i}\int(1-\omega)^2[\omega(\omega-z)(1-z)(1-z\omega)]^{-1}\,d\omega, \qquad (7a)$$

$$I^{(1)} = \frac{-1}{4\pi i}\int(1-\omega)^2(1+\omega+\omega^2) \times [\omega^2(\omega-z)(1-z)(1-z\omega)]^{-1}\,d\omega. \qquad (7b)$$

Evaluation through the residue formula yields for these

$$I^{(0)} = \sum_{n=0}^{\infty} z^{2n}, \qquad (8a)$$

$$I^{(1)} = \sum_{n=0}^{\infty} z^{(2n+1)}, \qquad (8b)$$

and the same considerations apply as in the SO(2) case treated above, so that for the fundamental representation of SO(3) any vector of any order is the gradient of a scalar.

Now we note that, in order to prove gradient property at order N (namely for vectors of order N), it suffices in general to prove that the number of (completely symmetric independent) vectors of order N is equal to the number of (completely symmetric independent) scalars of order $N+1$. In fact, the gradient of a scalar is surely a vector, and gradients of independent scalars are themselves independent; and all the vectors of order N are, therefore, in this case, gradients of scalars of order $N+1$. Consider now the product $[(\Gamma^{\mu})^N]_S$. The multiplicity of vectors and scalars is given, as said before, by the coefficients $c_N^{(\mu)}$ and $c_N^{(0)}$ in

$$I^{(\mu)} = \int \det[1 - z\Gamma(g)]^{-1}\,\overline{\chi^{(\mu)}(g)}\,d\nu(g) = \sum c_N^{\mu} z^N,$$

$$I^{(0)} = \int \det[1 - z\Gamma(g)]^{-1}\,d\nu(g) = \sum c_N^{(0)} z^N.$$

Therefore the gradient property holds for the representation Γ^{μ} at order N if and only if

$$c_N^{(\mu)} = c_{N+1}^{(0)}. \qquad (9)$$

Now we wish to consider the general case for $\mathscr{G} = \mathrm{SO}(3)$, namely, in the language used just above, that in which Γ is any of the D^l's. The integrals $I^{(0)}$ and $I^{(l)}$ will be written

$$I^{(0)} = \frac{1}{2\pi}\int_0^{2\pi}(1-\cos\theta)\Big(\prod_{m=-l}^{l}(1-ze^{i\theta})\Big)^{-1}d\theta, \qquad (10a)$$

$$I^{(l)} = \frac{1}{2\pi}\int_0^{2\pi}(1-\cos\theta)\Big(\sum_{m=-l}^{l} e^{im\theta}\Big) \times\Big(\prod_{m=-l}^{l}(1-ze^{i\theta})\Big)^{-1}d\theta. \qquad (10b)$$

Now, using the fact that

$$(1-\cos\theta)\sum_{m=-l}^{l} e^{im\theta} = \cos(l\theta) - \cos[(l+1)\theta], \qquad (11)$$

and writing $S(\theta,z) = \Pi_{m=-l}^{l}(1-ze^{im\theta})^{-1}$ we write (10) as

$$I^{(0)} = \frac{1}{2\pi}\int_0^{2\pi}(1-\cos\theta)S(\theta,z)\,d\theta, \qquad (12a)$$

$$I^{(l)} = \frac{1}{2\pi}\int_0^{2\pi}\{\cos(l\theta) - \cos[(l+1)\theta]\}S(\theta,z)\,d\theta. \qquad (12b)$$

Using the formula for geometrical series we get

$$S(\theta,z) = \prod_{m=-l}^{l}(1-ze^{im\theta})^{-1} = \prod_{m=-l}^{l}\sum_{n_m=0}^{\infty}(ze^{im\theta})^{n_m}$$
$$= \sum_{n_{-l},\dots,n_l=0}^{\infty} z^a e^b$$
$$= \sum_{N,S}\sigma(N,l,S)z^N e^{iS\theta},$$
$$a = \sum_k n_k, \quad b = \sum_k k n_k \theta, \quad k = -l,\dots,l,$$

and where $\sigma(N,l,S)$ is the number of possible ways of arranging $2l+1$ integer positive numbers n_k, $k = -l,\dots,l$ so that

$$\sum_{k=-l}^{l} n_k = N, \qquad \sum_{k=-l}^{l} k n_k = S.$$

Now we can rewrite (12) as

$$I^{(0)} = \frac{1}{2\pi}\sum_N z^N\int_0^{2\pi}(1-\cos\theta)\sum_S\sigma(N,l,S)e^{iS\theta}\,d\theta, \qquad (13a)$$

$$I^{(l)} = \frac{1}{2\pi}\sum_N z^N\int_0^{2\pi}\{\cos(l\theta) - \cos[(l+1)\theta]\} \times\sum_S\sigma(N,l,S)e^{iS\theta}\,d\theta, \qquad (13b)$$

which, using the orthogonality property of Fourier basis in $[0, 2\pi]$, gives finally

$$I^{(0)} = \sum_N [\sigma(N,l,0) - \sigma(N,l,1)] z^N = \sum_N c_N^{(0)} z^N, \tag{14a}$$

$$I^{(l)} = \sum_N [\sigma(N,l,l) - \sigma(N,l,l+1)] z^N = \sum_N c_N^{(l)} z^N, \tag{14b}$$

so that we can give a simple necessary and sufficient condition in order to have the gradient property holding for vectors of order N following D^l representation, namely it holds if and only if

$$\sigma(N+1,l,0) - \sigma(N+1,l,1) = \sigma(n,l,l) - \sigma(N,l,l+1). \tag{15}$$

By the way, during this computation we have proved that the multiplicity of D^s in the symmetrized product $\{(D^l)^N\}_S$ is given by

$$\sigma(N,l,S) - \sigma(N,l,S+1). \tag{16}$$

As a check for our result, we notice that, for $l = 1$, we have

$$\begin{aligned} &\sigma(N+1,l,0) = [(N+1)/2 + 1], \\ &\sigma(N+1,l,1) = [N/2 + 1], \\ &\sigma(N,l,l) = [(N-1)/2 + 1], \\ &\sigma(N,l,l+1) = [N/2], \end{aligned} \tag{17}$$

where the square brackets mean integer part. It is immediate to see that in this case Eq. (15) is satisfied at any order, so we have anew the result obtained before by direct computation, namely that the gradient property holds at any order for the fundamental representation of SO(3).

In order to show our result is not trivial, and it really gives a useful criterion and not one which is always (or never) satisfied, we consider here the case $l = 2$. Using the usual technique to evaluate the integrals $I^{(0)}$ and $I^{(2)}$ given by

$$I^{(0)} = \frac{1}{2\pi} \int_0^{2\pi} (1 - \cos\theta) \left(\prod_{m=-2}^{2} (1 - z e^{im\theta}) \right)^{-1} d\theta,$$

$$I^{(2)} = \frac{1}{2\pi} \int_0^{2\pi} (1 - \cos\theta) \left(\prod_{m=-2}^{2} (1 - z e^{im\theta}) \right)^{-1} \times \left(\sum_{m=-2}^{2} e^{im\theta} \right) d\theta,$$

we get

$$I^{(0)} = \sum_N c_N^{(0)} z^N = [(1-z^2)(1-z^3)]^{-1}, \tag{18a}$$

$$I^{(2)} = \sum_N c_N^{(2)} z^N = z[(1-z)(1-z^3)]^{-1}, \tag{18b}$$

so that at first orders we have the following situation:

N	1	2	3	4	5	6	7
$c_N^{(2)}$	1	1	1	2	2	2	3 .
$c_{N+1}^{(0)}$	1	1	1	1	2	1	2

Also, we have that at order 4 the gradient property is not true, but it is true again at order 5.

We recall that for $N \leqslant 3$, it has been shown[1,2] that the gradient property holds for any D^l. The table above shows it is not possible to extend those results in such a generality.

In the case $l = 2$ it is easy to convince oneself, from (18), that the gradient property does not hold for $N > 5$. It is natural to ask if it is possible to give a function of l, $F(l)$, such that gradient property for the representation D^l does not hold at order N for any $N > F(l)$.

As for what concerns the bifurcation equation, we have that (a) for any system with SO(3) symmetry, the reduced bifurcation equation up to order 3 is a gradient equation[1,2]; (b) if SO(3) acts on the space in which the bifurcation equation acts through the three-dimensional representation D^1, the full bifurcation equation is a gradient equation; and (c) if SO(3) acts on the space in which the bifurcation equation acts through another representation D^l, the gradient property holds for reduced bifurcation equations up to order M if for any order N at which we have nonvanishing terms, with $N \leqslant M$, condition (15) is satisfied.

In the case of systems with SO(2) symmetry, as we showed before, the bifurcation equation is always a gradient equation.

ACKNOWLEDGMENTS

One of the authors (G. G.) wishes to thank Professor G. F. Dell'Antonio and Professor D. H. Sattinger for useful discussions in an early stage of this work.

[1] D. H. Sattinger, *Group Theoretic Methods in Bifurcation Theory, Lecture Notes in Mathematics*, No. 762 (Springer, New York, 1979).
[2] G. Gaeta, J. Phys. A **16**, L607 (1983).

Construction of extremal vectors for Verma submodules of Verma modules

Miguel Lorente[a)]
Departamento de Metodos Matemáticos de la Física, Facultad de Ciencias Físicas, Universidad Complutense, Madrid -3, Spain
Bruno Gruber
Physics Department, Southern Illinois University, Carbondale, Illinois 62901

(Received 7 June 1983; accepted for publication 14 October 1983)

In this paper we discuss certain aspects of the theory of Verma submodules of Verma modules of simple Lie algebra. We describe a simple algorithm for the complete determination of the highest weights which are associated with the Verma submodules. Moreover, we give the proof for a method which permits an explicit determination of the extremal vectors which define the Verma submodules. For the case of the simple Lie algebra A_l the Verma modules are obtained in explicit form.

PACS numbers: 02.20. + b

1. INTRODUCTION

In this paper we discuss certain aspects of the theory of Verma modules of simple Lie algebras. By making use of the fundamental results of Verma's theory as given by Verma[1] and Bernshtein, Gel'fand, and Gel'fand[2] we obtain an algorithm for the determination of the weights which are associated with the extremal vectors which define the Verma submodules. This algorithm is discussed in detail for the case of the simple Lie algebras A_l.

In Ref. 3 it was observed that the extremal vectors which define the Verma submodules factorize. Based upon this fact we make use of Verma's theory to prove the factorization of the extremal vectors for the general case of a simple Lie algebra. It is shown that every extremal vector can be obtained in the form of a product of simple extremal vectors. That is, every extremal vector is given as a product of extremal vectors each of which belongs to a Verma submodule $V_M \subset V_\Lambda$ whose weight M is related to the weight Λ of another Verma submodule V_Λ by a single reflection S_α where α is a (positive) root of the algebra. If S_α is a reflection which corresponds to a nonsimple root then S_α can be broken up into a product of an odd number of reflections on simple roots. Thus the extremal vector will be the product of simple root vectors. If a simple root vector has negative power for some stage one can also define the extremal vector formally, by the product of simple root vectors with possible negative powers and then obtain its expression in terms of the basis vectors by simple commutation relations.

This then permits an explicit determination of the extremal vectors which define the Verma submodules of the Verma modules of the simple Lie algebras. The method employed is again an algorithm and thus a given extremal vector is constructed in stages from previously constructed extremal vectors. The extremal vectors which are obtained in this manner are in general not in standard form. Use has to be made of the Lie products to bring them into standard form and thus the extremal vector is in general obtained as a nontrivial linear combination over the basis elements of the weight subspace associated with it.

[a)] Work supported in part by a grant of the U.S.–Spanish Joint Committee for Scientific and Technological Cooperation.

As an additional result we determine in this paper the Verma modules for the case of the simple Lie algebras A_l in explicit form.

2. THE LIE ALGEBRA A_l AND ITS UNIVERSAL ENVELOPING ALGEBRA

Let A_l denote the complex simplex Lie algebra, which corresponds to the group SU(l + 1). In its Cartan canonical form a basis is given by the elements

$$h_i \equiv e_{ii}, \quad i = 1,2,\ldots,l+1, \quad \sum_{i=1}^{l+1} h_i = 0,$$

$$_i f_j \equiv e_{ij}, \quad i \neq j, \; i,j = 1,2,\ldots,l+1,$$

where e_{ij} represents the matrix with 1 in the position (ij) and zero elsewhere; the h_i are the diagonal elements of the Cartan subalgebra, the $_i f_j$ denote the root vector associated to the root $e_i - e_j$, and the e_i denote the Cartesian basis in the (l + 1)-dimensional root space. A suffix to the left of the root vector $_i f_j$ indicates that the i-component of the root is + 1; while a suffix to the right indicates that the j-component of the root is − 1. Obviously $i<j$ holds for the positive roots and $i>j$ for the negative ones.

The vectors $h_i, {}_i f_j$ satisfy the Lie products

$$\begin{aligned} &[h_i, h_j] = 0, \quad i,j = 1,2,\ldots,l+1, \\ &[h_i, {}_j f_k] = \delta_{ij}\, {}_j f_k - \delta_{ik}\, {}_j f_k, \\ &[{}_i f_j, {}_k f_l] = \delta_{jk}\, {}_i f_l - \delta_{il}\, {}_k f_j. \end{aligned} \tag{1}$$

According to the Poincaré–Birkchoff–Witt theorem a basis for the universal enveloping algebra of A_l can be chosen as the following set of ordered tensor products of the vectors h_i and $_i f_j$, namely,

$$\Omega : \{ \mathbb{1}, {}_2 f_1^m\, {}_3 f_1^n \cdots {}_1 f_2^r\, {}_1 f_3^s \cdots h_1^t h_2^u \cdots \}, \tag{2}$$

where $m,n,r,s,r,t,u \cdots$ are nonnegative integers, not equal to zero simultaneously. The symbol $\mathbb{1}$ represents the identity element of the enveloping algebra, i.e., for all exponents equal to zero.

The basis for the universal enveloping algebra can be written as

$$\Omega = \Omega_- \Omega_+ \Omega_H,$$

where Ω_- is the enveloping algebra of the vectors ${}_if_j$ $(i>j)$ associated with the negative roots, Ω_+ is the enveloping algebra of the vectors ${}_if_j$ $(i<j)$ associated with the positive roots, and Ω_H is the enveloping algebra of the elements h_i of the Cartan subalgebra.

It is easy to prove by induction in the algebra Ω the following commutation relations

$$\begin{aligned}
&[h_i, {}_jf_k^m] = m(\delta_{ij}\, {}_jf_k^m - \delta_{ik}\, {}_jf_k^m),\\
&[{}_jf_k, {}_kf_j^m] = m\, {}_kf_j^{m-1}(h_j - h_k - m + 1),\\
&[{}_if_j, {}_kf_l^m] = m\, {}_kf_l^{m-1}(\delta_{jk}\, {}_if_l - \delta_{il}\, {}_kf_j).
\end{aligned} \tag{3}$$

3. REPRESENTATIONS OF A_l IN THE SUBSPACE Ω_+

Let us define a basis for the enveloping subalgebra of all the vectors associated with positive roots. This basis is contained in (2) if we take all the exponents of the vectors h_i and ${}_if_j$ $(i>j)$ equal to zero and then choose an arbitrary, but fixed, order product for the vectors ${}_if_j$ $(i<j)$. For the sake of simplicity we choose the following ordering for our basis in Ω_+:

$$\begin{aligned}
&{}_1f_2^{m_{12}}\,{}_1f_3^{m_{13}}\cdots{}_1f_n^{m_{1n}}\,{}_2f_3^{m_{23}}\,{}_2f_4^{m_{24}}\cdots{}_2f_n^{m_{2n}}\cdots{}_{n-1}f_n^{m_{n-1,n}}\\
&\equiv X(m_{12},m_{13},\dots,m_{1n},m_{23},m_{24},\dots,m_{2n},\dots,m_{n-1,n}),
\end{aligned} \tag{4}$$

where $n = l+1$, and the exponents m_{ij} are nonnegative integers. The identity element $\mathbb{1}$ corresponds to the zero value for all the exponents. This basis defines a subspace of Ω, where a linear representation ρ of the elements of the algebra A_l can be defined with the condition

$$\begin{aligned}
&\rho(h_i)\mathbb{1} = \Lambda_i\mathbb{1}, \quad \Lambda_i \in \mathbb{C}, \quad \sum_{i=1}^{l+1}\Lambda_i = 0,\\
&\rho({}_if_j)\mathbb{1} = 0, \quad i>j, \quad i,j = 1,2,\dots,l+1.
\end{aligned} \tag{5}$$

A straight-forward calculation with the help of relations (3) and conditions (5) gives the following linear representations for the elements of the Cartan subalgebra h_i and the root vectors ${}_if_j$ associated with the simple roots:

$$\begin{aligned}
&\rho(h_1)X(m_{12},m_{13},\dots)\\
&\quad = (\Lambda_1 + m_{12} + m_{13} + \cdots + m_{1n})X(m_{12},m_{13},\dots),\\
&\rho(h_2)X(m_{12},m_{13},\dots)\\
&\quad = (\Lambda_2 - m_{12} + m_{23} + m_{24} + \cdots + m_{2n})X(m_{12},m_{13},\dots),\\
&\rho(h_n)X(m_{12},m_{13},\dots)\\
&\quad = (\Lambda_n - m_{1n} - m_{2n} - \cdots - m_{n-1,n})X(m_{12},m_{13},\dots);\\
&\rho({}_1f_2)X(m_{12},m_{13},\dots) = X(m_{12}+1,m_{13},\dots),\\
&\rho({}_2f_3)X(m_{12},m_{13},\dots) = X(m_{12},m_{13},\dots,m_{1n},m_{23}+1,m_{24},\dots)\\
&\qquad\qquad - m_{12}X(m_{12}-1,m_{13}+1,m_{14},\dots),\\
&\rho({}_3f_4)X(m_{12},m_{13},\dots)\\
&\quad = X(m_{12},\dots,m_{1n},m_{23},\dots,m_{2n},m_{34}+1,\dots)\\
&\qquad - m_{13}X(m_{12},m_{13}-1,m_{14}+1,m_{15},\dots)\\
&\qquad - m_{23}X(m_{12},\dots,m_{1n},m_{23}-1,m_{24}+1,m_{25},\dots),\\
&\qquad\vdots\\
&\rho({}_{n-1}f_n)X(m_{12},m_{13},\dots)\\
&\quad = X(m_{12},m_{13},\dots,m_{n-2,n},m_{n-1,n}+1)\\
&\qquad - m_{1,n-1}X(m_{12},\dots,m_{1,n-1}-1,m_{1n}+1,m_{23},\dots)\\
&\qquad - m_{2,n-1}X(m_{12},\dots,m_{2,n-1}-1,m_{2n}+1,m_{34},\dots)\\
&\qquad\vdots\\
&\qquad - m_{n-2,n-1}X(m_{12},\dots,m_{n-2,n-1}-1,\\
&\qquad m_{n-2,n}+1,m_{n-1,n});\\
&\rho({}_2f_1)X(m_{12},m_{13},\dots)\\
&\quad = m_{12}(\Lambda_2 - \Lambda_1 + 1 + m_{23} + m_{24} + \cdots + m_{2n}\\
&\qquad - m_{12} - m_{13} - \cdots - m_{1n})X(m_{12}-1,m_{13},\dots)\\
&\qquad + m_{13}X(m_{12},m_{13}-1,\dots,m_{1n},m_{23}+1,\dots)\\
&\qquad + \cdots\\
&\qquad + m_{1n}X(m_{12},\dots,m_{1n}-1,\dots,m_{2n}+1,\dots),\\
&\rho({}_3f_2)X(m_{12},m_{13},\dots)\\
&\quad = -m_{13}X(m_{12}+1,m_{13}-1,m_{14},\dots)\\
&\qquad + m_{23}(\Lambda_3 - \Lambda_2 + 1 + m_{34} + m_{35} + \cdots + m_{3n}\\
&\qquad - m_{23} - m_{24} - \cdots - m_{2n})X(m_{12},\dots,m_{23}-1,\dots)\\
&\qquad + m_{24}X(m_{12},\dots,m_{23},m_{24}-1,\dots,m_{34}+1,\dots)\\
&\qquad + \cdots + m_{2n}X(m_{12},\dots,m_{1n},m_{23},\dots,m_{2n}-1,\\
&\qquad m_{34},\dots,m_{3n}+1,\dots),\\
&\qquad\vdots\\
&\rho({}_nf_{n-1})X(m_{12},m_{13},\dots)\\
&\quad = -m_{1n}X(m_{12},\dots,m_{1,n-1}+1,m_{1n}-1,m_{23},\dots)\\
&\qquad - m_{2n}X(m_{12},\dots,m_{23},\dots,m_{2,n-1}+1,m_{2n}+1,\dots)\\
&\qquad - \cdots - m_{n-2,n}X(m_{12},\dots,m_{n-2,n-1}+1,\\
&\qquad m_{n-2,n}-1,\dots)\\
&\qquad - \cdots + m_{n-1,n}(\Lambda_n - \Lambda_{n-1} + 1 - m_{n-1,n})\\
&\qquad \times X(m_{12},\dots,m_{n-1,n}-1).
\end{aligned} \tag{6}$$

For the rest of the vectors with positive roots we can use the following relations:

$${}_jf_k = [[[\cdots[{}_jf_{j+1},{}_{j+1}f_{j+2}],{}_{j+2}f_{j+3}],\cdots],{}_{k-1}f_k].$$

These representations are called elementary representations on the subspace Ω_+ and are either irreducible or reducible and indecomposible. They have by definition an extremal vector $\mathbb{1}$, but they can have additional extreme vectors and, if this is the case, these extremal vectors will induce invariant subspaces. The invariant subspaces themselves serve as carrier spaces for elementary subrepresentations. Besides, other types of representations can be obtained on the quotient space with respect to the invariant subspaces. The general method to calculate the extremal vectors will be given in the following section.

4. INVARIANT SUBSPACES ON Ω_-

We can construct a basis for the enveloping algebra of the vectors associated with the negative roots and obtain the linear representations of the elements of A_l by the same method which we used in Sec. 3. In order to make applications to particular examples we restrict ourselves to the algebra A_3 and thereafter generalize to other simple algebras. We choose as our basis on Ω_- the ordered product of vectors

$$_4f_3^l\, {}_4f_2^m\, {}_4f_1^n\, {}_3f_2^r\, {}_3f_1^s\, {}_2f_1^t \equiv X(l,m,n,r,s,t) \tag{7}$$

with l,m,n,r,s,t nonnegative integers, and the identity element $\mathbb{1}$ given by $l=m=n=r=s=t=0$.

If we impose the conditions

$$\begin{aligned}&\rho(h_i)\mathbb{1} = \Lambda_i \mathbb{1}, \quad \Lambda_i \in \mathbb{C}, \ \Lambda_1+\Lambda_2+\Lambda_3+\Lambda_4 = 0,\\ &\rho({}_1f_2)\mathbb{1} = \rho({}_2f_3)\mathbb{1} = \rho({}_3f_4)\mathbb{1} = 0,\end{aligned} \tag{8}$$

then we can define an elementary representation for the vectors of the algebra A_3, namely,

$$\begin{aligned}
&\rho(h_1)X(l,m,n,r,s,t) = (\Lambda_1 - n - s - t)X(l,m,n,r,s,t),\\
&\rho(h_2)X(l,m,n,r,s,t) = (\Lambda_2 - m - r + t)X(l,m,n,r,s,t),\\
&\rho(h_3)X(l,m,n,r,s,t) = (\Lambda_3 - l + r + s)X(l,m,n,r,s,t),\\
&\rho(h_4)X(l,m,n,r,s,t) = (\Lambda_4 + l + m + n)X(l,m,n,r,s,t),\\
&\rho({}_2f_1)X(l,m,n,r,s,t)\\
&\quad = X(l,m,n,r,s,t+1) - rX(l,m,n,r-1,s+1,t)\\
&\qquad - mX(l,m-1,n+1,r,s,t),\\
&\rho({}_3f_2)X(l,m,n,r,s,t) = X(l,m,n,r+1,s,t)\\
&\qquad\qquad - lX(l-1,m+1,r,s,t),\\
&\rho({}_4f_3)X(l,m,n,r,s,t) = X(l+1,m,n,r,s,t),\\
&\rho({}_1f_2)X(l,m,n,r,s,t)\\
&\quad = t(\Lambda_1 - \Lambda_2 + 1 - t)X(l,m,n,r,s,t-1)\\
&\qquad - nX(l,m+1,n-1,r,s,t) - sX(l,m,n,r+1,s-1,t),\\
&\rho({}_2f_3)X(l,m,n,r,s,t)\\
&\quad = r(\Lambda_2 - \Lambda_3 + 1 - r - s + t)X(l,m,n,r-1,s,t)\\
&\qquad - mX(l+1,m-1,n,r,s,t) + sX(l,m,n,r,s-1,t+1),\\
&\rho({}_3f_4)X(l,m,n,r,s,t)\\
&\quad = l(\Lambda_3 - \Lambda_4 + 1 - l - m - n + r + s)\\
&\qquad \times X(l-1,m,n,r,s,t) + mX(l,m-1,n,r+1,s,t)\\
&\qquad + nX(l,m,n-1,r,s+1,t).
\end{aligned} \tag{9}$$

An extremal vector $Y \in \Omega_-$ is defined by the conditions

$$\rho({}_1f_2)Y = \rho({}_2f_3)Y = \rho({}_3f_4)Y = 0 \tag{10}$$

and the corresponding weight $M = (M_1,M_2,M_3,M_4)$ by

$$\rho(h_i)Y = M_i Y, \quad M_1 + M_2 + M_3 + M_4 = 0. \tag{11}$$

From (9) and (10) we find the most simple solutions for the extremal vectors, namely,

(i) $t = \Lambda_1 - \Lambda_2 + 1, \quad l = m = n = r = s = 0,$
$Y = {}_2f_1^{\Lambda_1 - \Lambda_2 + 1},$

(ii) $r = \Lambda_2 - \Lambda_3 + 1, \quad l = m = n = s = t = 0,$
$Y = {}_3f_2^{\Lambda_2 - \Lambda_3 + 1},$

(iii) $l = \Lambda_3 - \Lambda_4 + 1, \quad m = n = r = s = t = 0,$
$Y = {}_4f_3^{\Lambda_3 - \Lambda_4 + 1}.$

From (9) and (11) we obtain for the corresponding weights (we write $M_{\rm I} \equiv \Lambda$)

(i) $M_{\rm II} = (\Lambda_2 - 1, \Lambda_1 + 1, \Lambda_3, \Lambda_4)$,

(ii) $M_{\rm III} = (\Lambda_1, \Lambda_3 - 1, \Lambda_2 + 1, \Lambda_4)$,

(iii) $M_{\rm IV} = (\Lambda_1, \Lambda_2, \Lambda_4 - 1, \Lambda_3 + 1)$.

In order to calculate other extremal vectors we could apply conditions (10) to a general element of Ω_- given as a linear combination of basis vectors with the same weight, as has been worked out by Gruber, Klimyk and Smirnov[3] for the algebra A_2. Instead, we will use another method which is based on properties of extremal vectors given in the same work. Suppose we know some extremal vector $Z_1 \in \Omega_-$, which defines an ideal $I_1 \subset \Omega_-$. If the weight associated with Z_1 is M_1, we can take M_1 as the highest weight of an elementary representation defined on the ideal $I_1 = \Omega_- Z_1$. If this representation contains another extremal vector Z_2 with associate weight M_2, we can construct this new subrepresentation on the ideal $I_2 = \Omega_- Z_2$ with highest weight M_2, and so on. If we have a sequence of ideals $I_i \subset I_{i-1} \subset \cdots \subset I_2 \subset I_1 \subset \Omega_-$, the extremal vectors defining the ideal I_i as a subspace of I_{i-1} will be Z_i, and as a consequence the vector

$$Y_i = Z_i Z_{i-1} \cdots Z_2 Z_1 \tag{12}$$

will be an extremal vector with respect to the whole space Ω_-. An example of this inclusion relation among ideals and the defining extremal vectors is given in Ref. 3.

In our case, we can apply the conditions (i), (ii), and (iii) to the representation defined by the weight $M_{\rm II}$ substituting Λ by $M_{\rm II}$ in (9), and thus we will obtain

(iv) $r = \Lambda_1 - \Lambda_3 + 2, \quad l = m = n = s = t = 0,$
$Y = {}_3f_2^{\Lambda_1 - \Lambda_3 + 2}$

with

$M_{\rm V} = (\Lambda_2 - 1, \Lambda_3 - 1, \Lambda_1 + 2, \Lambda_4)$,

(v) $l = \Lambda_3 - \Lambda_4 + 1, \quad m = n = r = s = t = 0,$
$Y = {}_4f_3^{\Lambda_3 - \Lambda_4 + 1}$

with

$M_{\rm VI} = (\Lambda_2 - 1, \Lambda_1 + 1, \Lambda_4 - 1, \Lambda_3 + 1)$.

Cases (iv) and (v) will give the following extremal vectors with respect to Ω_-:

(iv) $M_{\rm V}$: $Y = {}_3f_2^{\Lambda_1 - \Lambda_3 + 2}\, {}_2f_1^{\Lambda_1 - \Lambda_2 + 1}$,

(v) $M_{\rm VI}$: $Y = {}_4f_3^{\Lambda_3 - \Lambda_4 + 1}\, {}_2f_1^{\Lambda_1 - \Lambda_2 + 1}$.

The same calculation can be applied to $M_{\rm III}$ and $M_{\rm IV}$. In Table I we give all the extremal vectors of A_3 and their weights using this method. In order that an extremal vector exists it is necessary that in each step the values t,r,l should be given by nonnegative integers. This is equivalent to saying that each weight M should satisfy

$$M = \Lambda + l\beta_{43} + m\beta_{42} + n\beta_{41} + r\beta_{32} + s\beta_{31} + t\beta_{21}, \tag{13}$$

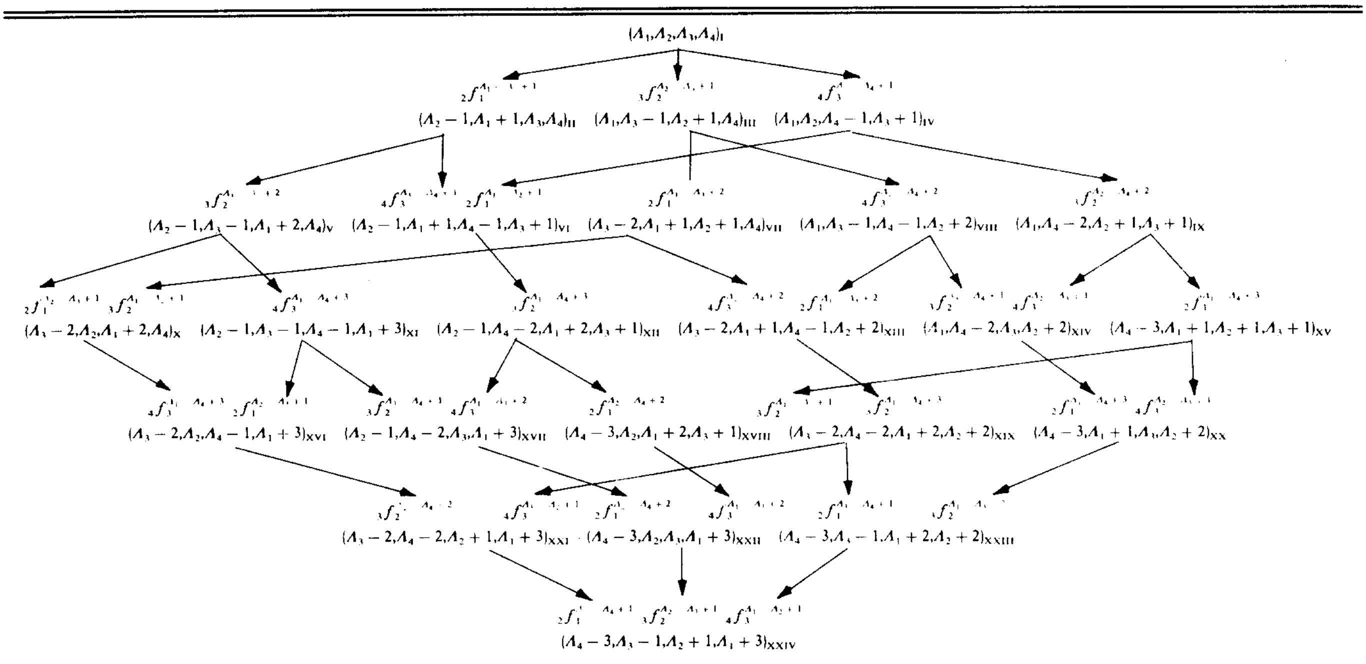
l,m,n,r,s,t, being nonnegative integers and β_{ij} the negative roots.

It will be shown in Sec. 5 that all weights obtained by this method form a subset of the set of weights defined by

$$M = S(\Lambda + R) - R, \tag{14}$$

where R is half the sum of all positive roots of A_3 and S

TABLE I. Extremal vectors in A_3.



represents an element of the Weyl group for A_3. In fact, the cases (i), (ii), and (iii) correspond to the reflections of the weight $\Lambda + R$ on the hyperplanes perpendicular to the simple roots $e_1 - e_2$, $e_2 - e_3$, $e_3 - e_4$, respectively. If the representation defined by the highest weight Λ satisfies condition (13) an invariant subspace is obtained and the representation is indecomposable. If there is no Λ such that (13) is satisfied then the representation is irreducible.

The method we have outlined in order to calculate the extremal vectors for A_3 can be generalized easily to all the algebras A_l, and similar tables as Table I can be calculated by the following algorithm:

(i) Given a particular highest weight $\Lambda = (\Lambda_1, \Lambda_2, ..., \Lambda_{l+1})$ of the algebra A_l, with $\Lambda_1 + \Lambda_2 + \cdots + \Lambda_{l+1} = 0$, we can immediately write down the l weights $M_{\rm II}$, $M_{\rm III}$,...,M_l and extremal vectors corresponding to the simple roots of A_l in the following way:

$$M_{\rm II} = (\Lambda_2 - 1, \Lambda_1 + 1, \Lambda_3, ..., \Lambda_{l+1}),$$
$$Y = {}_2f_1^{\Lambda_1 - \Lambda_2 + 1},$$
$$M_{\rm III} = (\Lambda_1, \Lambda_3 - 1, \Lambda_2 + 1, ..., \Lambda_{l+1}),$$
$$Y = {}_3f_2^{\Lambda_2 - \Lambda_3 + 1},$$
$$\vdots$$
$$M_l = (\Lambda_1, \Lambda_2, ..., \Lambda_{l+1} - 1, \Lambda_l + 1),$$
$$Y = {}_{l+1}f_l^{\Lambda_l - \Lambda_{l+1} + 1},$$

provided that the exponents of the vectors are nonnegative integers. We call these the weights and vectors of the first layer.

(ii) Starting from each of the weights $M_{\rm II}$, $M_{\rm III}$,...,M_l, we will obtain the weights and extremal vectors of the second layer provided the exponents are always nonnegative integers. It may happen that weights of the second layer will be obtained from different weights of the first layer, as can be seen in Table I.

(iii) From the second layer we calculate the weights and vectors of the third layer and so on until we exhaust all the possible weights. To each weight will correspond a vector which will be given with respect to Ω_- (i.e., with respect to the space defined by Λ) by the product of all the extremal vectors given in each layer, taking care that the inclusion relation of the ideals, as explained in (12) must be taken in the correct order.

(iv) If the extremal vector obtained in this fashion does not have the chosen standard order in the basis for the algebra A_l, then it must be brought into standard order by making use of the commutation relations (3). Thus the extremal vectors will in general be given in the form of a nontrivial linear combination over basis elements which belong to a fixed weight subspace.

5. EXTREMAL VECTOR IN SIMPLE LIE ALGEBRAS

Let L denote a simple Lie algebra of rank l with diagonal elements H_i ($i = 1,2,...,l$) for its Cartan subalgebra, shift operators E_α associated with each positive root α, and shift operators $E_{-\alpha}$ associated to each negative root; then the following canonical commutation relations hold:

$$[H_i, E_{\pm\alpha}] = \pm \alpha_i E_{\pm\alpha},$$
$$[E_\alpha, E_{-\alpha}] = \sum_{i=1}^{l} \alpha_i H_i \equiv (\alpha, H), \tag{15}$$

where α_i denotes the ith component of the root α.

We can construct an elementary representation d_Λ defined on the subspace Ω_- of the enveloping algebra of L, with basis

$$\Omega_- = \{1, E^m_{-\alpha} E^n_{-\beta} \cdots E^r_{-\delta}\}, \tag{16}$$

where $m,n,...,r$ are nonnegative integers, and the following conditions hold:

TABLE II. Extremal vectors in simple algebras of rank 2.

A_2

$$\begin{tikzcd}
 & (\Lambda_1,\Lambda_2,\Lambda_3)_{\mathrm{I}} \arrow[dl] \arrow[dr] & \\
\begin{array}{c} {}_2f_1^{\Lambda_1-\Lambda_2+1} \\ (\Lambda_2-1,\Lambda_1+1,\Lambda_3)_{\mathrm{II}} \end{array} \arrow[d] & & \begin{array}{c} {}_3f_2^{\Lambda_2-\Lambda_3+1} \\ (\Lambda_1,\Lambda_3-1,\Lambda_2+1)_{\mathrm{III}} \end{array} \arrow[d] \\
\begin{array}{c} {}_3f_2^{\Lambda_1-\Lambda_3+2} \\ (\Lambda_2-1,\Lambda_3-1,\Lambda_1+2)_{\mathrm{IV}} \end{array} \arrow[dr] & & \begin{array}{c} {}_2f_1^{\Lambda_1-\Lambda_3+2} \\ (\Lambda_3-2,\Lambda_1+1,\Lambda_2+1)_{\mathrm{V}} \end{array} \arrow[dl] \\
 & \begin{array}{c} {}_2f_1^{\Lambda_2-\Lambda_3+1}\,{}_3f_2^{\Lambda_1-\Lambda_2+1} \\ (\Lambda_3-2,\Lambda_2,\Lambda_1+2)_{\mathrm{VI}} \end{array} &
\end{tikzcd}$$

B_2

$$\begin{tikzcd}
 & (\Lambda_1,\Lambda_2)_{\mathrm{I}} \arrow[dl] \arrow[dr] & \\
\begin{array}{c} {}_2f_1^{\Lambda_1-\Lambda_2+1} \\ (\Lambda_2-1,\Lambda_1+1)_{\mathrm{II}} \end{array} \arrow[d] & & \begin{array}{c} f_2^{2\Lambda_2+1} \\ (\Lambda_1,-\Lambda_2-1)_{\mathrm{III}} \end{array} \arrow[d] \\
\begin{array}{c} f_2^{2\Lambda_1+3} \\ (\Lambda_2-1,-\Lambda_1-2)_{\mathrm{IV}} \end{array} \arrow[d] & & \begin{array}{c} {}_2f_1^{\Lambda_1+\Lambda_2+2} \\ (-\Lambda_2-2,\Lambda_1+1)_{\mathrm{V}} \end{array} \arrow[d] \\
\begin{array}{c} {}_2f_1^{\Lambda_1+\Lambda_2+2} \\ (-\Lambda_1-3,\Lambda_2)_{\mathrm{VI}} \end{array} \arrow[dr] & & \begin{array}{c} f_2^{2\Lambda_1+3} \\ (-\Lambda_2-2,-\Lambda_1-2)_{\mathrm{VII}} \end{array} \arrow[dl] \\
 & \begin{array}{c} f_2^{2\Lambda_2+1}\,{}_2f_1^{\Lambda_1-\Lambda_2+1} \\ (-\Lambda_1-3,-\Lambda_2-1)_{\mathrm{VIII}} \end{array} &
\end{tikzcd}$$

$$\begin{tikzcd}
 & G_2\colon (\Lambda_1,\Lambda_2,\Lambda_3)_{\mathrm{I}} \arrow[dl] \arrow[dr] & \\
\begin{array}{c} E_{-1}^{\Lambda_1-\Lambda_2+1} \\ (\Lambda_2-1,\Lambda_1+1,\Lambda_3)_{\mathrm{II}} \end{array} \arrow[d] & & \begin{array}{c} E_{-2}^{3\Lambda_2+1} \\ (-\Lambda_3+\frac{1}{3},-\Lambda_2-\frac{2}{3},-\Lambda_1+\frac{1}{3})_{\mathrm{III}} \end{array} \arrow[d] \\
\begin{array}{c} E_{-2}^{3\Lambda_1+4} \\ (-\Lambda_3+\frac{1}{3},-\Lambda_1-\frac{5}{3},-\Lambda_2+\frac{4}{3})_{\mathrm{IV}} \end{array} \arrow[d] & & \begin{array}{c} E_{-1}^{\Lambda_2-\Lambda_3+2} \\ (-\Lambda_2-\frac{5}{3},-\Lambda_3+\frac{4}{3},-\Lambda_1+\frac{1}{3})_{\mathrm{V}} \end{array} \arrow[d] \\
\begin{array}{c} E_{-1}^{\Lambda_1-\Lambda_3+3} \\ (-\Lambda_1-\frac{8}{3},-\Lambda_3+\frac{4}{3},-\Lambda_2+\frac{4}{3})_{\mathrm{VI}} \end{array} \arrow[d] & & \begin{array}{c} E_{-2}^{3\Lambda_1+3\Lambda_2+5} \\ (\Lambda_1,\Lambda_3-2,\Lambda_2+2)_{\mathrm{VII}} \end{array} \arrow[d] \\
\begin{array}{c} E_{-2}^{3\Lambda_1+3\Lambda_2+5} \\ (\Lambda_2-1,\Lambda_3-2,\Lambda_1+3)_{\mathrm{VIII}} \end{array} \arrow[d] & & \begin{array}{c} E_{-1}^{\Lambda_1-\Lambda_3+3} \\ (\Lambda_3-3,\Lambda_1+1,\Lambda_2+2)_{\mathrm{IX}} \end{array} \arrow[d] \\
\begin{array}{c} E_{-1}^{\Lambda_2-\Lambda_3+2} \\ (\Lambda_3-3,\Lambda_2,\Lambda_1+3)_{\mathrm{X}} \end{array} \arrow[dr] & & \begin{array}{c} E_{-2}^{3\Lambda_1+4} \\ (-\Lambda_2-\frac{5}{3},-\Lambda_1-\frac{5}{3},-\Lambda_3+\frac{10}{3})_{\mathrm{XI}} \end{array} \arrow[dl] \\
 & \begin{array}{c} E_{-2}^{3\Lambda_2+1}\,E_{-1}^{\Lambda_2-\Lambda_3+2} \\ (-\Lambda_1-\frac{8}{3},-\Lambda_2-\frac{2}{3},-\Lambda_3+\frac{10}{3})_{\mathrm{XII}} \end{array} &
\end{tikzcd}$$

$$\rho(H_i)\mathbf{1} = \Lambda_i \mathbf{1}, \quad \Lambda_i \in \mathbb{C}, \; i = 1,2,...,l,$$
$$\rho(E_\alpha)\mathbf{1} = 0 \quad \text{for all positive roots } \alpha \tag{17}$$

An extremal vector is defined by those elements Y of d_Λ which satisfy

$$\rho(H_i)Y = M_i Y, \quad (M_1, M_2, ..., M_l) = M,$$
$$\rho(E_\alpha)Y = 0 \quad \text{for all positive roots } \alpha. \tag{18}$$

In order to construct the extremal vectors we start with those vectors whose weights correspond to simple roots of L.

Let Π: $\{\alpha, \beta, ...\}$ denote the set of simple roots. It is easy to prove by induction that the following commutation relations hold:

$$[H_i, E^m_{\pm\alpha}] = \pm \alpha_i E^m_{\pm\alpha},$$
$$[E_\alpha, E^m_{-\beta}] = m\delta_{\alpha\beta} E^{m-1}_{-\alpha}\{(\alpha,H) - \tfrac{1}{2}(m-1)(\alpha,\alpha)\}. \tag{19}$$

Suppose that $Y = E^m_{-\alpha}$ is an extremal vector. Applying (19) to conditions (18) and (17) one obtains

$$m\{(\alpha,\Lambda) - \tfrac{1}{2}(m-1)(\alpha,\alpha)\} E^{m-1}_{-\alpha} = 0.$$

Since $m = 0$ gives the trivial solution $Y = 1$, we get

$$m = 2(\Lambda,\alpha)/(\alpha,\alpha) + 1 \in \mathbb{N}, \quad \alpha \in \Pi \tag{20}$$

as a condition that $E^m_{-\alpha}$ is an extremal vector, corresponding to the simple root α. N denotes the set of all nonnegative integers.

The weight M, which is associated with this vector Y is given by (18). One obtains with the help of (17)

$$\rho(H_i) E^m_{-\alpha} = (\Lambda_i - (2(\Lambda,\alpha)/(\alpha,\alpha))\alpha_i - \alpha_i) E^m_{-\alpha},$$

or

$$M = \Lambda - (2(\Lambda,\alpha)/(\alpha,\alpha))\alpha - \alpha. \tag{21}$$

Using the property $2(R,\alpha)/(\alpha,\alpha) = 1$ for any simple root α and R equal to half of the sum of all positive roots, we have

$$M = \Lambda - \frac{2(\Lambda + R,\alpha)}{(\alpha,\alpha)}\alpha$$

or

$$M = S_\alpha(\Lambda + R) - R, \tag{22}$$

where S_α denotes the Weyl reflection on that hyperplane which is perpendicular to the simple root α.

In the same fashion as was done for the case of A_l one can make use of an algorithm in order to calculate the extremal vectors for a given representation d_Λ of any simple Lie algebra, using the different layers and inclusion relations among ideals generated by the extremal vectors.

In Table II we give the extremal vectors, the associate highest weights, and the inclusion relations among them for the simple Lie algebras of rank 2. We choose the following basis:

In A_2: $\alpha_1 = (1,-1,0)$, $\alpha_2 = (0,1,-1)$,

In B_2: $\alpha_1 = (1,-1)$, $\alpha_2 = (0,1)$,

In G_2: $\alpha_1 = (1,-1,0)$, $\alpha_2 = \tfrac{1}{3}(-1,2,-1)$.

6. VERMA SUBMODULES OF VERMA MODULES

In the last section we have given a method to calculate the extremal vector of elementary representations; these extremal vectors will induce invariant subspaces and therefore elementary subrepresentations. These type of subrepresentations are called Verma submodules of Verma modules.[1]

The question now arises whether or not the method of the last section gives all possible extremal vectors and thus all the Verma submodules of a given Verma module (of highest weigth Λ) defined on the enveloping algebra of a simple Lie algebra. In order to answer this question we must recall a theorem proved by Bernshtein, Gel'fand, and Gel'fand.[2,4]

BGG theorem: Let d_Λ and d_M be two Verma modules of highest weights Λ and M, respectively; the necessary and sufficient conditions that d_M is contained in d_Λ is that there exists a sequence of positive roots $\gamma_1, \gamma_2, ..., \gamma_k$, such that they satisfy

(i) $M + R = S_{\gamma_k} \cdots S_{\gamma_2} \sigma_{\gamma_1}(\Lambda + R)$,

(ii) $2(S_{\gamma_{i-1}} \cdots S_{\gamma_2} S_{\gamma_1}(\Lambda + R), \gamma_i)/(\gamma_i, \gamma_i) \in \mathbb{N}$,

where $S_{\gamma i}$ is the Weyl reflection for γ_i $(i = 1,2,...,k)$, $S_{\gamma_0} \equiv 1$.

The BGG theorem gives a complete classification of the weights M which correspond to all the possible extremal vectors Y. Thus, given a highest weight Λ, defining a Verma module, we can find all the positive roots that satisfy condition (ii) and the resulting weights M obtained by (i) will give the Verma submodules. Starting from the weights M we can find again all the positive roots satisfying condition (ii) and so we will obtain new weights from (i) defining new Verma submodules, and so on until all the possibilities are exhausted, since the Weyl group is finite.

Nevertheless the BGG theorem does not give the extremal vectors explicitly. We will now show how the extremal vectors can be obtained using the results of Sec. 5. In order to do that it is sufficient to restrict ourselves to the use of one positive root γ in the BGG theorem, with the conditions

(i) $M = S_\gamma(\Lambda + R) - R$,

(ii) $2(\Lambda + R,\gamma)/(\gamma,\gamma) \equiv n \in \mathbb{N}$.

It is known[5] that any element of the Weyl group can be decomposed in the product of simple reflections S_i on the hyperplanes perpendicular to the simple roots α_i $(i = 1,2,...,l)$. Since the product of two reflections does not correspond to a reflection, the element S_γ of the Weyl group must be equivalent to the product of an odd number of simple reflections.

For the sake of simplicity let us suppose that S_γ is equivalent to the product of three simple reflections. We have two cases:

(A) *The simple roots are all different*: for instance, $S_\gamma = S_3 S_2 S_1$.

Define

$$m_1 = \frac{2(\Lambda + R,\alpha_1)}{(\alpha_1\alpha_1)},$$
$$m_2 = \frac{2(S_1(\Lambda + R),\alpha_2)}{(\alpha_2,\alpha_2)}, \tag{23}$$
$$m_3 = \frac{2(S_2S_1(\Lambda + R),\alpha_3)}{(\alpha_3,\alpha_3)}.$$

From condition (i) of the BGG theorem holds

$$M = S_3S_2S_1(\Lambda + R) - R = \Lambda - m_1\alpha_1 - m_2\alpha_2 - m_3\alpha_3. \quad (24)$$

The last equation together with condition (ii) requires

$$n\gamma = m_1\alpha_1 + m_2\alpha_2 + m_3\alpha_3.$$

Since the simple roots are linearly independent we must have

$$\gamma = k_1\alpha_1 + k_2\alpha_2 + k_3\alpha_3, \quad k_i \in \mathbb{N},$$

and therefore

$$m_1 = nk_1, \quad m_2 = nk_2, \quad m_3 = nk_3$$

which means that m_1, m_2, m_3 are nonnegative integers. If we construct the vector

$$Y = E^{m_3}_{-\alpha_3} E^{m_2}_{-\alpha_2} E^{m_1}_{-\alpha_1}. \quad (25)$$

with the associated weight M given by (24), it can be proved that

$$E_{\alpha_1} Y = E_{\alpha_2} Y = E_{\alpha_3} Y = 0,$$

and therefore Y is the extremal vector that satisfies conditions (i) and (ii).

(B) *Two simple roots are equal*: $S_\gamma = S_1S_2S_1$, say. Define

$$m_1 = \frac{2(\Lambda + R, \alpha_1)}{(\alpha_1, \alpha_1)},$$
$$m_2 = \frac{2(S_1(\Lambda + R), \alpha_2)}{(\alpha_2, \alpha_2)}, \quad (26)$$
$$m_3 = \frac{2(S_2S_1(\Lambda + R), \alpha_1)}{(\alpha_1, \alpha_1)}.$$

From condition (i) we have

$$M = \Lambda - m_1\alpha_1 - m_2\alpha_2 - m_3\alpha_1 \quad (27)$$

and from condition (ii) we have

$$n\gamma = (m_1 + m_3)\alpha_1 + m_2\alpha_2;$$

therefore

$$\gamma = k_1\alpha_1 + k_2\alpha_2, \quad k_1, k_2 \in \mathbb{N}$$

or

$$m_1 + m_3 = nk_1, \quad m_2 = nk_2. \quad (28)$$

In this case m_1 and m_3 can be arbitrary numbers. From condition (ii) and Eq. (26) it can be proved that $m_3 = 0$ (trivial case) or $m_3 \neq 0$ and

$$\frac{k_1}{k_2} = -\frac{2(\alpha_1, \alpha_2)}{(\alpha_1, \alpha_1)} \equiv -A_{12} \in \mathbb{N}$$

which, after substitution in (27) and (28), yields

$$\gamma = S_1\alpha_2, \quad m_2 = n. \quad (29)$$

The most general vector satisfying (27) would be

$$Y = \sum_{k=0}^{n} a_k\, E^{n-k}_{-\gamma} E^{k}_{-\alpha_2} E^{(1+A_{12})n - A_{12}k}_{\alpha_1}.$$

The conditions on Y to be an extremal vector lead to some recurrence relations for a_k, which is a function of m_1, m_2, m_3, and since these relations must stop, m_1 and m_3 should be integers. This fact has been proved explicitly in Ref. 3, Theorems 1 and 2, in the case of a Weyl reflection on $\gamma = (1,0,-1)$, where $m_1 = \Lambda_1 - \Lambda_2 + 1$ and $m_3 = \Lambda_2 - \Lambda_3 + 1$ are always integers. This property simplifies the construction of the extremal vectors. If $m_3 > 0$, then the vector

$$Y = E^{m_3}_{-\alpha_1} E^{m_2}_{-\alpha_2} E^{m_1}_{-\alpha_1} \quad (30)$$

satisfies (using (19) and (26))

$$E_{\alpha_1} Y = E_{\alpha_2} Y = 0.$$

If $m_3 < 0$ ($m_1 < 0$ is excluded because $M < \Lambda$) the vector Y defined by

$$E^{-m_3}_{-\alpha_1} Y = E^{m_2}_{-\alpha_2} E^{m_1}_{-\alpha_1} \quad (31)$$

satisfies

$$E^{-m_3}_{-\alpha_1} E_{\alpha_1} Y = E^{-m_3}_{-\alpha_1} E_{\alpha_2} Y = 0,$$

and since Ω_- has no divisors of zero (Ref. 2, Lemma 1)

$$E_{\alpha_1} Y = E_{\alpha_2} Y = 0.$$

The general case should be treated similarly. Suppose that there exists a positive root γ, not equal to a simple root, such that conditions (i) and (ii) of the BGG theorem are satisfied. Then, if

$$S_\gamma = S_{i_k} \cdots S_{i_2} S_{i_1}$$

we define

$$m_1 = \frac{2(\Lambda + R, \alpha_{i_1})}{(\alpha_{i_1}, \alpha_{i_1})},$$
$$m_2 = \frac{2(S_1(\Lambda + R), \alpha_{i_2})}{(\alpha_{i_2}, \alpha_{i_2})}, \quad (32)$$
$$m_k = \frac{2(S_{i_{k-1}} \cdots S_{i_2} S_{i_1}(\Lambda + R), \alpha_{i_k})}{(\alpha_{i_k}, \alpha_{i_k})}$$

provided $S_1(\Lambda + R) - R < \Lambda$ and so on. Then

$$M = \Lambda - m_1\alpha_{i_1} - m_2\alpha_{i_2} - \cdots - m_k\alpha_{i_k} = \Lambda - n\gamma.$$

If some of the simple reflections are equal, the m's can be arbitrary numbers but the recurrence relations imply, as before, that they should be integers.

Another argument to prove that the m's are integer numbers is the following: The extremal vectors can be expressed always in terms of the (nonordered) products of simple root vectors. We choose for our extremal vector Y a product of vectors corresponding to the simple roots which appear in the decomposition of S_γ in terms of simple reflections, $E^{n_k}_{-\alpha_{i_k}} \cdots E^{n_2}_{-\alpha_{i_2}} E^{n_1}_{-\alpha_{i_1}}$. From the BGG theorem, if M satisfies conditions (i) and (ii) an extremal vector must exist satisfying

$$E_{\alpha_1} Y = E_{\alpha_2} Y = \cdots = E_{\alpha_l} Y = 0.$$

This requires that the powers of the simple roots vectors, $n_1, n_2, \ldots, n_k$ must be equal to the coefficient m's given in (32), and therefore all the m's are integer numbers.

Since there are, in general, several descompositions of a Weyl reflection in terms of simple reflections, there will be different products of simple root vectors for the same extremal vector, but all these vectors are equal, as Verma has shown in Ref. 1, Theorem 4.

If all m's are nonnegative, the extremal vector corresponding to M will be

$$Y = E^{m_k}_{-\alpha_{i_k}} \cdots E^{m_2}_{-\alpha_{i_2}} E^{m_1}_{\alpha_{i_1}}. \quad (33)$$

If $m_k < 0$, then the extremal vector is defined by

$$E_{-\alpha_{i_k}}^{-m_k} Y = E_{-\alpha_{i_{k-1}}}^{m_{k-1}} \cdots E_{-\alpha_{i_1}}^{m_1}.$$

It should be noticed that no more than one negative coefficient will appear in (32) because, if m_k and m_{k-1}, for instance, were negative then the vector X defined by

$$E_{\alpha_{i_{k-1}}}^{-m_{k-1}} X = E_{-\alpha_{i_{k-2}}}^{m_{k-2}} \cdots E_{-\alpha_{i_2}}^{m_2} E_{-\alpha_{i_1}}^{m_1}$$

would also be an extremal vector, but its associated weight would not correspond to the reflection of a positive root.

7. SOME EXAMPLES

Example 1: In algebra A_2, $\Lambda = (-1,1,0)$ satisfies the BGG theorem with respect to the positive root $\gamma = (1,0,-1)$, namely, (i) $M = S_\gamma(\Lambda + R) - R = (-2,1,1)$, (ii) $2(\Lambda + R,\gamma)/(\gamma,\gamma) = 1$. Take $S_\gamma = S_2S_1S_2$. Then $m_1 = 2$, $m_2 = 1$, $m_3 = -1$. So we construct the extremal vector as

$${}_3f_2 Y = {}_2f_1\, {}_3f_2^2 = {}_3f_2({}_3f_2\, {}_2f_1 - 2\, {}_2f_1);$$

hence

$$Y = {}_3f_2\, {}_2f_1 - 2\, {}_2f_1.$$

Formally we could have written

$$Y = {}_3f_2^{-1}\, {}_2f_1\, {}_3f_2^2$$

and used the appropriate commutation relation to cancel the negative power.

Example 2: General case in A_2: $\Lambda = (\Lambda_1,\Lambda_2,\Lambda_3)$. The only positive root different from simple ones is $\gamma = (1,0,-1)$. We have two possibilities:

(a) $S_\gamma = S_2S_1S_2$,

$$M = (\Lambda_3 - 2, \Lambda_2, \Lambda_1 + 2),$$

$$Y = {}_3f_2^{\Lambda_1 - \Lambda_2 + 1}\, {}_2f_1^{\Lambda_1 - \Lambda_3 + 2}\, {}_3f_2^{\Lambda_2 - \Lambda_3 + 1};$$

(b) $S_\gamma = S_1S_2S_1$,

$$M = (\Lambda_3 - 2, \Lambda_2, \Lambda_1 + 2),$$

$$Y = {}_2f_1^{\Lambda_2 - \Lambda_3 + 1}\, {}_3f_2^{\Lambda_1 - \Lambda_3 + 2}\, {}_2f_1^{\Lambda_1 - \Lambda_2 + 1}.$$

These two cases correspond in Ref. 3, Corollary 1 to the weight M_{VI}, when (a) $m_3 = \Lambda_1 - \Lambda_2 + 1 < 0$ or (b) $m_3 = \Lambda_2 - \Lambda_3 + 1 < 0$. If m_3 is nonnegative we have case (c) with two possible inclusions;

$$I_{\mathrm{VI}} \subset I_{\mathrm{V}} \subset I_{\mathrm{III}} \subset I_{\mathrm{I}} \quad \text{or} \quad I_{\mathrm{VI}} \subset I_{\mathrm{IV}} \subset I_{\mathrm{II}} \subset I,$$

where I_M represents the ideal I associated with the weights M.

In case (a) with $m_3 < 0$, we can calculate the extremal vector using the relation

$$[{}_2f_1^n, {}_3f_2^m] = \sum_{s=1}^{n} \binom{n}{s} \frac{m!}{(m-s)!}\, {}_3f_1^s\, {}_3f_2^{m-s}\, {}_2f_1^{n-s}$$

and we obtain, with $n = m_2$, $m = m_1$,

$$Y = \sum_{s=0}^{n_2} \binom{m_2}{s} \frac{m_1!}{(m_1 - s)!}\, {}_3f_2^{m_2 - s}\, {}_3f_1^s\, {}_2f_1^{m_2 - s}.$$

With the substitution $m_2 - s = k$, the last expression can be written as

$$Y = \sum_{k=0}^{m_2} \frac{m_2(m_2-1)\cdots(m_2-k+1)}{k!} \times \frac{m_1!}{(k-m_3)!}\, {}_3f_1^{m_2-k}\, {}_3f_2^k\, {}_2f_1^k$$

which is equivalent to the case VI (a) in Corollary 1 of Ref. 3.

Example 3: Let $\Lambda = (-2,2,0,0)$ in A_3. Λ satisfies conditions of the BGG theorem for $\gamma = (1,0,0,-1)$:

(i) $M = S_\gamma(\Lambda + R) - R = (-3,2,0,1)$,

(ii) $\dfrac{2(\Lambda + R,\gamma)}{(\gamma,\gamma)} = 1.$

Take $S_\gamma = S_3S_2S_1S_2S_3$; then

$$m_1 = 1, m_2 = 4, m_3 = 1, m_4 = -3, m_5 = 0,$$

$$Y = {}_3f_2^{-3}\, {}_2f_1\, {}_3f_4^4\, {}_4f_3 = {}_3f_2\, {}_2f_1\, {}_4f_3 - 4\, {}_3f_1\, {}_4f_3.$$

Notice that $m_5 = 0$ means that the last reflection is the identity, and therefore m_4 could be negative.

Example 4: General case in A_3: $\Lambda = (\Lambda_1,\Lambda_2,\Lambda_3,\Lambda_4)$. In Table I we have given all the inclusion relations among the ideals defined by the simple root vectors. The only cases in which the Weyl reflections correspond to (nonsimple) positive roots are the following:

$$M_{\mathrm{X}}:\ \gamma = (1,0,-1,0),\ S_\gamma = S_1S_2S_1 = S_2S_1S_2,$$

$$M_{\mathrm{XVI}}:\ \gamma = (0,1,0,-1),\ S_\gamma = S_2S_3S_2 = S_3S_2S_3,$$

$$\begin{aligned} M_{\mathrm{XXII}}:\ \gamma = (1,0,0,-1),\ S_\gamma &= S_1S_2S_3S_2S_1 \\ &= S_3S_2S_1S_2S_3 \\ &= S_1S_3S_2S_3S_1 \\ &= S_1S_3S_2S_1S_3 \\ &= S_3S_1S_2S_3S_1 \\ &= S_3S_1S_2S_1S_3. \end{aligned}$$

In the last case, care must be taken that the corresponding weights in each simple reflection be always less than Λ. The extremal vectors are given by the product of the simple root vectors of Table I if all the exponents are nonnegatives. The exceptional case is when the last exponent different from zero is negative, in which case this exponent must be canceled by appropriate commutation relations.

ACKNOWLEDGMENT

One of the authors (M.L.) wishes to express his gratitude to the authorities of the Southern Illinois University, in particular to Dean N. J. Doorenbos, for the hospitality extended to him during the visit during which part of this work was done.

[1]D. N. Verma, Bull. Am. Math. Soc. **74**, 160–6 (1968).
[2]I. N. Bernshtein, I. M. Gel'fand, and S. I. Gel'fand, Funkcional. Anal. i Prilozen. **5**, 1–19 (1971) (in Russian).
[3]B. Gruber, A. U. Klimyk, and Y. F. Smirnov, Nuovo Cimento A **69**, 97–127 (1982).
[4]J. Dixmier, *Enveloping Algebras* (North-Holland, Amsterdam, 1977).
[5]N. Jacobson, *Lie Algebras* (Addison-Wesley, New York, 1962).

Lattice degeneracies of fermions

H. Raszillier
Physikalisches Institut, Universität Bonn, D-5300 Bonn 1, West Germany

(Received 17 November 1983; accepted for publication 20 January 1984)

We present a detailed description of the minimal degeneracies of geometric (Kähler) fermions on all the lattices of maximal symmetries in $n = 1,...,4$ dimensions. We also determine the isolated orbits of the maximal symmetry groups, which are related to the minimal numbers of "naive" fermions on the reciprocals of these lattices. It turns out that on the self-reciprocal lattices the minimal numbers of naive fermions are equal to the minimal numbers of degrees of freedom of geometric fermions. The description we give relies on the close connection of the maximal lattice symmetry groups with (affine) Weyl groups of root systems of (semi-) simple Lie algebras.

PACS numbers: 02.20. + b, 05.50. + q, 11.15.Ha, 11.30.Er

I. INTRODUCTION

The present paper is devoted to the investigation of the minimal degeneracies of fermions, if put on those lattices in the n-dimensional Euclidean space E^n ($n = 1,...,4$) which correspond to the maximal discrete symmetries. There are 1, 2, 4, and 9 such lattices in 1, 2, 3, and 4 dimensions, respectively. Of these lattices the primitive (hyper) cubic ones have so far mostly been used in the investigation of lattice field theories. However, recently there has been an increasing interest in other lattices too (simplicial[1] and body-centered hypercubic[2] lattices), for reasons which are related to the question of universality and to a possibly more rapid approach to the continuum limit due to the increased number of degrees of freedom or to the higher order of the symmetry group.

In a lattice (gauge) field theory the (minimal) number of degrees of freedom of the fermions included will depend both on the lattice and on the description chosen for them. There are two attractive descriptions of lattice fermions, which start from slightly different points of view: the "naive" description[3] on the one side and the geometric or Kähler description[4–6] on the other.

The naive description comes from a finite difference approximation of the Dirac equation, invariant under a discrete group, and is thus related to invariant smooth vector fields on the torus T^n, the (first) Brillouin zone of the lattice.[7] The minimal numbers of naive fermions are related to the minimal numbers of (nondegenerate) critical points which any invariant smooth vector field has to have on the Brillouin zones, i.e., on the Dirichlet cells (Wigner–Seitz cells) of the reciprocal lattices.[7]

The geometric description relates[8] the minimal lattice fermionic degrees of freedom to the translationally inequivalent k-faces ($k = 0,...,n$) of some of those parallelohedra [polyhedra with pairwise parallel $(n-1)$-faces] which give normal partitions [i.e., partitions with only complete overlap of $(n-1)$-faces] of the Euclidean space E^n. On the lattices of maximal symmetries the normal partitions are unique and are given by the Dirichlet cells of the lattices.

The minimal numbers of naive fermions and their dependence on the lattice symmetry have been described, for dimensions $n \leqslant 3$, in Ref. 7. Our concern in this paper, which is very closely related to Ref. 8, is therefore mainly in the Kähler description of lattice fermions and in its comparison with the naive description, extended here also to $n = 4$. We thus determine, on the one side, the minimal numbers of degrees of freedom of Kähler fermions on the lattices of maximal symmetries from a detailed combinatorial description of the Dirichlet cells. On the other side, we investigate the actions of the maximal symmetry groups (precisely of their factor groups with respect to the lattice translations) on the Dirichlet cells and determine also for $n = 4$ their isolated orbits which are known[7] to give the minimal sets already mentioned of critical points of invariant smooth vector fields on the Dirichlet cells. Since any lattice of maximal symmetry is reciprocal to a (generally different) lattice of also a maximal symmetry, we thereby get, as far as no obstruction appears from the additional, the physical requirement that the critical points be nondegenerate, the minimal numbers of naive fermions on the reciprocals of the lattices, and can compare them with the minimal numbers of degrees of freedom of Kähler fermions on the lattices. On the self-reciprocal lattices there is no obstruction, and the two numbers turn out to be equal.

Since Kähler fermions are related[9–13] to differential forms their number of degrees of freedom on a manifold M_n of dimension n is *a priori* equal to the dimension 2^n of the space of these forms. For free fermions it can be reduced because this space can be reduced conveniently into invariant subspaces of the Kähler–Dirac operator $d + \delta$ acting in the Hilbert space of square-integrable differential forms (d = exterior differentiation, $\delta = d^*$ its adjoint). In the presence of interaction this is no longer possible, and one remains with an object of 2^n degrees of freedom to which one has to give a physical interpretation. In spite of this problem, the Kähler interpretation of fermions is very attractive as a *lattice* formulation of fermions because of the essentially algebraic nature of its mathematical background, the exterior calculus. Among the steps which one would have to perform in order to give the lattice analog of the exterior calculus as the formulation of the Kähler fermions on lattices in the n-dimensional Euclidean space E^n corresponding to a given discrete symmetry group are:

(i) to give a (minimal) normal partition of the space E^n into polyhedra, which is invariant under the action of the group and minimizes the number

 0022-2488/84/061682-12$02.50

$$N_g = \sum_{k=0}^{n} C_k \tag{1.1}$$

of degrees of freedom of the Kähler lattice fermions, where C_k is the number of k-faces of the polyhedra, which are not equivalent under lattice translations;

(ii) to define on the k-faces of the polyhedra linear functions and to transfer the boundary operation ∂ of the polyhedral complex to the space of functions as the coboundary operator Δ, the analog of the adjoint d^* of the exterior differentiation d;

(iii) to give on this linear space of functions a scalar product, which makes it to a Hilbert space and defines the lattice adjoint Δ^* of Δ, as well as the lattice Dirac–Kähler operator $\Delta + \Delta^*$;

(iv) to perform a spectral analysis of the operator $\Delta + \Delta^*$, in order to find out how the N_g degrees of freedom carried by the lattice fermions manifest themselves in the spectrum;

(v) to investigate the behavior of the spectrum of $\Delta + \Delta^*$ when the minimal lattice translation tends to zero (continuum limit), eventually without going through the detailed spectral analysis.

These steps are, of course, easily performed for primitive hypercubic lattices,[4–6] where the polyhedra are hypercubes, and where $C_k = \binom{n}{k}$ and $N_g = 2^n$ as in (the continuum) E^n. For other lattices, especially in $n = 4$, this is no more so; in this paper we perform step (i) in some detail. The next step, (ii), which refers essentially to the incidence structure of the k-faces of the polyhedra, can be performed by further application of the methods used in this paper. On the remaining steps we shortly comment in the last section of the paper.

We may limit ourselves, as far as geometric lattice fermions are concerned, to the maximal symmetry groups because from their Dirichlet cells one also gets according to a simple rule,[8] by linear deformation, the minimal normal partitions which are invariant only with respect to their subgroups. Naive fermions, however, are much stronger connected to individual groups because of their dependence on the orbit structure. In order to get a complete picture of naive fermions also in $n = 4$ dimensions, one would have to investigate the orbit structure of a very large number of groups, because the total number of symmetry groups increases very fast with dimension n: it is 2 for $n = 1$, 13 for $n = 2$, 73 for $n = 3$, and 710 for $n = 4$.[14] The numbers of maximal symmetry groups are, on the contrary, fairly low: 1, 2, 4, and 9 in $n = 1, 2, 3$, and 4 dimensions, respectively, and equally low is therefore the number of normal parallelohedral partitions of E^n which are of relevance for geometric lattice fermions.

We describe in Sec. II the maximal groups in precise (arithmetic) terms and relate them to quadratic forms (geometry of numbers) and an associated general theory of normal parallelohedral partitions of E^n,[15] as well as to root spaces of semisimple Lie algebras.[16] In Secs. III–VI we describe the Dirichlet cells corresponding to these groups. For most of the maximal groups we do this by exploiting the simple and detailed insight one gains from the Lie algebraic point of view. But in some cases we find it convenient to use also arguments from the theory of normal parallelohedral partitions of E^n.[15] In Sec. VII we compare the minimal numbers N_g of geometric fermions for the Dirichlet cells given by (1.1), with the minimal numbers of naive fermions on the corresponding reciprocal lattices. Finally we summarize (Sec. VIII) the results and make a few comments on some points not touched in the body of the paper.

II. CRYSTALLOGRAPHIC GROUPS

There are various (equivalent) ways to define these groups and several ways to classify them. In our choice[17,18] we will be guided by the wish to arrive in an economic way to our purpose. We consider a lattice $L_n(a_1,..., a_n)$, spanned in the Euclidean space E^n by the n linearly independent vectors $a_i (i = 1,..., n)$,

$$L_n(a_1,..., a_n) = \left\{ x \,\middle|\, x = \sum_{i=1}^{n} n_i a_i,\ n_i \in \mathbb{Z} \right\}. \tag{2.1}$$

With it we associate a group of symmetry operations which leave the point $x = 0$ (chosen as origin) fixed (point group). This group is finite and to it there corresponds a finite group G of $n \times n$ matrices with integer elements and determinant ± 1, i.e., a finite subgroup of $\mathrm{GL}(n,\mathbb{Z})$. The correspondence is not one-to-one since it depends on the chosen basis $a_1,..., a_n$. A change of the basis corresponds to a conjugation of the matrices G by the matrix $A \in \mathrm{GL}(n,\mathbb{Z})$ of the basis transformation,

$$G \to A^{-1} G A, \quad A \in \mathrm{GL}(n,\mathbb{Z}). \tag{2.2}$$

The group G leaves the positive definite quadratic form

$$F(x) = \sum_{i,j=1}^{n} a_{ij}\, x_i x_j, \tag{2.3}$$

with $a_{ij} = a_i \cdot a_j$, invariant; its conjugate (2.2) leaves the form $F(Ax)$ invariant.

To any point group there thus corresponds an equivalence class (2.2) of finite subgroups of $\mathrm{GL}(n,\mathbb{Z})$ and a corresponding equivalence class of invariant positive definite quadratic forms. Inversely, to any finite subgroup G of $\mathrm{GL}(n,\mathbb{Z})$ there corresponds a point group of a lattice. For, if $F(x)$ is an arbitrary positive definite quadratic form in E^n, then the form

$$\varphi(x) = \sum_{g_i \in G} F(g_i x) \equiv \sum_{i,j=1}^{n} b_{ij}\, x_i\, x_j \tag{2.4}$$

is invariant under G; $\varphi(gx) = \varphi(x)$, and G indeed corresponds to a point group of the lattice $L_n(b_1,..., b_n)$, $b_i \cdot b_j = b_{ij}$.

This correspondence between point groups and equivalence classes (2.2) of finite subgroups of $\mathrm{GL}(n,\mathbb{Z})$ is the basis for the *arithmetic* classification of point groups. Point groups are considered to be *arithmetically* distinct, if the equivalence classes (2.2) corresponding to them do not coincide. The number of arithmetically distinct point groups for $n = 1, 2, 3$, and 4 is 2, 13, 73, and 710, respectively.

Not all of these groups are equally interesting for our subject. We may first concentrate on those groups, which are full symmetry groups of a lattice; they are sometimes called *Bravais groups* or also *arithmetic holohedries*. These groups give a classification of lattices; their number is 1, 5, 14, and 64 in one, two, three, and four dimensions, respectively. But the groups which are actually relevant for us are those Bravais groups which are *maximal* finite subgroups of $GL(n,\mathbb{Z})$, i.e., those finite subgroups of $GL(n,\mathbb{Z})$ which are not subgroups of any other finite subgroup of $GL(n,\mathbb{Z})$. These subgroups are sometimes called *maximal arithmetic holohedries* or *maximal Bravais groups*. Their numbers for $n = 1, 2, 3, 4$ are 1, 2, 4, 9, respectively. As we have seen they can be specified by their corresponding lattices. In dimension $n = 2$ these lattices are the quadratic and hexagonal ones, whereas in dimension $n = 3$ the corresponding lattices are hexagonal, cubic primitive, cubic face-centered, and cubic body-centered. A lattice terminology for $n = 4$ is given in Ref. 14.

Coming back to quadratic forms we observe that the Bravais groups are, by their very definition, the full symmetry groups [in $GL(n,\mathbb{Z})$] of the positive definite quadratic forms. The description of all forms corresponding to the same Bravais group falls in the realm of reduction theory (theory of numbers). It is by this theory, in fact, that the maximal Bravais groups in higher dimensions ($n \leqslant 7$) have been determined in terms of their equivalence classes of quadratic forms. We give in Table I for the dimensions $n = 1 - 4$, of interest to us, a complete list of representatives of quadratic forms for all maximal Bravais groups, together with the terminology of the corresponding lattices and the orders of the Bravais groups.[19,20]

All the crystallographic (space) groups can be constructed, as group extensions, starting from the groups of lattice translations and the lattice point groups.[14] But we will be concerned here only with the *symmorphic space groups*, which are the semidirect products of point groups with the lattice translations and thus are in one-to-one correspondence with the arithmetic crystal classes.

There is an approach to part of the crystallographic groups which is related to *root systems R of semisimple Lie algebras*[16] as well as to their automorphism groups $A(R)$, which in most cases coincide with the *Weyl groups* $W(R)$. The lattice generated by a basis of the root system is the affine root space $Q(R)$, the group $A(R)$ is the point group of this lattice [the Weyl group $W(R)$ is its subgroup generated by reflections], and the semidirect product $A(R)\cdot Q(R)$ is the corresponding symmorphic space group [with the *affine Weyl group* $W_a = W(R)\cdot Q(R)$ as subgroup]. There is a corresponding group $A \cdot P^V$ acting on the reciprocal lattice of $Q(R)$, the weight lattice P^V.

One cannot expect to relate in a simple manner all arithmetic crystal classes to semisimple Lie algebras, but for the *maximal* Bravais groups this relation is indeed simple. One can easily identify[21] the Lie algebras related to the maximal Bravais groups in terms of the quadratic forms generated by the root (weight) bases. These identifications are presented in Table II. There is, however, one maximal Bravais group in four dimensions which does not act on a root or weight lat-

TABLE I. Quadratic forms of maximal arithmetic holohedries.

Dimension	Lattice	Quadratic Form	Order of Bravais group
1		$\lambda x_1^2, \quad \lambda>0$	2
2	square, primitive	$\lambda(x_1^2+x_2^2), \quad \lambda>0$	8
	hexagonal, primitive	$\lambda(x_1^2+x_1x_2+x_2^2), \quad \lambda>0$	12
3	hexagonal, primitive	$\lambda x_1^2+\mu(x_2^2+x_2x_3+x_3^2), \quad \lambda,\mu>0$	24
	cubic, primitive	$\lambda(x_1^2+x_2^2+x_3^2), \quad \lambda>0$	48
	cubic, face-centered (F)	$\lambda(x_1^2+x_2^2+x_3^2+x_1x_2+x_1x_3+x_2x_3), \quad \lambda>0$	48
	cubic, body-centered (I)	$\lambda(3x_1^2+3x_2^2+3x_3^2-2x_1x_2-2x_1x_3-2x_2x_3), \quad \lambda>0$	48
4	hexagonal tetragonal	$\lambda(x_1^2+x_2^2)+\mu(x_3^2+x_3x_4+x_4^2), \quad \lambda,\mu>0$	96
	cubic orthogonal, $F(2,3,4)$-centered	$\lambda x_1^2+\mu(x_2^2+x_3^2+x_4^2+x_2x_3+x_2x_4+x_3x_4), \quad \lambda,\mu>0$	96
	cubic orthogonal, $I(2,3,4)$-centered	$\lambda x_1^2+\mu(3x_2^2+3x_3^2+3x_4^2-2x_2x_3-2x_2x_4-2x_3x_4), \quad \lambda,\mu>0$	96
	hypercubic, primitive	$\lambda(x_1^2+x_2^2+x_3^2+x_4^2), \quad \lambda>0$	384
	hypercubic, Z-centered	$\lambda(x_1^2+x_2^2+x_3^2+x_4^2+x_1x_2+x_1x_4+x_2x_3-x_3x_4), \quad \lambda>0$	1152
	diisohexagonal orthogonal, primitive	$\lambda(x_1^2-x_1x_2+x_2^2+x_3^2-x_3x_4+x_4^2), \quad \lambda>0$	288
	diisohexagonal orthogonal RR_2-centered	$\lambda(2x_1^2+2x_2^2+2x_3^2+2x_4^2+x_1x_2-2x_1x_3-2x_1x_4-2x_2x_3-2x_2x_4+x_3x_4), \quad \lambda>0$	144
	icosahedral, primitive	$\lambda(x_1^2+x_2^2+x_3^2+x_4^2+x_1x_2+x_1x_3+x_1x_4+x_2x_3+x_2x_4+x_3x_4), \quad \lambda>0$	240
	icosahedral, SN-centered	$\lambda(2x_1^2+2x_2^2+2x_3^2+2x_4^2-x_1x_2-x_1x_3-x_1x_4-x_2x_3-x_2x_4-x_3x_4), \quad \lambda>0$	240

TABLE II. Identifications of lattices and maximal Bravais groups as root lattices $Q(R)$ [weight lattices $P = P(R^\vee)$] and as automorphism groups of root systems (R) of semisimple Lie algebras.

Dimension	Lattice	Lie algebra (R)	Dual algebra $(R^\vee)$	Bravais group	Lattice type
1		$B_1 (= A_1 = C_1)$	C_1	$A(B_1) = W(B_1)$	$Q = P^\vee$
2	square, primitive	$B_2 (= C_2)$	C_2	$A(B_2) = W(B_2)$	$Q = P^\vee$
	hexagonal, primitive	G_2	G_2	$A(G_2) = W(G_2)$	$Q = P^\vee$
3	hexagonal, primitive	$B_1 + G_2$	$C_1 + G_2$	$A(B_1 + G_2) = W(B_1) \cdot W(G_2)$	$Q = P^\vee$
	cubic, primitive	B_3	C_3	$A(B_3) = W(B_3)$	$Q = P^\vee$
	cubic, face-centered	C_3	B_3	$A(C_3) = W(C_3)$	Q
	cubic, body-centered	C_3	B_3	$A(B_3) = W(B_3)$	$P^\vee$
4	hexagonal, tetragonal	$B_2 + G_2$	$C_2 + G_2$	$A(B_2 + G_2) = W(B_2) \cdot W(B_2)$	$Q = P^\vee$
	cubic orthogonal, $F(2, 3, 4)$-centered	$B_1 + C_3$	$C_1 + B_3$	$A(B_1 + C_3) = W(B_1) \cdot W(C_3)$	Q
	cubic orthogonal, $I(2, 3, 4)$-centered	$B_1 + C_3$	$C_1 + B_3$	$A(C_1 + B_3) = W(C_1) \cdot W(B_3)$	$P^\vee$
	hypercubic, primitive	B_4	C_4	$A(B_4) = W(B_4)$	$Q = P^\vee$
	hypercubic, Z-centered	F_4	F_4	$A(F_4) = W(F_4)$	$Q = P^\vee$
	diisohexagonal orthogonal, primitive	$G_2 + G_2$	$G_2 + G_2$	$A(G_2 + G_2)$	$Q = P^\vee$
	diisohexagonal orthogonal, RR_2-centered				$Q(A_2) \otimes Q(A_2)$
	icosahedral, primitive	A_4	A_4	$A(A_4)$	Q
	icosahedral, SN-centered	A_4	A_4	$A(A_4)$	$P^\vee$

tice, as all the others do, but on a lattice in a tensor product of two root spaces.

III. NORMAL PARTITIONS OF E^n ($n \leqslant 4$)

Around any point of a lattice in the Euclidean space E^n, we can build its *Dirichlet* (or Voronoi) *cell*: It is the (closed) convex polyhedron whose points lie at a distance to the chosen lattice point, which is not larger than their distance to any other lattice point. For $n = 3$ these cells are also known in solid state physics as Wigner–Seitz cells. It is evident that the Dirichlet cells of a lattice give a normal partition of the Euclidean space; this partition is invariant under the Bravais group of the lattice, as well as under translation.

It turns out that at least in E^1, E^2, E^3, and E^4 all normal partitions are affine images of Dirichlet partitions; in these spaces all parallelohedra are affine images of Dirichlet cells.[22,23] The number of combinatorially distinct parallelohedra in E^n is 1, 2, 5, 52 for $n = 1, 2, 3, 4$, respectively. The purely geometric description of the parallelohedra for dimensions $n \leqslant 4$ is known for a relatively long time[22]: for $n = 2$ the corresponding Dirichlet cells are the square and the regular hexagon, whereas for $n = 3$ they are the cube, the cuboctahedron, the rhombic dodecahedron, the regular right hexagonal prism, and the elongated rhombic dodecahedron. The first four are the Dirichlet cells of the primitive, body-centered, and face-centered cubic lattice, and of the hexagonal lattice, respectively, which are lattices of maximal Bravais groups, whereas the fifth is not related to a maximal Bravais group. For *maximal* Bravais groups the only normal partitions invariant under the associated symmorphic space group are the Dirichlet partitions. This is generally not true for the other Bravais groups.

As far as possible we shall derive the combinatorial properties of the Dirichlet cells by assembling them from fundamental domains of space groups. Thereby we can limit ourselves to the subgroups $W \cdot Q$ and $W \cdot P^\vee$ of the Bravais (space) groups $A \cdot Q$ and $A \cdot P^\vee$, respectively. We shall start with the affine Weyl groups $W_a = W \cdot Q$, because they are the simplest ones. The action of the groups $W \cdot P^\vee$ reduces to them and to that of a finite factor group which describes the P-lattice as a "centered" Q-lattice. We shall also consider only irreducible groups (leaving no subspace invariant), which correspond to irreducible root systems, because for the reducible groups the Dirichlet cells are the (set theoretic) products of the cells of the irreducible components (as, e.g., the hexagonal prism is the product of a hexagon and a segment).

We have mentioned that the Dirichlet cell transforms into itself under the action of the Bravais group. Therefore, it consists of whole orbits of the Weyl group $W(R)$. But since the cells transform into each other under the action of lattice translations $Q(R)$, they are sets of nonequivalent orbit points (fundamental domains) of the translation group. So they are, in fact, built out of a number $|W|$ of fundamental domains (Weyl alcoves) of the affine Weyl group $W(R) \cdot Q(R)$, where by $|W|$ we denote the order of the Weyl group $W(R)$. We refer to Ref. 16 for the background to the geometry of the affine Weyl groups and their fundamental domains, the Weyl alcoves. We will describe the minimum of facts we need for the construction of the Dirichlet cells.

We start from a reduced root system R in a real n-dimensional Euclidean space V with scalar product $(\cdot,\cdot)$, i.e., from a set R of vectors, which obeys the following conditions:

(1) R is finite and generates V;

(2) for any vector $\alpha \in R$ there exists a vector α^V in the dual space V^* of V (which we often identify with V through the given scalar product), such that $(\alpha, \alpha^V) = 2$ and such that the automorphism $x \rightarrow x - (x, \alpha^V)\alpha$, $\alpha \in R$, of V transforms R into R;

(3) $(x, \alpha^V) \in \mathbb{Z}$ for all $\alpha, x \in R$;

(4) if $\alpha \in R$, then $2\alpha \notin R$.

The vectors α^V are given by $\alpha^V = 2\alpha/(\alpha,\alpha)$ and form a root system R^V, called dual to R. Further one defines the group $A(R)$ of all automorphisms of V, which transform R into itself, and its subgroup generated by reflections $x \rightarrow x - 2[(x,\alpha)/(\alpha,\alpha)]\alpha$, $x \in V$, $\alpha \in R$, the Weyl group $W(R)$. The groups $W(R)$ and $W(R^V)$ as well as the groups $A(R)$ and $A(R^V)$ are isomorphic.

In the space V we consider the hyperplanes $L_\alpha = \{x | (x,\alpha) = 0\}$, $x \in V$, $\alpha \in R$; the set $V - \{L_\alpha\}$ consists of $|W|$ connected pieces, which are simplicial cones, the *Weyl chambers*. One can choose in the root system R a basis $B = \{\alpha_1,\dots,\alpha_n\}$, which spans V and has the property that

$$\mathrm{Ch} = \{x \in V | (x,\alpha_i) > 0, \quad i = 1,\dots, n\} \tag{3.1}$$

is a Weyl chamber. A basis has the property that any root vector can be expressed in terms of its vectors with integer, either all nonnegative or all nonpositive, coefficients. There is a maximal root $\tilde{\alpha} = \Sigma_{i=1}^n n_i\alpha_i$ such that for any root $\Sigma_{i=1}^n p_i\alpha_i$ one has $n_i \geqslant p_i$, $1 \leqslant i \leqslant n$.

We will exploit the facts that (a) the Weyl group $W(R)$ acts simply transitive on the set of Weyl chambers and (b) the closure $\overline{\mathrm{Ch}}$ of a Weyl chamber is a fundamental domain of the Weyl group $W(R)$. But we are, in fact, interested in the fundamental domain of the affine Weyl group $W_a = W(R) \cdot Q(R)$, where

$$Q(R) = \left\{x \in V \,\middle|\, x = \sum_{i=1}^n n_i\alpha_i, \quad n_i \in \mathbb{Z}\right\} \equiv L_n(\alpha_1,\dots,\alpha_n) \tag{3.2}$$

is the (group of) lattice (translations). This domain is the closure $\overline{C}$ of the *Weyl alcove C* and is described by the set of inequalities

$$\begin{aligned} &(\alpha_i, x) \geqslant 0, \quad i = 1,\dots, n \quad (\text{the closure } \overline{\mathrm{Ch}} \\ &\qquad\qquad \text{of the Weyl chamber Ch}), \\ &(\tilde{\alpha}^V, x) \leqslant 1, \end{aligned} \tag{3.3}$$

where $\tilde{\alpha}^V$ is the maximal root of the dual root system R^V, considered in the dual basis $B^V = \{\alpha_1^V,\dots, \alpha_n^V\}$, $\alpha_i^V = 2\alpha_i/(\alpha_i,\alpha_i)$. Since $(\alpha_i, x) > 0$ is equivalent to $(\alpha_i^V, x) > 0$, the Weyl alcove C is described by

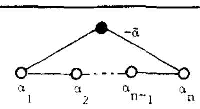
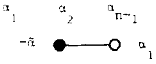
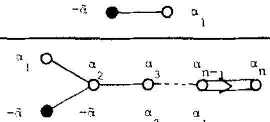
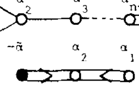
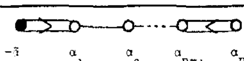
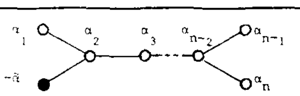
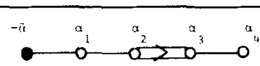
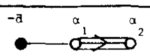
$$\begin{aligned} &(\alpha_i^V, x) > 0, \quad i = 1,\dots, n, \\ &(\tilde{\alpha}^V, x) < 1, \end{aligned} \tag{3.4}$$

in terms of the dual root system R^V. Its closure $\overline{C}$ is

$$\begin{aligned} &(\alpha_i^V, x) \geqslant 0, \quad i = 1,\dots, n, \\ &(\tilde{\alpha}^V, x) \leqslant 1. \end{aligned} \tag{3.5}$$

The geometric properties of the root systems R, with the maximal roots $\tilde{\alpha}$ added to the corresponding bases $B = \{\alpha_1,\dots,\alpha_n\}$, are given by the *extended Dynkin diagrams* (Table III).

TABLE III. Extended Dynkin diagrams of simple Lie algebras, including the root bases $\{\alpha_1,\dots,\alpha_n\}$ and the maximal roots $\tilde{\alpha}$: two vertices are connected by h (= 1, 2 or 3) lines if the corresponding roots are not orthogonal; one of the two roots is h times longer than the other (as the inequality sign on the lines shows).

R	Type of R^V	Extended Dynkin diagram (R)	Order of $W(R)$	Order of $A(R)$	Stability group of $\tilde{\alpha}$ in $W(R)$
A_n, $n \geqslant 2$	A_n		$(n+1)!$	$2\|W(R)\|$	$W(A_{n-2})$
A_1	A_1		2	2	
B_n $n \geqslant 3$	C_n		$2^n \cdot n!$	$\|W(R)\|$	$W(A_1) \cdot W(B_{n-2})$
B_2	$C_2 (= B_2)$		8	8	$W(A_1)$
C_n, $n \geqslant 2$	B_n		$2^n \cdot n!$	$\|W(R)\|$	$W(C_{n-1})$
D_n, $n \geqslant 4$	D_n		$2^{n-1} \cdot n!$	$2\|W(R)\|$, $n \neq 4$ $3\|W(R)\|$, $n = 4$	$W(A_1) \cdot W(D_{n-2})$
F_4	F_4		$2^7 \cdot 3^2 = 1152$	$\|W(R)\|$	$W(C_3)$
G_2	G_2		12	$\|W(R)\|$	$W(A_1)$

The inequalities (3.4) describe simplexes in E^n: the first inequalities represent the positive orthant (a simplicial cone) and the last one reduces the cone by cutting it with the hyperplane

$$H(C):\ (\tilde{\alpha}^V, x) = 1. \tag{3.6}$$

The Dirichlet cell $D_Q \equiv D_Q(R)$ of the lattice $Q(R)$ is

$$D_Q = \bigcup_{w \in W(R)} w\overline{C}. \tag{3.7}$$

Its combinatorics lies in its boundary ∂D_Q, which is entirely described in terms of the hyperplanes $H(wC)$, $w \in W(R)$, generated from $H(C)$ by the Weyl group. The Dirichlet cell can be described analytically as

$$D_Q = \bigcap_{w \in W(R)} D_w(R),$$
$$D_w(R) = \{x \mid (w\tilde{\alpha}^V, x) \leqslant 1\}, \tag{3.8}$$

the intersection of the half-spaces generated from $(\tilde{\alpha}^V, x) \leqslant 1$ by the Weyl group.

IV. COMBINATORIAL STRUCTURE OF THE DIRICHLET CELLS IN $Q(R)$

(i) $(n-1)$-boundaries: The number of $(n-1)$-boundaries [$(n-1)$-faces] of the cell is equal to the number of distinct hyperplanes $H(wC)$, i.e., to the numbers of vectors in the orbit of the maximal root $\tilde{\alpha}^V$ under the Weyl group. Since the Weyl group acts transitively on the set of roots of equal length, the number of $(n-1)$-boundaries equals the number of roots of squared length $(\tilde{\alpha}^V, \tilde{\alpha}^V)$, which is equal to the order $|W|$ of the Weyl group divided by the order $|G|$ of the stability group G of $\tilde{\alpha}^V$: $|W|/|G|$. The stability group of $\tilde{\alpha}^V$ can be read off the Dynkin diagrams. Namely, if $w = w_{\alpha^V}$, $\alpha^V \in R^V$, then

$$w_{\alpha^V}\tilde{\alpha}^V = \tilde{\alpha}^V - 2\frac{(\tilde{\alpha}^V,\alpha^V)}{(\alpha^V,\alpha^V)}\alpha^V \equiv \tilde{\alpha}^V \leftrightarrow (\tilde{\alpha}^V,\alpha^V) = 0. \tag{4.1}$$

The subset of a root system R^V, which lies in a subspace of E^n (here the space orthogonal to $\tilde{\alpha}^V$) is itself a (perhaps reducible) root system; its basis is given in our case by the subset of $B(R^V) = \{\alpha_1^V,...,\alpha_n^V\}$ which obeys $(\tilde{\alpha}^V,\alpha_i^V) = 0$. G is the Weyl group of this root system.

(ii) $(n-2)$-boundaries: These boundaries are determined by two equalities of the type

$$(w\alpha_i^V, x) = 0, \quad i = 1,..., n, \quad w \in W,$$
$$(w\tilde{\alpha}^V, x) = 1, \tag{4.2}$$

supplemented by the rest of inequalities

$$(w\alpha_j^V, x) \geqslant 0, \quad j = 1,..., n, \quad j \neq i,$$

of (3.5). Since the closed Weyl chamber $\overline{\mathrm{Ch}}$: $(\alpha_i^V, x) \geqslant 0$, $i = 1,..., n$, is a fundamental domain of the Weyl group, the orbits of $(\alpha_i^V, x) = 0$ and $(\alpha_j^V, x) = 0$ will be distinct for $i \neq j$. Therefore, we shall investigate successively the number of distinct pairs (4.2) for fixed values of $i = 1,..., n$. But not all values have to be considered, since some of them disappear as boundaries because the simplexes they separate belong to the same $(n-1)$-boundary. Suppose that $(\tilde{\alpha}^V,\alpha_i^V) = 0$; then $(w_{\alpha_i^V}\tilde{\alpha}^V, x) \equiv (\tilde{\alpha}^V, x)$ and $(\alpha_i^V, x) \geqslant 0 \rightarrow (w_{\alpha_i^V}\alpha_i^V, x) = -(\alpha_i^V, x) \leqslant 0$ and the contiguous Weyl alcoves C and $w_{\alpha_i^V}C$ belong to the same $(n-1)$-boundary $(\tilde{\alpha}^V, x) = 1$. We need thus be concerned only with those α_i^V's which are not orthogonal to $\tilde{\alpha}^V$; these are very few: two for root systems of type A_n and one for all the others.

In order to determine the number of $(n-2)$-boundaries we have first to determine the number of distinct pairs

$$(w\alpha_i^V, x) = 0,$$
$$(w\tilde{\alpha}^V, x) = 1, \quad w \in W, \tag{4.3}$$

and then divide it by the order of the equivalence classes of pairs corresponding to the same $(n-2)$-boundaries. The number of equations (4.3) is $|W(R^V)|/|G(\tilde{\alpha}^V,\alpha_i^V)|$, where $G(\tilde{\alpha}^V,\alpha_i^V) = G(\tilde{\alpha}^V) \cap G(\alpha_i^V)$ is the stability group of $\tilde{\alpha}^V,\alpha_i^V$ in $W(R^V)$. The equivalence class mentioned has two elements, corresponding to the Weyl group generated by α_i^V. All these numbers can be read from the extended Dynkin diagram and the results are given in Table IV.

(iii) $(n-3)$-boundaries: Similar reasoning is valid here;

TABLE IV. k-boundaries (F_k) of Dirichlet cells D_Q of the lattices $Q(R)$.

Dimension n	Lattice	F_0	F_1	F_2	F_3	F_4	$\chi_s = \sum_{k=0}^{n-1}(-1)^k F_k$
1	$Q(B_1)$	2	1	—	—	—	2
2	$Q(B_2)$	4	4	1	—	—	0
	$Q(G_2)$	6	6	1	—	—	0
3	$Q(B_1+G_2)$	12	18	8	1	—	2
	$Q(B_3)$	8	12	6	1	—	2
	$Q(C_3)$	14	24	12	1	—	2
4	$Q(B_1+C_3)$	28	62	48	14	1	0
	$Q(B_2+G_2)$	24	48	34	10	1	0
	$Q(G_2+G_2)$	36	72	48	12	1	0
	$Q(A_4)$	30	70	60	20	1	0
	$Q(B_4)$	16	32	24	8	1	0
	$Q(F_4)$	24	96	96	24	1	0

TABLE V. Equivalence classes (C_k) of k-boundaries of Dirichlet cells D_Q and minimal numbers N_g of degrees of freedom of Kähler fermions on the lattices $Q(R)$.

Dimension n	Lattice	C_0	C_1	C_2	C_3	C_4	$\chi = \sum_{k=0}^{n} (-1)^k C_k$	$N_g = \sum_{k=0}^{n} C_k$
1	$Q(B_1)$	1	1	—	—	—	0	2
2	$Q(B_2)$	1	2	1	—	—	0	4
	$Q(G_2)$	2	3	1	—	—	0	6
3	$Q(B_1+G_2)$	2	5	4	1	—	0	12
	$Q(B_3)$	1	3	3	1	—	0	8
	$Q(C_3)$	3	8	6	1	—	0	18
4	$Q(B_1+C_3)$	3	11	14	7	1	0	36
	$Q(B_2+G_2)$	2	7	9	5	1	0	24
	$Q(G_2+G_2)$	4	12	13	6	1	0	36
	$Q(A_4)$	4	15	20	10	1	0	50
	$Q(B_4)$	1	4	6	4	1	0	16
	$Q(F_4)$	3	24	32	12	1	0	72

one only has to take those distinct pairs of α_i^V,α_j^V, $i\neq j$ for which $(\tilde{\alpha}^V,\alpha_i^V,\alpha_j^V)$ does not fall into orthogonal subsystems, and one gets the number of such (i,j)-boundaries as $|W(R^V)|/(|G(\tilde{\alpha}^V,\alpha_i^V,\alpha_j^V)|\cdot|W(\alpha_i^V,\alpha_j^V)|)$, where $W(\alpha_i^V,\alpha_j^V)$ is the Weyl group of the root system generated by α_i^V,α_j^V and $G(\tilde{\alpha}^V,\alpha_i^V,\alpha_j^V)$ is the stability group of $\tilde{\alpha}^V,\alpha_i^V,\alpha_j^V$ in $W(R^V)$. The group $W(\alpha_i^V,\alpha_j^V)$ gives the size of the equivalence classes of hyperplanes

$$(w\tilde{\alpha}^V, x) = 1,$$
$$(w\alpha_i^V, x) = 0,$$
$$(w\alpha_j^V, x) = 0,$$

which correspond to the same $(n-3)$-boundaries.

(iv) $(n-4)$-boundaries: Their number is given by $|W(R^V)|/(|G(\tilde{\alpha}^V,\alpha_i^V,\alpha_j^V,\alpha_k^V)||W(\alpha_i^V,\alpha_j^V,\alpha_k^V)|)$, where the magnitudes involved are completely analogous to those described at the other dimensions.

So far we have derived the combinatorial structure of the Dirichlet cells of the $Q(R)$-type lattices, looked upon from the spherical topology; the number χ_s computed in Table IV equals the Euler characteristics of the sphere S_{n-1} in E^n as it should. We will now be concerned, in addition, with the equivalence structure of the boundaries induced by the lattice translations of $Q(R)$. This structure is described, as can easily be seen, by a region homeomorphic to an open half-sphere of S_{n-1}. The lattice translations which realize the equivalences, are given by the set of roots of minimal length in the root system R which generates the lattice $Q(R)$. Namely, the vector which touches the hyperplane $(\tilde{\alpha}^V, x) = 1$ orthogonally is $\beta = \tilde{\alpha}^V/(\tilde{\alpha}^V,\tilde{\alpha}^V) = \frac{1}{2}(\tilde{\alpha}^V)^V$. Since $2\beta = (\tilde{\alpha}^V)^V \in R$ and $(\beta,\beta) = 1/(\tilde{\alpha}^V,\tilde{\alpha}^V)$, $\tilde{\alpha}^V$ being a root in R^V of maximal length, $\alpha \equiv 2\beta$ is of minimal length in $(R^V)^V \equiv R$.

The essential group theoretic property which comes into play here is the simple transitivity of the *affine* Weyl group on the alcoves and the fact that their closures are fundamental domains. The necessary data for the equivalence calculations can be taken again from the extended Dynkin diagrams.

(i') $(n-1)$-boundaries: Here we know already that any $(n-1)$-boundary is common to two Dirichlet cells; therefore, the number of equivalence classes of these boundaries is half the number of boundaries. Group-theoretically, we would have to reason in the following way: the number of neighboring cells corresponding to $\tilde{\alpha}^V$ is given by the order of the Weyl group $W(\tilde{\alpha}^V)$, $|W(\tilde{\alpha}^V)| = 2$. This gives the desired reduction factor.

(ii') $(n-2)$-boundaries: Let us assume that such a boundary is given by $(\tilde{\alpha}^V, x) = 1$, $(\alpha_i^V, x) = 0$. From the extended Dynkin diagram we compute the number of alcoves contiguous to this boundary as $|W(\tilde{\alpha}^V,\alpha_i^V)|\cdot|G(\tilde{\alpha}^V,\alpha_i^V)|$, of which $|W(\alpha_i^V)|\cdot|G(\tilde{\alpha}^V,\alpha_i^V)|$ belong to one cell. The number of cells contiguous at this boundary is thus $|W(\tilde{\alpha}^V,\alpha_i^V)|/|W(\alpha_i^V)|$.

(iii') $(n-k)$-boundaries: The general result for any boundary of type $\tilde{\alpha}^V\alpha_{i_1}^V \cdots \alpha_{i_{k-1}}^V$ is that the number of contiguous cells is given by $|W(\tilde{\alpha}^V\alpha_{i_1}^V \cdots \alpha_{i_{k-1}}^V)|/|W\alpha_{i_1}^V \cdots \alpha_{i_{k-1}}^V)|$. Combining this with the number of boundaries computed under (i)–(iv), we get for the number of equivalence classes of $(n-k)$-boundaries $(k = 1,..., n)$ of type $\tilde{\alpha}^V\alpha_{i_1}^V \cdots \alpha_{i_{k-1}}^V$: $|W(\alpha_i^V \cdots \alpha_n^V)|/(|W(\tilde{\alpha}^V\alpha_{i_1}^V \cdots \alpha_{i_{k-1}}^V)|\cdot|G(\tilde{\alpha}^V\alpha_{i_1}^V \cdots \alpha_{i_{k-1}}^V)|)$. They are all collected in Table V.

V. LATTICES OF TYPE $P \equiv P(R)$

So far we have investigated the combinatorial properties of the Dirichlet cells of the root lattice $Q(R)$ by starting from the fundamental domain of the affine Weyl group $W_a = W(R)\cdot Q(R)$, which is a simplex. Now we discuss the cells of the weight lattices $P(R)$. A lattice $P(R)$ is defined as

$$L_n(\omega_1,...,\omega_n) = \left\{x \mid x = \sum_{i=1}^{n} n_i\omega_i\right\},$$
$$(\omega_i,\alpha_j^V) = \delta_{ij}, \quad i,j = 1,..., n, \tag{5.1}$$

where $\{\alpha_1^V,...,\alpha_n^V\}$ is the basis of the dual root system R^V. The lattice $Q(R)$ is a sublattice of $P(R)$ and the group $W_a = W(R)\cdot Q(R)$ is a subgroup of $W_a' = W(R)\cdot P(R)$.

TABLE VI. k-boundaries (F_k) of Dirichlet cells D_P of the lattices $P(R)$.

Dimension n	Lattice	F_0	F_1	F_2	F_3	F_4	$\chi_s = \sum_{k=0}^{n-1} (-1)^k F_k$
3	$P(B_3)$	24	36	14	1	—	2
4	$P(C_1+B_3)$	48	96	64	16	1	0
	$P(A_4)$	120	240	150	30	1	0

The order F of the factor group W_a'/W_a [which is isomorphic to $P(R)/Q(R)$] gives the number of fundamental domains of W_a' in a Weyl alcove $\overline{C}$ (the fundamental domain of W_a). The division of the alcove $\overline{C}$ into fundamental domains C_P of W_a' is performed by the subgroup Γ_C (or order f) of W_a', transforming C into itself,

$$\Gamma_C = \{\gamma_i | \gamma_i(C) = C\}, \quad \gamma_i \in W_a'. \tag{5.2}$$

Its $f-1$ nontrivial elements γ_i are described by

$$\gamma_i = t(\omega_i) w_i w_0, \tag{5.3}$$

where the index i exhausts the set J of values for which $n_i = 1$ in the maximal root $\tilde{\alpha}^V = \Sigma_{i=1}^n n_i \alpha_i^V$ of R^V; w_0 is the (unique) element of $W(R)$, which transforms the Weyl alcove C into $-C$, w_i is the same for the group $W(R_i)$ [considered as a subgroup of $W(R)$] of the root system R_i determined by the Dynkin diagram of R, if the positive root α_i is deleted. $t(\omega_i)$ is the translation by the corresponding weight ω_i. The weights ω_i, $i \in J$ form a system of representatives of $P(R)/Q(R)$ in $P(R)$. The operations $w_i w_0$ are most simply displayed by the automorphism group they induce on the extended Dynkin diagram of R^V.

There are two cases we have to analyze: $P(B_3)$ and $P(A_4)$.

$P(B_3)$: Since $B_3^V = C_3$ and $\tilde{\alpha}^V = 2\alpha_1^V + 2\alpha_2^V + \alpha_3^V$, there is only one element in the set J: $i = 3$. The group $P(B_3)/Q(B_3)$ is of order 2 ($f = 2$); ω_3 is a representative in $P(B_3)$. The group operation induced by $w_3 w_0$ in $B_3^V = C_3$ is $-\tilde{\alpha}^V \leftrightarrow \alpha_3^V$, $\alpha_1^V \leftrightarrow \alpha_2^V$. With these elements one easily observes that $\overline{C}_P$ is given by

$$\begin{aligned} &\overline{C}_P: \\ &(\tilde{\alpha}^V, x) \leqslant 1, \\ &(\omega_3, x) \leqslant \tfrac{1}{2}(\omega_3, \omega_3), \\ &(\alpha_1^V, x) \geqslant 0, \\ &(\alpha_2^V, x) \geqslant 0, \\ &(\alpha_3^V, x) \geqslant 0. \end{aligned} \tag{5.4}$$

Its boundary structure is as follows:

$$\begin{aligned} &\text{1-boundaries:} && \alpha_1^V\alpha_2^V, \alpha_1^V\alpha_3^V, \alpha_2^V\alpha_3^V, \tilde{\alpha}^V\alpha_2^V, \tilde{\alpha}^V\alpha_3^V, \\ &&& \alpha_1^V\omega_3, \alpha_2^V\omega_3, \alpha_3^V\omega_3, \tilde{\alpha}^V\omega_3; \\ &\text{0-boundaries:} && \alpha_1^V\alpha_2^V\alpha_3^V, \\ &&& \alpha_1^V\alpha_2^V\omega_3, \alpha_1^V\alpha_3^V\omega_3, \tilde{\alpha}^V\alpha_2^V\omega_3, \tilde{\alpha}^V\alpha_3^V\omega_3, \end{aligned} \tag{5.5}$$

expressed in terms of combinations of inequalities (5.4).

$P(A_4)$: Here we have $A_4^V = A_4$ and $\tilde{\alpha} = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$, so the set J includes all values $i = 1,...,4$ ($f = 5$) and we have $\omega_1,...,\omega_4$ as representatives of $P(A_4)/Q(A_4)$ in $P(A_4)$. The group operations $w_i w_0$ induce on the extended Dynkin diagram cyclic permutations, whereby for a given i, $-\tilde{\alpha} \to \alpha_i$. The domain $\overline{C}_P$ is

$$\begin{aligned} &(\omega_i, x) \leqslant \tfrac{1}{2}(\omega_i, \omega_i), \\ &(\alpha_j, x) \geqslant 0, \quad i,j = 1,..., n, \end{aligned} \tag{5.6}$$

with the boundary structure:

$$\begin{aligned} &\text{2-boundaries:} && \alpha_i\alpha_j \quad (i<j), \\ &&& \alpha_i\omega_j \quad (i \neq j), \\ &&& \omega_i\omega_j \quad (i<j), \\ &\text{1-boundaries:} && \alpha_{i_1}\alpha_{i_2}\alpha_{i_3} \quad (i_1<i_2<i_3), \\ &&& \alpha_{i_1}\alpha_{i_2}\omega_{i_3} \quad (i_1<i_2), \\ &&& \alpha_{i_1}\omega_{i_2}\omega_{i_3} \quad (i_2<i_3), \\ &&& \omega_{i_1}\omega_{i_2}\omega_{i_3} \quad (i_1<i_2<i_3), \\ &\text{0-boundaries:} && \alpha_1\alpha_2\alpha_3\alpha_4, \\ &&& \alpha_{i_1}\alpha_{i_2}\alpha_{i_3}\omega_{i_4} \quad (i_1<i_2<i_3), \\ &&& \alpha_{i_1}\alpha_{i_2}\omega_{i_3}\omega_{i_4} \quad (i_1<i_2, i_3<i_4), \\ &&& \alpha_{i_1}\omega_{i_2}\omega_{i_3}\omega_{i_4} \quad (i_2<i_3<i_4), \\ &&& \omega_1\omega_2\omega_3\omega_4, \end{aligned} \tag{5.7}$$

where $i_m = 1,...,4$, $i_m \neq i_{m'}$.

TABLE VII. Equivalence classes (C_k) of k-boundaries of Dirichlet cells D_P and minimal numbers N_g of degrees of freedom of Kähler fermions on the lattices $P(R)$.

Dimension n	Lattice	C_0	C_1	C_2	C_3	C_4	$\chi = \sum_{k=0}^{n} (-1)^k C_k$	$N_g = \sum_{k=0}^{n} C_k$
3	$P(B_3)$	6	12	7	1	—	0	26
4	$P(C_1+B_3)$	6	18	19	8	1	0	52
	$P(A_4)$	24	60	50	15	1	0	150

TABLE VIII. k-boundaries (F_k) of the Dirichlet cell D_T of the lattice $Q(A_2) \otimes Q(A_2)$.

Dimension n	Lattice	F_0	F_1	F_2	F_3	F_4	$\chi_s = \sum_{k=0}^{n-1} (-1)^k F_k$
4	$Q(A_2) \otimes Q(A_2)$	102	216	144	30	1	0

The construction of the Dirichlet cell $D_P \equiv D_P(R)$,

$$D_P = \bigcup_{w \in W(R)} w\overline{C}_P, \tag{5.8}$$

proceeds along the same lines as that of $D_Q(R)$, only has one to take into account an additional set of hyperplanes $(\omega_i, x) = \frac{1}{2}(\omega_i, \omega_i)$, besides or instead of $(\widetilde{\alpha}^V, x) = 1$. The relevant orthogonality properties now also include $(\omega_i, \alpha_j^V) = \delta_{ij}$. The numbers F_k of k-boundaries of D_P are given in Table VI. These numbers already show that the cells $D_P(B_3)$ and $D_P(A_4)$ are primitive parallelohedra,[15] i.e., any k-boundary ($k = 0,..., n$) is contiguous to $n - k + 1$ cells; this gives the equivalence classes of k-boundaries of Table VII.

VI. THE LATTICE $Q_T \equiv Q(A_2) \otimes Q(A_2)$

This is the diisohexagonal tetragonal RR_2-centered lattice. It is, from the point of view of root systems, exceptional. The Weyl group of the A_2 root systems induce on this lattice an automorphism group (of rotations) of order 36. The full automorphism (point) group A_T is of order $36 \cdot 4 = 144$; the additional generators are the reflection in one A_2 and the permutation of the two A_2 root systems. Instead of computing the fundamental domain $\overline{C}_T$ of the (symmorphic) lattice group $A_T \cdot Q_T$ generated by A_T and the lattice translations and deriving therefrom the combinatorial properties of the Dirichlet cell

$$D_T = \bigcup_{g \in A_T} g\overline{C}_T, \tag{6.1}$$

we use results from the theory of parallelohedra and of positive definite quadratic forms.[15] The quadratic form (metric) of the lattice, generated by $R(A_2) \times R(A_2)$ is, in the basis,

$$\begin{aligned} e_1 &= \alpha_1 \otimes \alpha_1, \\ e_2 &= \alpha_2 \otimes \alpha_2, \\ e_3 &= \alpha_1 \otimes \alpha_2, \\ e_4 &= \alpha_2 \otimes \alpha_1, \end{aligned} \tag{6.2}$$

$$\Big(\sum_{i=1}^{4} x_i e_i\Big)^2 = 2(2x_1^2 + 2x_2^2 + 2x_3^2 + 2x_4^2 + x_1 x_2 + x_3 x_4 - 2x_1 x_3 - 2x_1 x_4 - 2x_2 x_3 - 2x_2 x_4). \tag{6.3}$$

The parallelohedron corresponding to it turns out from this to be a limit case of a primitive parallelohedron of type III described in detail by Voronoi.[15] Its numbers of k-boundaries ($k = 0,...,4$) are listed in Table VIII and its equivalence classes of k-boundaries are given in Table IX.

VII. COMPARISON WITH NAIVE FERMIONS

The minimal number of naive fermions is related to the (equivalence class of) points on the Dirichlet cells, which belong to isolated orbits of the factor group of the crystallographic group with respect to the lattice translations. These orbits can be particularly easily determined for those groups which are affine Weyl groups $W_a = W(R) \cdot Q(R)$. Affine Weyl groups are generated by reflections on hyperplanes. Since their action on the alcoves is simply transitive, their isolated orbits are the extremal points, the 0-boundaries of the alcoves. These are the images under the Weyl group of the points symbolized by $\alpha_1^V \cdots \alpha_n^V$ and by $\widetilde{\alpha}^V \widetilde{\alpha}_{i_1}^V \cdots \alpha_{i_{n-1}}^V$. The first of these points is the center of the cell D_Q and the others are on its boundaries. The orbit of the point $\widetilde{\alpha}^V \alpha_{i_1}^V \cdots \alpha_{i_{n-1}}^V$ has $|W(R^V)|/|W(\alpha_{i_1}^V \cdots \alpha_{i_{n-1}}^V)|$ elements. If the set $\widetilde{\alpha}^V \alpha_{i_1}^V \cdots \alpha_{i_{n-1}}^V$ splits into (at most) two mutually orthogonal systems $\widetilde{\alpha}^V \alpha_{i_1}^V \cdots \alpha_{i_{k-1}}^V$ and $\alpha_{i_k}^V \cdots \alpha_{i_{n-1}}^V$ the orbit lies on the $(n-k)$-boundaries of D_Q of type $\widetilde{\alpha}^V \alpha_{i_1}^V \cdots \alpha_{i_{k-1}}^V$ ($k = 1,..., n$). The numbers p_{n-k} of Table X come from all the nonequivalent $(n-k)$-boundaries of such types.

If the group $A(R) \cdot Q(R)$ is larger than the affine Weyl group, as it happens for $R = G_2 + G_2$ and $R = A_4$, where $|A|/|W| = 2$, the fundamental domain is half the Weyl alcove. We may, however, look simply at the fixed points of the factor group $A \cdot Q / W \cdot Q$ on the Weyl alcove.

For the root system A_4 the nontrivial element of this group corresponds to the inversion $\alpha_i \to -\alpha_i$ ($i = 1,...,4$). Its fixed points are the images under W of the point $(\alpha_1, x) = \frac{1}{2}$, $(\alpha_4, x) = \frac{1}{2}$ on the $\widetilde{\alpha}\alpha_2\alpha_3$-boundary. These are the centers of the 3-boundaries of D_Q. Thus, for $W \cdot Q$ the points of isolated orbits are the center and the 0-boundaries of D_Q (the W-images of the 0-boundaries of the Weyl alcove) and for $A \cdot Q$ the centers of the 3-boundaries add to them.

In $G_2 + G_2$ the fundamental domain of the affine Weyl group is the product of two simplexes (triangles) with the

TABLE IX. Equivalence classes (C_k) of k-boundaries (and of isolated orbit points) of the Dirichlet cell D_T and minimal number N_g of degrees of freedom of Kähler fermions on the lattice $Q(A_2) \otimes Q(A_2)$.

Dimension n	Lattice	C_0	C_1	C_2	C_3	C_4	$\chi = \sum_{k=0}^{n} (-1)^k C_k$	$N_g = \sum_{k=0}^{n} C_k$
4	$Q(A_2) \otimes Q(A_2)$	20	54	48	15	1	0	138

TABLE X. Equivalence classes (p_k) of isolated orbit points of the groups on k-boundaries of the Dirichlet cells D_Q, i.e., on the Brillouin zones of the reciprocal lattices $P(R^V)$, compared to N_g of $D_P(R^V)$.

Dimension n	Lattice	Bravais group	p_0	p_1	p_2	p_3	p_4	$N_C = \sum_{k=0}^{n} p_k$	N_g	Reciprocal lattice	
1	$Q(B_1)$	$A(B_1) = W(B_1)$	1	1	—	—	—	2	2	$P(C_1)$	$= Q(B_1)$
2	$Q(B_2)$	$A(B_2) = W(B_2)$	1	2	1	—	—	4	4	$P(C_2)$	$= Q(B_2)$
	$Q(G_2)$	$A(G_2) = W(G_2)$	2	3	1	—	—	6	6	$P(G_2)$	$= Q(G_2)$
3	$Q(B_1) + Q(G_2)$	$A = W(B_1) \cdot W(G_2)$	2	5	4	1	—	12	12	$P(C_1) + P(G_2)$	$= Q(B_1) + Q(G_2)$
	$Q(B_3)$	$A = W(B_3)$	1	3	3	1	—	8	8	$P(C_3)$	$= Q(B_3)$
	$Q(C_3)$	$A = W(C_3)$	3	—	6	1	—	10	26	$P(B_3)$	
4	$Q(B_1) + Q(C_3)$	$A = W(B_1) \cdot W(C_3)$	3	3	6	7	1	20	52	$P(C_1) + P(B_3)$	
	$Q(B_2) + Q(G_2)$	$A = W(B_2) \cdot W(G_2)$	2	7	9	5	1	24	24	$P(C_2) + P(G_2)$	$= Q(B_2) + Q(G_2)$
	$Q(G_2) + Q(G_2)$	$A(G_2 + G_2)$	4	12	13	6	1	36	36	$P(G_2) + P(G_2)$	$= Q(G_2) + Q(G_2)$
	$Q(A_4)$	$A(A_4)$	4	—	—	10	1	15	150	$P(A_4)$	
	$Q(B_4)$	$A = W(B_4)$	1	4	6	4	1	16	16	$P(C_4)$	$= Q(B_4)$
	$Q(F_4)$	$A = W(F_4)$	3	24	32	12	1	72	72	$P(F_4)$	$= Q(F_4)$

extreme points of higher stability group. The additional symmetry operation in $A \cdot Q$ is the interchange of the two root systems G_2; it has no isolated fixed points.

The corresponding numbers for A_4 and $G_2 + G_2$ are, again, given in Table X. We consider now the groups $A \cdot P$ and start from their subgroups $W \cdot P$. For B_3, where we have $A = W$ and where the factor group $W \cdot P / W \cdot Q$ has order 2, we get the fixed points of $W \cdot Q$ and, in addition, the images of the points $\omega_3 \alpha_1^V \alpha_2^V$ and $\tilde{\alpha}^V \omega_3 \alpha_3^V$. Thus, the points of isolated orbits are the center of the cell, those of its 2-boundaries, and its 0-boundaries.

In the fundamental domain of $W \cdot P(A_4)$ the only fixed points of the factor group $W \cdot P / W \cdot Q$ are $\alpha_1 \alpha_2 \alpha_3 \alpha_4$ and $\omega_1 \omega_2 \omega_3 \omega_4$. The additional fixed points, coming from the group $A \cdot P / W \cdot P$, are $\omega_1 \alpha_2 \alpha_3 \alpha_4, \omega_2 \alpha_1 \alpha_3 \alpha_4, \omega_3 \alpha_1 \alpha_2 \alpha_4$, and $\omega_4 \alpha_1 \alpha_2 \alpha_3$. The points of isolated orbits of the Dirichlet cell $D_P(A_4)$ thus are for $W \cdot P$ the center of the cell and its 0-boundaries, whereas for $A \cdot P$ the centers of the 3-boundaries add to them.

The numbers p_k for the cells D_P are given in Table XI.

There remains the (self-reciprocal) lattice $Q_T = Q(A_2) \otimes Q(A_2)$. From its full symmetry group we can determine, with a certain amount of computation, the stability groups of the various (inequivalent) k-boundaries of the Dirichlet cell D_T. The result is that on any k-boundary ($k = 0,...,4$) there is precisely one point of an isolated orbit.

VIII. COMMENTS

We have derived for the lattices of maximal symmetries in $n = 1,...,4$ dimensions the numbers N_g of degrees of freedom of geometric fermions, which are connected with the translationally inequivalent k-boundaries ($k = 0,..., n$) of their Dirichlet cells. These numbers, summarized in Tables V, VII, and IX, are minima by the construction of the cells from the fundamental domains of the symmorphic space groups associated to the maximal Bravais groups. The cells do not provide the *unique* minimal *partition*: the dual partition, which corresponds to $C_k \leftrightarrow C_{n-k}$ ($k = 0,..., n$) in these tables, is another such partition. It is given by the polyhedra which are the convex hulls of the centers of the Dirichlet cells contiguous at 0-boundaries. The dual partitions can be described with the methods used in this paper. We only mention here that for the hexagonal lattice $Q(G_2)$, the body-centered cubic lattice $P(B_3)$, and the SN-centered icosahedral lattice $P(A_4)$ they are simplicial partitions (simplicial lattices). For fermions (either naive or geometric) the choice between Dirichlet or dual partition is not so important, but it is important for objects with degrees of freedom sitting on boundary elements of given order k, as, e.g., vector (gauge) bosons ($k = 1$).

We have also determined, for the actions of the symmorphic space groups corresponding to the maximal Bra-

TABLE XI. Equivalence classes (p_k) of isolated orbit points of the groups on k-boundaries of the Dirichlet cells D_P, i.e., on the Brillouin zones of the reciprocal lattices $Q(R^V)$, compared to N_g of $D_Q(R^V)$.

Dimension n	Lattice	Bravais group	p_0	p_1	p_2	p_3	p_4	$N_C = \sum_{k=0}^{n} p_k$	N_g	Reciprocal lattice
3	$P(B_3)$	$A = W(B_3)$	6	—	7	1	—	14	18	$Q(C_3)$
4	$P(C_1) + P(B_3)$	$A = W(C_1) \cdot W(B_3)$	6	6	7	8	1	28	36	$Q(B_1) + Q(C_3)$
	$P(A_4)$	$A(A_4)$	24	—	—	15	1	40	50	$Q(A_4)$

vais groups in $n = 4$ dimensions, the orbits which are isolated in their strata. For reasons of completeness we have rederived these orbits here also for $n \leqslant 3$. The points of these orbits (Tables X and XI) are minimal critical sets for any smooth invariant vector field on the Dirichlet cell[24] and thereby give lower bounds N_c for the numbers of naive fermions on the reciprocal lattice, for which the Dirichlet cell plays the role of the Brillouin zone.[7] The physical (particle) interpretation of the vector field requires, however, that its critical points should not be degenerate. On several of the lattices which are not self-reciprocal, e.g., on those of type A_4, this requirement cannot be imposed on the minimal set of critical points since the orbit structure does not allow them to obey the Poincaré–Hopf theorem. On them the lower bound N_c cannot be saturated by physically acceptable vector fields. The minimal numbers of naive fermions for the corresponding groups are then necessarily higher than the numbers in Tables X and XI show; their precise values have to be determined from a detailed orbit analysis of the group action. Anyhow the location of these fermions on the Brillouin zone will then no more be completely specified by symmetry. This phenomenon has already been encountered for certain three-dimensional nonmaximal groups; the groups of the lattices $Q(A_4)$, $P(A_4)$ are similar to the space groups T_d^2, T_d^3.

On the self-reciprocal lattices the situation is much simpler. Here we have $p_k = C_k$ and therefore the equality $\Sigma_{k=0}^n(-1)^k p_k = \Sigma_{k=0}^n(-1)^k C_k = 0$, by which the Poincaré–Hopf theorem allows us to saturate the bound N_c by physically interesting vector fields with only nondegenerate critical points. For the maximal groups corresponding to the self-reciprocal lattices, one can therefore make a precise statement: the minimal numbers of naive fermions are given by N_c of Table X and the location of these fermions is on the centers of the (translationally inequivalent) k-boundaries of the Dirichlet cells (= Brillouin zones); for these lattices $N_c = N_g$.

In order to have the complete computational instrument (the appropriate l^2-cohomologies) of a lattice theory, one has to know, in addition to the partitions giving the minimal degrees of freedom, also the incidence structure of the boundaries of the polyhedra and the appropriate scalar product for the (Hilbert) space of linear functions defined on the boundaries. The incidence structure can easily be derived from the geometric insight gained by the Lie algebraic methods used in this paper. The appropriate scalar product should be the one induced[25,26] by that of the Euclidean space E^n.

We have not discussed in this paper the continuum limit. It has been commented upon in Refs. 6 and 8, and a two-dimensional example has been worked out in Ref. 27. The mathematical background for a discussion can be traced from Refs. 26 and 28.

There remains, finally, the problem of determining the minimal numbers of degrees of freedom for lattice fermions for the *nonsymmorphic* space groups corresponding to the maximal Bravais groups. We expect that, analogously to naive fermions,[7] one will get a strong increase of these numbers compared to the symmorphic groups. We illustrate in an appendix on the (single) two-dimensional case the situation to be expected generally.

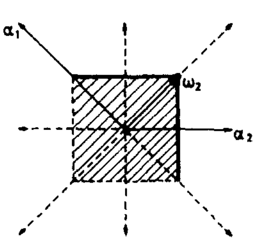


FIG. 1.

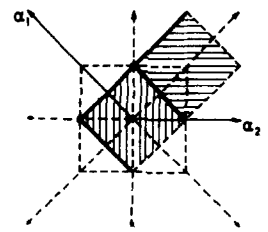


FIG. 2.

ACKNOWLEDGMENTS

My insight into the combinatorics of Dirichlet partitions has much gained from the suggestion of Professor E. Brieskorn and Dr. W. Plesken to look at them from the point of view of root systems of Lie algebras. I am much indebted also to Dr. P. Slodowy for his assistance in finding my way through the geometry of root spaces. I am grateful to R. Flume and V. Rittenberg for their comments.

APPENDIX

The two-dimensional maximal Bravais groups are the Weyl groups $W(B_2)$ and $W(G_2)$. To $W(G_2)$ there corresponds a single space group, which is symmorphic, whereas to $W(B_2)$ there correspond two, one of which is nonsymmorphic. Its factor group with respect to lattice translations may be represented by the four proper rotations (even number of Weyl reflections) and by the four improper rotations (odd number of Weyl reflections) combined with a translation by the weight ω_2. In Fig. 1 we display the location of the minimal numbers $N_g = 4$ of degrees of geometric lattice fermions for the symmorphic group ($C_0 = 1$, $C_1 = 2$, $C_2 = 1$) and in Fig. 2 the same for the nonsymmorphic group ($N_g = 8$: $C_0 = 2$, $C_1 = 4$, $C_2 = 2$).

[1]J. M. Drouffe and K. J. M. Moriarty, Nucl. Phys. B **220**, 253 (1983); J. M. Drouffe and K. J. M. Moriarty, "U(2) Four-Dimensional Simplicial Lattice Gauge Theory," Preprint Ref. TH. 3624-CERN, 1983; J. M. Drouffe, K. J. M. Moriarty, and C. N. Monhas, "U(1) Four-Dimensional Gauge Theory on a Simplicial Lattice," Preprint Ref. TH. 3653-CERN, 1983; J. M. Drouffe, K. J. M. Moriarty, and C. N. Monhas, Comput. Phys. Commun. (Netherlands) **30**, 249 (1983).
[2]W. Celmaster, Phys. Rev. D **26**, 2955 (1982); W. Celmaster and F. Krausz, Phys. Rev. D **28**, 1527 (1983); W. Celmaster, Phys. Rev. D **28**, 2076 (1983).
[3]L. Susskind, Phys. Rev. D **16**, 3031 (1977).
[4]P. Becher, Phys. Lett. B **104**, 221 (1981).

[5]P. Becher and H. Joos, Z. Phys. C **15**, 343 (1982).
[6]J. Rabin, Nucl. Phys. B **201**, 315 (1982).
[7]H. Raszillier, J. Math. Phys. **24**, 642 (1983).
[8]H. Raszillier, Phys. Lett. B **131**, 145 (1983).
[9]E. Kähler, Rend. Mat. (Roma) **21**, 425 (1962).
[10]W. Graff, Ann. Inst. H. Poincaré A **29**, 85 (1978).
[11]I. M. Benn and R. W. Tucker, Commun. Math. Phys. **89**, 341 (1983).
[12]T. Banks, Y. Dothan, and D. Horn, Phys. Lett. B **117**, 413 (1982).
[13]D. Hestenes, *Space-Time Algebra* (Gordon and Breach, New York, 1966).
[14]H. Brown, R. Bülow, J. Neubüser, H. Wondratschek, and H. Zassenhaus, *Crystallographic Groups in Four-Dimensional Space* (Wiley, New York, 1978).
[15]G. Voronoi, J. Reine Angew. Math. **134**, 198 (1908); **136**, 67 (1909).
[16]N. Bourbaki, *Groupes et algèbres de Lie* (Hermann, Paris, 1968), Chap. IV–VI.
[17]R. L. E. Schwarzenberger, *N-Dimensional Crystallography* (Pitman, San Francisco, 1980).
[18]J. J. Burckhardt, *Die Bewegungsgruppen der Kristallographie* (Birkhäuser, Basel, 1966), 2nd ed.
[19]E. C. Dade, Ill. J. Math. **9**, 99 (1965).
[20]S. S. Ryshkov, Proc. Steklov Inst. Math. **128**, 217 (1972).
[21]G. Maxwell, J. Algebra **35**, 159 (1975).
[22]B. Delaunay, Izv. Akad. Nauk SSSR **1**, 79 (1929); **2**, 147 (1929).
[23]A. D. Alexandrow, *Konvexe Polyeder* (Akademie-Verlag, Berlin, 1958).
[24]L. Michel, Rev. Mod. Phys. **52**, 617 (1980).
[25]H. Whitney, *Geometric Integration Theory* (Princeton U. P., Princeton, NJ, 1957).
[26]J. Dodziuk, Am. J. Math. **98**, 79 (1976).
[27]M. Göckeler, Z. Phys. C **18**, 323 (1983).
[28]V. K. Patodi, in the Proceedings of the International Congress on Mathematics, Vancouver, 1974.

Orbit spaces of low-dimensional representations of simple compact connected Lie groups and extrema of a group-invariant scalar potential[a]

Jai Sam Kim[b]
California Institute of Technology, Pasadena, California 91125

(Received 7 September 1983; accepted for publication 4 November 1983)

Orbit spaces of low-dimensional representations of classical and exceptional Lie groups are constructed and tabulated. We observe that the orbit spaces of some single irreducible representations (adjoints, second-rank symmetric and antisymmetric tensors of classical Lie groups, and the defining representations of F_4 and E_6) are warped polyhedrons with (locally) more protrudent boundaries corresponding to higher level little groups. The orbit spaces of two irreducible representations have different shapes. We observe that dimension and concavity of different strata are not sharply distinguished. We explain that the observed orbit space structure implies that a physical system tends to retain as much symmetry as possible in a symmetry breaking process. In Appendix A, we interpret our method of minimization in the orbit space in terms of conventional language and show how to find all the extrema (in the representation space) of a general group-invariant scalar potential monotonic in the orbit space. We also present the criterion to tell whether an extremum is a local minimum or maximum or an inflection point. In Appendix B, we show that the minimization problem can always be reduced to a two-dimensional one in the case of the most general Higgs potential for a single irreducible representation and to a three-dimensional one in the case of an even degree Higgs potential for two irreducible representations. We explain that the absolute minimum condition prompts the boundary conditions enough to determine the representation vector.

PACS numbers: 02.20.Qs, 11.15.Ex

1. INTRODUCTION

Since the discovery of the Higgs mechanism,[1] it has been employed almost exclusively in the gauge symmetry breaking problem because it breaks a local gauge symmetry without damaging the renormalizability.[2] Though it is not the only known mechanism to do such a job,[3,4] certainly it is the only tractable one. Partly due to its tractability it drew considerable attention from theoretical physicists despite some ugly features. There is a consensus that, though it may not be a fundamental mechanism, it would describe the effective phenomena arising from some unknown fundamental interactions. It was applied to the unification of electromagnetic and weak interactions[5] with great success and subsequently to fancier grand unification theories.[6] Here some major difficulties arose, namely, the gauge hierarchy problem[7] and proliferation of Higgs parameters, etc. The spontaneous symmetry breaking problem in supersymmetric theories[8] is one of the most popular problems these days. Since the spontaneous symmetry breaking mechanism was devised by Landau[9] to explain continuous structural phase transitions in crystals, the mechanism has been widely employed in condensed matter physics.[10] Spontaneous symmetry breaking is one of the most fundamental phenomena observed in nature. It is no wonder that there exist several extensive review articles on this subject.[11,12]

The technical problem of minimizing the scalar potential or the thermodynamic potential and finding the symmetry of the vacuum or the equilibrium state has been considered to be a formidable task among theoretical physicists. Unification theorists could only check the list of possible symmetry breaking directions without knowing whether and when the symmetry is broken in certain directions. Condensed matter theorists had to use an abbreviated potential. Our geometrical method provides the most appropriate language for the problem. It gives accurate minimizing solutions for a general Higgs potential of single irreducible representations and for a general even-degree Higgs potential of two irreducible representations. We leave to an interested reader the analysis of a general even-degree Higgs potential of three irreducible representations, which will give needed solutions to some unification models, e.g., the E_6 model with a Higgs assignment in 27, 78, and 351 representations.[13]

The geometrical method of minimizing the Landau–Higgs potentials, devised by the author,[14–17] reduces the problem to one of finding "contours" of directional minima. It is based on the observation[18] that the orbits and the conjugacy classes of subgroups are the relevant quantities to describe the minimum of the Landau–Higgs potential which is invariant under a linear transformation of a compact Lie group on the scalar fields (or a finite group on the order parameters). Hilbert[19] proved that there is a basic set of invariants such that all other invariants are expressed as their polynomials and provided a systematic method to find all

[a] Work supported in part by the U. S. Department of Energy under Contract No. DE-AC-03-81-ER40050.

[b] Address after 1 September 1983: Department of Physics, Washington University, St. Louis, MO 63130, after 1 November 1983: Institute for Theoretical Physics, SUNY, Stony Brook, NY 11794, and after 1 January 1984: Department of Physics, Johns Hopkins University, Baltimore, MD 21218.

the basic invariants. It has been known[20] that invariants specify orbits, i.e., one can view an orbit as a point in an $(l+1)$-dimensional vector space, $(l+1)$ being the number of independent basic invariants. How can we describe a direction in such a space? Indeed there is a set of parameters[21] that can be used for such purpose. "Orbit parameters" are defined to be dimensionless ratios of invariant polynomials. These parameters can be considered as some set of generalized angles specifying a direction in the representation space. Their ranges being bounded, they occupy a localized region (called the "orbit space") in the orbit parameter space, which can be regarded as an l-dimensional vector space.

Since the scalar potential is a group invariant function, it can be expressed in terms of the basic invariants. But a classical Higgs potential is restricted to be a fourth-degree polynomial of the scalar fields due to renormalizability. Because of this restriction it is normal that only a subset of all the basic invariants appear in the Higgs potential, and that the orbit parameters formed from this subset appear linearly.[22] The potential can be written in terms of the norm of the field and a few orbit parameters. For a given set of orbit parameters we can survey the behavior, particularly extrema, of the potential along the corresponding direction in the field component space. By varying the orbit parameters, we can survey the whole space in search of the absolute minimum. Because of the linearity, the absolute minimum of the potential occurs on the most protrudent portions of the boundary of the orbit space formed from the fourth-degree Higgs potential, which is a projection of the complete orbit space.

The potential can be minimized abstractly for a general representation of a general compact group. The difficult part of extremizing the potential in the conventional methods[23,24] has as its counterpart in our method the problem of finding the orbit space boundary, which is unique for each different representation. In our original works our method for constructing the projected orbit space was empirical and we used the Michel–Radicati conjecture[25] for one irreducible representation (irrep)[26] and the Gell-Mann–Slansky conjecture[27] for two irreps as a guide for finding the orbit space boundary. Then our results were tested with the boundary conditions. It was realized that we need not solve high degree algebraic equations to find orbit space boundaries. The procedure is facilitated by some general mathematical results.[28,29,16] Using these results, we look for branching rules[27,30] and singlet forms of the given representation under various subgroups, starting from the highest level to successively lower levels. In any case we need to know at least this much information to specify the absolute minimum. "Usable boundaries" (where the potential may have a minimum) correspond to higher symmetry groups. In practice one finds the whole boundary before he reaches the lowest level.

Much work has been done by mathematicians[31,32,29] on the structure of the complete orbit space. Their results were originally derived from the properties of linear actions of compact transformation groups. However, Ref. 32 deals with the relationship between orbits and invariants. Recently a comprehensive review article[29] has been published for physicists. Our formalism is entirely based on the invariant polynomials and requires less mathematical background, presenting a concrete and intuitive picture without losing generality. The main result is that the orbit space consists of some l-dimensional volume occupied by the generic stratum of the lowest level symmetry group with all the other strata of higher symmetries forming the singular boundaries. Equivalently, the generic stratum occupies an open, dense, topologically connected region and thus the boundaries must belong to the lower dimensional strata. It was explicitly shown[12,29] that a lower dimensional stratum of a higher symmetry is a subspace which is spanned by the gradients of basic invariant polynomials. (This is equivalent to our boundary conditions.[16])

In this paper we present concrete examples, showing that lower dimensional strata of higher symmetry groups always form the boundaries of higher dimensional strata of lower symmetry groups. We also observe that high symmetry strata are normally[33] more protrudent than lower symmetry ones, which was conjectured in our earlier works.[14–16] This protrusiveness of orbit spaces (defined in terms of ratios of invariant polynomials) is important physically because it indicates that a physical system tries to retain the highest symmetries possible when the spontaneous symmetry breaking takes place, which is the spirit of the Michel–Radicati and the Gell-Mann–Slansky conjectures. It also makes our method powerful. The hierarchy of protrusiveness on the boundaries is essential to predict how small the little groups can be in the presence of nonmonotonic orbit parameters.[17]

In Sec. 2, we briefly review the minimization problem in orbit space. In Sec. 3, we construct orbit spaces of adjoint representations. We observe that they form polyhedrons as conjectured in Ref. 16. In Sec. 4, second-rank symmetric and antisymmetric tensors of all the classical Lie groups and other low-dimensional single irreps are analyzed. We observe that the tensors have the same orbit spaces as adjoint representations of other groups and that there is an interesting relationship between the number of maximal little groups and the degrees of basic invariants. We discuss the implications of the observed properties in the minimization problem. In Sec. 5, orbit spaces of two irreps are shown. We find that the generic stratum is semiclosed. It is shown that dimension and concavity of different strata are not sharply distinguished. In Appendix A, we compare our method to the conventional one to help the reader to understand the workings of our method. We also show how to find all the extrema (in the representation space) of a general Higgs potential. If a Higgs potential contains more than four independent invariant polynomials it seems difficult to locate the absolute minimum visually. In Appendix B, we show that the minimization problem can always be reduced to a two-dimensional one in the case of the most general Higgs potential for a single irrep and to a three-dimensional one in the case of an even-degree Higgs potential for two irreps. Thus we can visually minimize the Higgs potentials of these two types using the contours of directional minima we derived previously.

Once the orbit space is constructed, the absolute minimum of the Landau–Higgs potential for a given representation can be read off the list right away using the results derived in Refs. 14–16.

2. HIGGS PROBLEM AND ORBIT PARAMETERS

In a nonabelian gauge theory, where the scalar potential has a symmetry $G \times$reflection and the scalars transform as an n-dimensional irreducible representation $\bar{R}$ of a semi-simple compact Lie group G, the Higgs potential can be written as

$$V(\varphi) = -\frac{1}{2} m^2 \sum_{i=1}^{n} \varphi_i^* \varphi_i + \frac{1}{4} A \left(\sum_{i=1}^{n} \varphi_i^* \varphi_i \right)^2 + \frac{1}{4} A_1 f_{ijkl} \varphi_i^* \varphi_j \varphi_k^* \varphi_l + \frac{1}{4} A_2 g_{ijkl} \varphi_i^* \varphi_j \varphi_k^* \varphi_l + \cdots . \quad (1)$$

$V(\varphi)$ is invariant under a group transformation, $\varphi_j' = \Sigma_{i=1}^n T(\vartheta)_{ji} \varphi_i$. $T(\vartheta)$ is an $n \times n$ matrix representing a group element.[34] It can be written in general, $T(\vartheta) = \exp(-i \Sigma_{i=1}^N \vartheta_L X_L)$. X_L are $n \times n$ matrices representing the generators of the group, and ϑ_L are real or complex parameters specifying an element of the group. Our objective is to find the field configuration and the corresponding symmetry that yield the minimum energy. We set the scalar fields constant in space-time and minimize the resulting potential.

We introduce some useful group theoretical concepts, nicely explained by O'Raifeartaigh.[11] The orbit of φ_a is defined to be the set of vectors $\varphi^{(a)}$ that can be expressed as $\varphi^{(a)} = T(\vartheta)\varphi_a$ with $T(\vartheta)$ an element of G. The little group of φ_a is defined to be the subgroup G_a' of G that leaves φ_a invariant: $T(\vartheta)\varphi_a = \varphi_a$ for $T(\vartheta) \in G_a' \subset G$. The vectors on an orbit are in one-to-one correspondence with the coset G/G_a'. It can easily be shown that the little group G_b' of any vector φ_b on the orbit of φ_a is conjugate to G_a'. If the $T(\vartheta)$ are unitary, then all the vectors $\varphi^{(a)}$ have the same norm $\varphi_a^* \varphi_a$. In general, there is a continuum of distinct orbits respecting the same little group up to conjugation. The set of all such orbits is called the stratum of the little group. Note that if the little groups of two orbits are distinct, then the orbits are distinct. However, the converse is not true, i.e., if two orbits are distinct, their little groups are not necessarily different.

By definition an invariant polynomial is constant on an orbit and thus is a function of orbits. A classical Higgs potential is a polynomial of some algebraically independent invariant polynomials. When we seek a solution to the Higgs problem, we are actually seeking the orbit that minimizes the potential, and its little group.

However we need to find a better way to specify an orbit, because to an orbit there corresponds a trajectory of vectors in the φ-space. Aronhold[20] realized that invariant polynomials specify orbits, which we adopt for our purposes. One is naturally led to the fact that there are only a finite number of independent invariants because there are only a finite number of real parameters specifying a vector in the representation space. Mathematicians have more to say. Hilbert[19] proved that there exists a set of invariant polynomials $I_a(\varphi)$, called the integrity basis, such that every invariant polynomial $P(\varphi)$ can be expressed as a polynomial of I_a: $P(\varphi) = \bar{P}[I_a(\varphi)]$. The invariants in the integrity basis are not necessarily independent, and indeed, for some representations, called noncoregular representations, there are polynomial identities among them, called syzygies. We will call the complete set of lowest degree independent invariants, "basic invariants." The number $(l+1)$ of basic invariants is different for each different representation. We can visualize an orbit as a point in the $(l+1)$-dimensional space of I_a.

The dimensionless ratios of invariants to the magnitude of the φ vector, for example,

$$\lambda = f_{ijkl} \varphi_i^* \varphi_j \varphi_k^* \varphi_l \Big/ \left(\sum_{i=1}^{n} \varphi_i^* \varphi_i \right)^2 , \quad (2)$$

can be used to specify strata, and yield a powerful tool in the minimization problem. We will call the dimensionless ratios *orbit parameters*. They can be considered as a set of generalized angles containing all the directional information. From the definition we can readily see that their ranges are bounded, and thus they occupy a localized region (called the *orbit space*)[35] in the orbit parameter space, which can be regarded as an l-dimensional vector space.

Our method reduces the minimization problem to one of finding "contours" of directional minima (the minimum of the potential in the direction specified by a set of orbit parameters). The "contour" for the most general Higgs potential of one irrep has been analyzed in Ref. 16. It is a curve somewhat similar to a parabola or a surface made by translating the curve. The absolute minimum of the potential occurs at the most protrudent portions of the projected orbit space boundary, corresponding to higher level little groups. The "contour" for the most general even degree Higgs potential of two irreps has been analyzed in Ref. 15. It is a cone and again the absolute minimum occurs at the most protrudent portions of the projected orbit space boundary. This result yields a powerful method for locating the absolute minimum.

When the orbit space dimension is less than four, one can visually locate the absolute minimum. When the dimension is higher than three, one would compare the potential values at different extremum points and pick the lowest one to find the absolute minimum. In Appendix A, we show, by derivation, that there are only a finite number of extrema (in the representation space) of a general Higgs potential, including the ones corresponding to lower level little groups. The absolute minimum normally occurs at the stratum of one of the maximal or maximaximal little groups because they normally correspond to the most protrudent portions of the orbit space boundary. However, we can still visually locate the potential minimum by projecting the orbit space further. In Appendix B, we show that the minimization problem can be reduced to a two-dimensional one in the case of the most general Higgs potential for a single irrep and to a three-dimensional one in the case of an even-degree Higgs potential for two irreps.

When the potential is not monotonic in orbit parameters, the situation is more complicated. We have shown[17] that, in a nontrivial case, each time an orbit parameter appears in the potential nonmonotonically the problem reduces to the same form on the constraint "surface" introduced by the nonmonotonicity. Thus the absolute minimum is now most likely to occur on the less (by one level) protrudent portions of the orbit space boundary. However, we cannot totally exclude trivial cases where the maximal or maximaximal little groups are still favored.

Deeper knowledge of the orbit space structure is essential to understand the Michel–Radicati and the Gell-Mann–Slansky conjectures on the minimal symmetry breaking principle and to see how the principle (not the conjectures themselves) works in the presence of nonmonotonic orbit parameters. The above conjectures seem to hold most of the time but counterexamples[36,37,17] have already been found recently.

In the following we tabulate orbit spaces of all the coregular representations that admit less than (or equal to) five independent basic invariants. Although the orbit space in its original sense is a uniquely defined mathematical object, there is some arbitrariness in defining it using invariant functions. Mathematicians would say that any $(l+1)$ independent smooth invariants would do the job. Physicists would try to be more specific and to make a definition useful for their own needs, preferably a visual and compact one. Our definition was formed in ignorance: dimensionless ratios of lowest degree independent invariant polynomials in the integrity basis to the unique quadratic invariant. This set itself is not uniquely defined because any linear combinations of the same-degree invariant polynomials are equally qualified. Since the concavity of a geometrical object does not change upon linear transformations of the coordinates, our definition is safe.

3. COMPLETE ORBIT SPACES OF ADJOINT REPRESENTATIONS

As we shall see, the orbit spaces of adjoint representations are prototypes for many other representations. We describe them in detail. Let us briefly review some group theoretical results[38] to set up our notation. For the algebra of order N and rank $(l+1)$ we choose a Cartan–Weyl basis, so that the commutation relations assume the standard form:

$$[H_i,H_j]=0,\quad i,j=1,2,...,(l+1), \tag{3a}$$

$$[H_i,E_{\pm\alpha}]=\pm r_i(\alpha)E_{\pm\alpha'}\quad \alpha=1,2,...,(N-l-1)/2, \tag{3b}$$

$$[E_\alpha,E_{-\alpha}]=\sum_{i=1}^{l+1} r_i(\alpha)H_i, \tag{3c}$$

$$[E_\alpha,E_\beta]=N_{\alpha\beta}E_{\alpha+\beta}, \tag{3d}$$

where $N_{\alpha\beta}\neq 0$ only if $\mathbf{r}(\alpha)+\mathbf{r}(\beta)$ is also a root. The Killing scalar products are

$$(H_i,H_i)=1,\quad (E_\alpha,E_{-\alpha})=1, \tag{4}$$

with all other scalar products being zero. Furthermore, the roots $\mathbf{r}(\alpha)$ satisfy the condition

$$\sum_\alpha r_i(\alpha)r_j(\alpha)=\delta_{ij}. \tag{5}$$

Using the generalized Casimir operators derived by Racah,[39] Gruber and O'Raifeartaigh[40] have derived forms for the Casimir invariants that are more useful in practice. [The field components can be reduced by a group transformation to $(l+1)$ (number of rank) irreducible components which correspond to H_i's in the Cartan–Weyl basis. Utilizing these results, we can readily write down the tractable form of each invariant.] The complete set of invariant polynomials for adjoint representations can be obtained by using the matrix form for the representation vector,

$$\varphi=\sum_{i=1}^{N}\varphi_i X_i, \tag{6}$$

where φ_i is the ith component of φ in vector notation and X_i is the matrix corresponding to the ith generator. Note that X_i can be based on any representation. Using the notation

$$I_m=\mathrm{Tr}\,\varphi^m, \tag{7}$$

we list the complete set of invariant polynomials in Table I along with other useful properties for each classical and exceptional Lie group. The I'_n of SO_{2n} is of a form similar to A_5 in Eq. (39) of Ref. 17.

Using the convention

$$\varphi=\sum_{i=1}^{l+1}\varphi_i H_i\equiv[a_1,a_2,\cdots], \tag{8}$$

where we have defined the square bracket as the diagonal elements of the matrix, we can directly write down the orbit parameters in the following generic form:

$$\alpha_m\equiv\frac{\mathrm{Tr}\,\varphi^m}{(\mathrm{Tr}\,\varphi^2)^{m/2}}, \tag{9a}$$

$$\alpha'_m\equiv\frac{2^m a_1a_2\cdots a_m}{(\mathrm{Tr}\,\varphi^2)^{m/2}}. \tag{9b}$$

A. Groups of rank two

There is only one orbit parameter for the adjoint representation of a group of rank 2, and thus the orbit space is a line.

SU(3)

We choose the vector representation for the basis of the matrices. Then the generic stratum and the orbit parameter are represented as follows:

$U_1\times U_1$:

$$\varphi=[a,b,-a-b],$$
$$\alpha_3=(a^3+b^3-(a+b)^3)/(a^2+b^2+(a+b)^2)^{3/2}. \tag{10}$$

TABLE I. List of Casimir invariants, order and rank of classical and exceptional Lie groups.

Group	Invariants	Order	Rank
SU_{n+1}	$I_2, I_3,...,I_{n+1}$	$n(n+2)$	n
SO_{2n+1}	$I_2, I_4,...,I_{2n}$	$n(2n+1)$	n
Sp_{2n}	$I_2, I_4, ...,I_{2n}$	$n(2n+1)$	n
SO_{2n}	$I_2, I_4, ...,I_{2n-2}, I'_n$	$n(2n-1)$	n
G_2	I_2, I_6	14	2
F_4	I_2, I_6, I_8, I_{12}	52	4
E_6	$I_2, I_5, I_6, I_8, I_9, I_{12}$	78	6
E_7	$I_2, I_6, I_8, I_{10}, I_{12}, I_{14}, I_{18}$	133	7
E_8	$I_2, I_8, I_{12}, I_{14}, I_{18}, I_{20}, I_{24}, I_{30}$	248	8

The stratum of each little group is represented as follows:

$SU_2\times U_1$:

$$3=1[-2]+2[1],\qquad \varphi=[a,a,-2a],\quad \alpha_3=\pm 1/\sqrt{6}. \tag{11}$$

The orbit space consists of two end points corresponding to $[SU_2\times U_1]$ and the interior corresponding to $[U_1\times U_1]$.

SO(5) and Sp(4)

We choose the five-dimensional vector representation for the basis. The generic stratum and the orbit parameter are represented as follows:

$U_1\times U_1$:

$$\varphi=[a,-a,b,-b,0],\qquad \alpha_4=(2a^4+2b^4)/(2a^2+2b^2)^2. \tag{12}$$

The stratum of each little group is represented as follows:

$SO_3\times U_1$:

$$5=1[1]+1[-1]+3[0],\qquad \varphi=[a,-a,0,0,0],\quad \alpha_4=\tfrac{1}{2}; \tag{13}$$

$SU_2\times U_1$:

$$5=1[0]+2[1]+2[-1],\qquad \varphi=[a,-a,a,-a,0],\quad \alpha_4=\tfrac{1}{4}. \tag{14}$$

The orbit space consists of two end points corresponding to $[SO_3\times U_1]$, $[SU_2\times U_1]$ and the interior corresponding to $[U_1\times U_1]$.

SO(4)

Although SO_4 can be considered to be a direct product group $SU_2\times SU_2$, we include it for completeness. We choose the vector representation for the basis. The generic stratum and the orbit parameter are represented as follows:

$U_1\times U_1$:

$$\varphi=[a,-a,b,-b],\qquad \alpha_2'=2^2ab/(2a^2+2b^2). \tag{15}$$

The stratum of each little group is represented as follows:

$SU_2\times U_1$:

$$(2,2)=2[1]+2[-1],\qquad \varphi=[a,-a,a,-a],\quad \alpha_2=\pm 1. \tag{16}$$

The orbit space consists of two end points corresponding to $[SU_2\times U_1]$ and the interior corresponding to $[U_1\times U_1]$.

G(2)

We choose the seven-dimensional representation for the basis. The generic stratum and the orbit parameter are represented as follows:

$U_1\times U_1$:

$$\varphi=[2a,0,-2a,a+b,a-b,-a+b,-a-b],\qquad \alpha_6=\frac{2(2a)^6+2(a+b)^6+2(a-b)^6}{[2(2a)^2+2(a+b)^2+2(a-b)^2]^3}. \tag{17}$$

The stratum of each little group is represented as follows:

$SO_3\times U_1$:

$$7=3[0]+2[1]+2[-1],\quad \varphi=[a,-a,a,-a,0,0,0],\quad \alpha_6=\tfrac{1}{16}; \tag{18}$$

$SU_2\times U_1$:

$$7=1[0]+1[2]+1[-2]+2[1]+2[-1],\quad \varphi=[2a,0,-2a,a,-a,a,-a,],\quad \alpha_6=\tfrac{33}{128}. \tag{19}$$

The orbit space consists of two end points corresponding to $[SO_3\times U_1]$, $[SU_2\times U_1]$ and the interior corresponding to $[U_1\times U_1]$.

B. Groups of rank three

There are two orbit parameters for the adjoint representation of a Lie group of rank 3. The orbit space turns out to be a warped triangle.

SU(4)

We choose the four-dimensional representation for the basis of the matrices. The generic stratum and the orbit parameters are represented as follows:

$U_1\times U_1\times U_1$:

$$\varphi=[a,b,c,-a-b-c],\quad \alpha_3=\frac{a^3+b^3+c^3-(a+b+c)^3}{[a^2+b^2+c^2+(a+b+c)^2]^{3/2}},\quad \alpha_4=\frac{a^4+b^4+c^4+(a+b+c)^4}{[a^2+b^2+c^2+(a+b+c)^2]^2}. \tag{20}$$

The stratum of each little group is represented as follows:

$SU_3\times U_1$:

$$4=1[-3]+3[1],\quad \varphi=[a,a,a,-3a],\quad \alpha_3=\pm 1/\sqrt{3},\ \alpha_4=\tfrac{7}{12}; \tag{21}$$

$SU_2\times SU_2\times U_1$:

$$4=(2,1)[1]+(1,2)[-1],\quad \varphi=[a,a,-a,-a],\quad \alpha_3=0,\ \alpha_4=\tfrac{1}{4}; \tag{22}$$

$SU_2\times U_1\times U_1$:

$$4=1[1,1]+1[1,-1]+2[-1,0],\qquad \varphi=[a,a,b,-2a-b]. \tag{23}$$

The orbit space is shown in Fig. 1. It is a warped triangle. Two cusps $\pm$ **P1** of $[SU_3\times U_1]$ and cusp **P2** of $[SU_2\times SU_2\times U_1]$ are connected by the curve of $[SU_2\times U_1\times U_1]$. The cusps and the curve together form the boundary of the generic stratum of $[U_1\times U_1\times U_1]$ which occupies the interior.

SO(6)

Since SO_6 is isomorphic to SU_4, the orbit spaces of their adjoints are identical up to scale factors and locations. If we choose the vector representation for the basis, the generic stratum and the orbit parameters are represented as follows:

$U_1\times U_1\times U_1$:

$$\varphi=[a,-a,b,-b,c,-c]\,,$$

$$\bar{\alpha}_4=\frac{2a^4+2b^4+2c^4}{(2a^2+2b^2+2c^2)^2}\,, \tag{20'}$$

$$\bar{\alpha}'_3=\frac{2^3abc}{(2a^2+2b^2+2c^2)^{3/2}}\,.$$

The stratum of each little group is represented as follows:

$SU_3\times U_1$:

$$6=3[2]+\bar{3}[-2]\,,$$

$$\varphi=[a,-a,a,-a,a,-a]\,, \tag{21'}$$

$$\bar{\alpha}_4=\tfrac{1}{6},\quad \bar{\alpha}'_3=\pm 4/3\sqrt{6}\,;$$

$SU_2\times SU_2\times U_1$:

$$6=(1,1)[2]+(1,1)[-2]+(2,2)[0]\,,$$

$$\varphi=[a,-a,0,0,0,0]\,, \tag{22'}$$

$$\bar{\alpha}_4=\tfrac{1}{2},\quad \bar{\alpha}'_3=0\,;$$

$SU_2\times U_1\times U_1$:

$$6=1[2,0]+1[-2,0]+2[0,1]+2[0,-1]\,,$$

$$\varphi=[a,-a,a,-a,b,-b]\,. \tag{23'}$$

The orbit space of the SO_6 adjoint is obtained from that of SU_4 by the following substitutions: $\bar{\alpha}_4=-\alpha_4+\frac{3}{4}$ and $\bar{\alpha}'_3=-\alpha_3(4/3\sqrt{2})$.

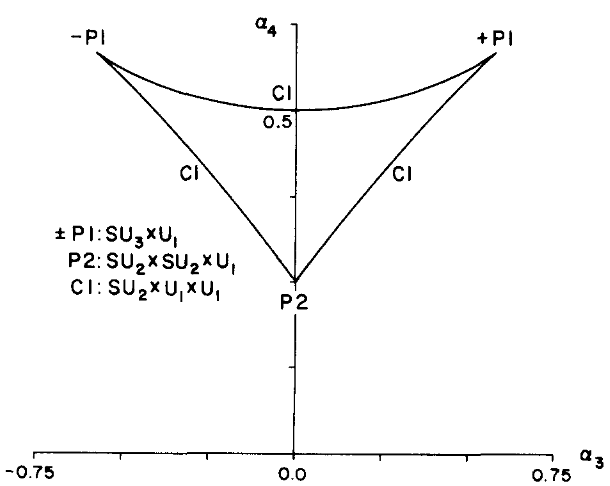


FIG. 1. The orbit space of the SU_4 (SO_6) adjoint representation.

SO(7)

We choose the vector representation for the basis. The generic stratum and the orbit parameters are represented as follows:

$U_1\times U_1\times U_1$:

$$\varphi=[a,-a,b,-b,c,-c,0]\,,$$

$$\alpha_4=\frac{2a^4+2b^4+2c^4}{(2a^2+2b^2+2c^2)^2}\,, \tag{24}$$

$$\alpha_6=\frac{2a^6+2b^6+2c^6}{(2a^2+2b^2+2c^2)^3}\,.$$

The stratum of each little group is represented as follows:

$SO_5\times U_1$:

$$7=1[1]+1[-1]+5[0]\,,$$

$$\varphi=[a,-a,0,0,0,0,0]\,, \tag{25}$$

$$\alpha_4=\tfrac{1}{2},\quad \alpha_6=\tfrac{1}{4}\,;$$

$SU_2\times SO_3\times U_1$:

$$7=(1,3)[0]+(2,1)[1]+(2,1)[-1]\,,$$

$$\varphi=[a,-a,a,-a,0,0,0]\,, \tag{26}$$

$$\alpha_4=\tfrac{1}{4},\quad \alpha_6=\tfrac{1}{16}\,;$$

$SU_3\times U_1$:

$$7=1[0]+3[1]+\bar{3}[-1]\,,$$

$$\varphi=[a,-a,a,-a,a,-a,0]\,, \tag{27}$$

$$\alpha_4=\tfrac{1}{6},\quad \alpha_6=\tfrac{1}{36}\,;$$

$SU_2\times U_1\times U_1$:

$$7=1[0,0]+1[0,1]+1[0,-1]+2[1,0]+2[-1,0]\,,$$

$$\varphi=[a,-a,a,-a,b,-b,0]\,; \tag{28}$$

$SO_3\times U_1\times U_1$:

$$7=1[1,1]+1[1,-1]+1[-1,1]+1[-1,-1]+3[0,0]\,, \tag{29}$$

$$\varphi=[a,-a,b,-b,0,0,0]\,.$$

The orbit space is shown in Fig. 2. It is again a warped triangle. Cusp **P1** of $[SO_5\times U_1]$ and cusp **P2** of $[SU_2\times SO_3\times U_1]$ are connected by straight line **L1** of $[SO_3\times U_1\times U_1]$. All three cusps including cusp **P3** of

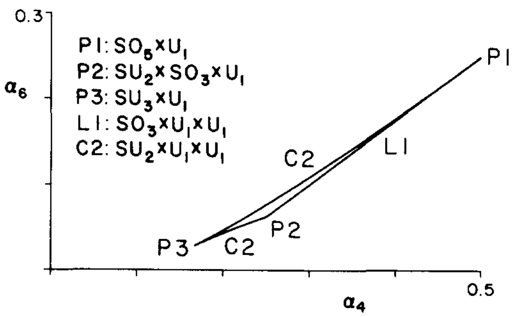


FIG. 2. The orbit space of the SO_7 adjoint representation.

$[SU_3 \times U_1]$ are connected by curve C2 of $[SU_2 \times U_1 \times U_1]$. All the cusps and L1 and C2 together form the boundary of the generic stratum $[U_1 \times U_1 \times U_1]$ which occupies the interior.

Sp(6)

We choose the vector representation for the basis. The generic stratum and the orbit parameters are represented as follows:

$$
\begin{aligned}
& U_1 \times U_1 \times U_1: \\
& \varphi = [a, -a, b, -b, c, -c]\,, \\
& \alpha_4 = \frac{2a^4 + 2b^4 + 2c^4}{(2a^2 + 2b^2 + 2c^2)^2}\,, \\
& \alpha_6 = \frac{2a^6 + 2b^6 + 2c^6}{(2a^2 + 2b^2 + 2c^2)^3}\,.
\end{aligned} \tag{30}
$$

The stratum of each little group is represented as follows:

$$
\begin{aligned}
& Sp_4 \times U_1: \\
& 6 = 1[1] + 1[-1] + 4[0]\,, \\
& \varphi = [a, -a, 0, 0, 0, 0]\,, \\
& \alpha_4 = \tfrac{1}{2}, \quad \alpha_6 = \tfrac{1}{4}\,;
\end{aligned} \tag{31}
$$

$$
\begin{aligned}
& SU_2 \times SU_2 \times U_1: \\
& 6 = (2,1)[0] + (1,2)[1] + (1,2)[-1]\,, \\
& \varphi = [a, -a, a, -a, 0, 0]\,, \\
& \alpha_4 = \tfrac{1}{4}, \quad \alpha_6 = \tfrac{1}{16}\,;
\end{aligned} \tag{32}
$$

$$
\begin{aligned}
& SU_3 \times U_1: \\
& 6 = 3[1] + \bar{3}[-1]\,, \\
& \varphi = [a, -a, a, -a, a, -a]\,, \\
& \alpha_4 = \tfrac{1}{6}, \quad \alpha_6 = \tfrac{1}{36}\,;
\end{aligned} \tag{33}
$$

$$
\begin{aligned}
& SU_2 \times U_1 \times U_1(A): \\
& 6 = 1[0,1] + 1[0,-1] + 2[1,0] + 2[-1,0]\,, \\
& \varphi = [a, -a, a, -a, b, -b]\,.
\end{aligned} \tag{34}
$$

$$
\begin{aligned}
& SU_2 \times U_1 \times U_1(B): \\
& 6 = 1[1,1] + 1[1,-1] + 1[-1,1] \\
& \qquad + 1[-1,-1] + 2[0,0]\,, \\
& \varphi = [a, -a, b, -b, 0, 0]\,.
\end{aligned} \tag{35}
$$

The orbit space is shown in Fig. 3. As we can see from Figs. 2 and 3 the orbit space of the Sp_6 adjoint is identical to that of the SO_7 adjoint. This identity persists between the Sp_{2n} adjoint and the SO_{2n+1} adjoint for any n because the orbit parameters are identically defined. Only the labeling of the little groups is different.

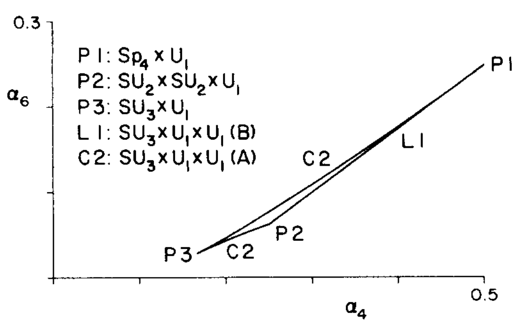


FIG. 3. The orbit space of the Sp_6 adjoint representation.

C. Groups of rank four

There are three orbit parameters for the adjoint representation of a Lie group of rank 4. The orbit space turns out to be a warped tetrahedron.

SU(5)

We choose the vector representation for the basis of the matrices. The generic stratum and the orbit parameters are represented as follows:

$$
\begin{aligned}
& U_1 \times U_1 \times U_1 \times U_1: \\
& \varphi = [a, b, c, d, -a-b-c-d]\,, \\
& \alpha_3 = \frac{a^3 + b^3 + c^3 + d^3 - (a+b+c+d)^3}{[a^2 + b^2 + c^2 + d^2 + (a+b+c+d)^2]^{3/2}}\,, \\
& \alpha_4 = \frac{a^4 + b^4 + c^4 + d^4 + (a+b+c+d)^4}{[a^2 + b^2 + c^2 + d^2 + (a+b+c+d)^2]^2}\,, \\
& \alpha_5 = \frac{a^5 + b^5 + c^5 + d^5 - (a+b+c+d)^5}{[a^2 + b^2 + c^2 + d^2 + (a+b+c+d)^2]^{5/2}}\,.
\end{aligned} \tag{36}
$$

The stratum of each little group is represented as follows:

$$
\begin{aligned}
& SU_4 \times U_1: \\
& 5 = 1[-4] + 4[1]\,, \\
& \varphi = [a, a, a, a, -4a]\,, \\
& \alpha_3 = \pm \frac{3}{2\sqrt{5}}, \quad \alpha_4 = \frac{13}{20}, \quad \alpha_5 = \pm \frac{51}{40\sqrt{5}}\,;
\end{aligned} \tag{37}
$$

$$
\begin{aligned}
& SU_3 \times SU_2 \times U_1: \\
& 5 = (3,1)[2] + (1,2)[-3]\,, \\
& \varphi = [2a, 2a, 2a, -3a, -3a]\,, \\
& \alpha_3 = \pm \frac{1}{\sqrt{30}}, \quad \alpha_4 = \frac{7}{30}, \quad \alpha_5 = \pm \frac{13}{30\sqrt{30}}\,;
\end{aligned} \tag{38}
$$

$$
\begin{aligned}
& SU_3 \times U_1 \times U_1: \\
& 5 = 1[0,1] + 1[-3,-1] + 3[1,0]\,, \\
& \varphi = [a, a, a, b, -3a-b]\,;
\end{aligned} \tag{39}
$$

$$
\begin{aligned}
& SU_2 \times SU_2 \times U_1 \times U_1: \\
& 5 = (1,1)[-2,-2] + (1,2)[1,0] + (2,1)[0,1]\,, \\
& \varphi = [a, a, b, b, -2a-2b]\,;
\end{aligned} \tag{40}
$$

$$
\begin{aligned}
& SU_2 \times U_1 \times U_1 \times U_1: \\
& 5 = 1[0,1,0] + 1[0,0,1] + 1[-2,-1,-1] \\
& \qquad + 2[1,0,0]\,, \\
& \varphi = [a, a, b, c, -2a-b-c]\,.
\end{aligned} \tag{41}
$$

The orbit space is shown in Fig. 4. It is a thin warped tetrahedron. Cusps $\pm$**P1** of $[SU_4 \times U_1]$ and cusps $\pm$**P2** of $[SU_3 \times SU_2 \times U_1]$ are connected by both curves **C1** of $[SU_3 \times U_1 \times U_1]$ and curves **C2** of $[SU_2 \times SU_2 \times U_1 \times U_1]$.

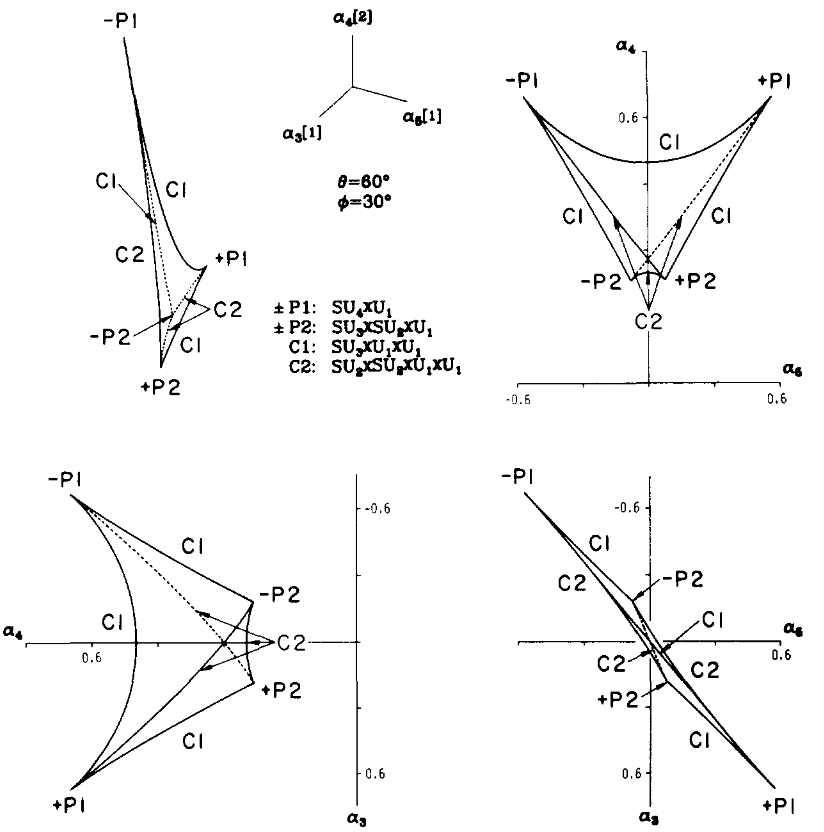


FIG. 4. The complete orbit space of the SU_5 adjoint representation. Shown at the upper left corner is a view from the direction oriented 30° from the α_3 axis and 60° from the α_4 axis. The numbers in the square brackets are the relative ratios of scale. Each projection is a view from the positive direction of the axis not shown in the picture. The dotted curves represent edges on the back (hidden) side of the orbit space.

The two curves lie on the warped surfaces of $[SU_2\times U_1\times U_1\times U_1]$. All these cusps, curves, and surfaces together form the boundary of the generic stratum $[U_1\times U_1\times U_1\times U_1]$ which occupies the interior.

The curves are all concave. One of the principal curvatures of each surface is zero (the surface is flat in this direction) and the other is negative (the surface is concave in this direction).

SO(9)

We choose the vector representation for the basis. The generic stratum and the orbit parameters are represented as follows:

$$
\begin{aligned}
&U_1\times U_1\times U_1\times U_1:\\
&\varphi=[a,-a,b,-b,c,-c,d,-d,0]\,,\\
&\alpha_4=\frac{2a^4+2b^4+2c^4+2d^4}{(2a^2+2b^2+2c^2+2d^2)^2}\,,\\
&\alpha_6=\frac{2a^6+2b^6+2c^6+2d^6}{(2a^2+2b^2+2c^2+2d^2)^3}\,,\\
&\alpha_8=\frac{2a^8+2b^8+2c^8+2d^8}{(2a^2+2b^2+2c^2+2d^2)^4}\,.
\end{aligned}
\tag{42}
$$

The stratum of each little group is represented as follows:

$$
\begin{aligned}
&SO_7\times U_1:\\
&9=1[1]+1[-1]+7[0]\,,\\
&\varphi=[a,-a,0,0,0,0,0,0,0]\,,\\
&\alpha_4=\tfrac{1}{2},\quad \alpha_6=\tfrac{1}{4},\quad \alpha_8=\tfrac{1}{8};
\end{aligned}
\tag{43}
$$

$$
\begin{aligned}
&SO_5\times SU_2\times U_1:\\
&9=(5,1)[0]+(1,2)[1]+(1,2)[-1]\,,\\
&\varphi=[a,-a,a,-a,0,0,0,0,0]\,,\\
&\alpha_4=\tfrac{1}{4},\quad \alpha_6=\tfrac{1}{16},\quad \alpha_8=\tfrac{1}{64};
\end{aligned}
\tag{44}
$$

$$
\begin{aligned}
&SU_3\times SU_2\times U_1:\\
&9=(3,1)[1]+(\bar{3},1)[-1]+(1,3)[0]\,,\\
&\varphi=[a,-a,a,-a,a,-a,0,0,0]\,,\\
&\alpha_4=\tfrac{1}{6},\quad \alpha_6=\tfrac{1}{36},\quad \alpha_8=\tfrac{1}{216};
\end{aligned}
\tag{45}
$$

$$
\begin{aligned}
&SU_4\times U_1:\\
&9=1[0]+4[1]+\bar{4}[-1]\,,\\
&\varphi=[a,-a,a,-a,a,-a,a,-a,0]\,,\\
&\alpha_4=\tfrac{1}{8},\quad \alpha_6=\tfrac{1}{64},\quad \alpha_8=\tfrac{1}{512};
\end{aligned}
\tag{46}
$$

$$
\begin{aligned}
&SU_3\times U_1\times U_1:\\
&9=1[0,0]+1[0,1]+1[0,-1]\\
&\qquad+3[1,0]+\bar{3}[-1,0]\,,\\
&\varphi=[a,-a,a,-a,a,-a,b,-b,0]\,;
\end{aligned}
\tag{47}
$$

$$
\begin{aligned}
&SU_2\times SU_2\times U_1\times U_1:\\
&9=(1,1)[0,0]+(2,1)[1,0]+(2,1)[-1,0]\\
&\qquad+(1,2)[0,1]+(1,2)[0,-1]\,,\\
&\varphi=[a,-a,a,-a,b,-b,b,-b,0]\,;
\end{aligned}
\tag{48}
$$

$$
\begin{aligned}
&SU_2\times SO_3\times U_1\times U_1:\\
&9=(1,1)[0,1]+(1,1)[0,-1]+(2,1)[1,0]\\
&\qquad+(2,1)[-1,0]+(1,3)[0,0]\,,\\
&\varphi=[a,-a,a,-a,b,-b,0,0,0]\,;
\end{aligned}
\tag{49}
$$

$SO_5 \times U_1 \times U_1$:

$$9 = 1[1,0] + 1[-1,0] + 1[0,1] + 1[0,-1] + 5[0,0]\,, \tag{50}$$
$$\varphi = [a, -a, b, -b, 0,0,0,0,0]\,.$$

$SU_2 \times U_1 \times U_1 \times U_1(A)$:

$$9 = 1[0,0,0] + 1[0,1,0] + 1[0,-1,0] + 1[0,0,1] + 1[0,0,-1] + 2[1,0,0] + 2[-1,0,0]\,, \tag{51}$$
$$\varphi = [a, -a, a, -a, b, -b, c, -c, 0]\,;$$

$SU_2 \times U_1 \times U_1 \times U_1(B)$:

$$9 = 1[1,0,0] + 1[-1,0,0] + 1[0,1,0] + 1[0,-1,0] + 1[0,0,1] + 1[0,0,-1] + 3[0,0,0]\,, \tag{52}$$
$$\varphi = [a, -a, b, -b, c, -c, 0,0,0]\,.$$

The orbit space is shown in Fig. 5. It is a thin and sharp tetrahedron. Cusp **P1** of $[SO_7 \times U_1]$ and cusp **P2** of $[SO_5 \times SU_2 \times U_1]$ are connected by curve **C1** of $[SO_5 \times U_1 \times U_1]$. Cusp **P3** of $[SU_3 \times SU_2 \times U_1]$ and cusp **P4** of $[SU_4 \times U_1]$ are connected by curve **C2** of $[SU_3 \times U_1 \times U_1]$ which connects also **P1** and **P4**. **P2** and **P4** are connected by curve **C3** of $[SU_2 \times SU_2 \times U_1 \times U_1]$. **P1**, **P2**, and **P3** are connected by curve **C4** of $[SU_2 \times SO_3 \times U_1 \times U_1]$. The stratum of $[SU_2 \times U_1 \times U_1 \times U_1(B)]$ occupies the warped triangular surface **P1–P2–P3** bounded by **C1** and **C4**. The stratum of $[SU_2 \times U_1 \times U_1 \times U_1(A)]$ closes the rest of the boundary of the generic stratum $[U_1 \times U_1 \times U_1 \times U_1]$ which occupies the interior.

Curves **C1** and **C3** are convex plane-curves and curves **C2** and **C4** are concave space-curves. Surface **P1–P2–P3** is convex along its length but it meets with a $\alpha_4 =$ const plane along a straight line. All the other surfaces meet with a $\alpha_4 =$ const plane along concave curves. Surface **P1–P3–P4** is totally concave. Each of the surfaces **P2–P3–P4** and **P1–P2–P4** have two principal curvatures of opposite sign, i.e., the surfaces are saddle-shaped.

Sp(8)

We choose the vector representation for the basis. The generic stratum and the orbit parameters are represented as follows:

$U_1 \times U_1 \times U_1 \times U_1$:

$$\varphi = [a, -a, b, -b, c, -c, d, -d]\,,$$
$$\alpha_4 = \frac{2a^4 + 2b^4 + 2c^4 + 2d^4}{(2a^2 + 2b^2 + 2c^2 + 2d^2)^2}\,,$$
$$\alpha_6 = \frac{2a^6 + 2b^6 + 2c^6 + 2d^6}{(2a^2 + 2b^2 + 2c^2 + 2d^2)^3}\,, \tag{53}$$
$$\alpha_8 = \frac{2a^8 + 2b^8 + 2c^8 + 2d^8}{(2a^2 + 2b^2 + 2c^2 + 2d^2)^4}\,.$$

The stratum of each little group is represented as follows:

$Sp_6 \times U_1$:

$$8 = 1[1] + 1[-1] + 6[0]\,,$$
$$\varphi = [a, -a, 0,0,0,0,0,0]\,, \tag{54}$$
$$\alpha_4 = \tfrac{1}{2}, \quad \alpha_6 = \tfrac{1}{4}, \quad \alpha_8 = \tfrac{1}{8}\,;$$

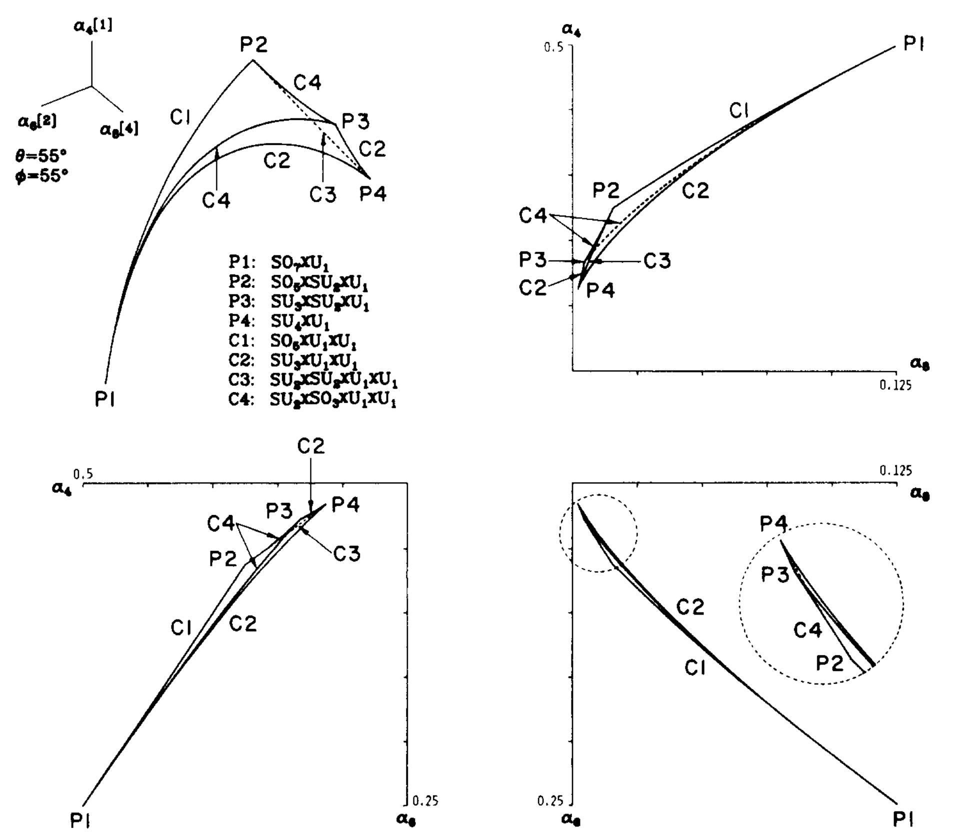


FIG. 5. The complete orbit space of the SO_9 adjoint representation. Shown at the upper left corner is a view from the direction oriented 55° from the α_6 axis and 55° from the α_4 axis. The numbers in the square brackets are the relative ratios of scale. Each projection is a view from the positive direction of the axis not shown in the picture. The dotted curves represent edges on the back (hidden) side of the orbit space.

$Sp_4 \times SU_2 \times U_1$:

$$8 = (4,1)[0] + (1,2)[1] + (1,2)[-1],$$
$$\varphi = [a, -a, a, -a, 0,0,0,0], \tag{55}$$
$$\alpha_4 = \tfrac{1}{4}, \quad \alpha_6 = \tfrac{1}{16}, \quad \alpha_8 = \tfrac{1}{64};$$

$SU_3 \times SU_2 \times U_1$:

$$8 = (3,1)[1] + (\bar{3},1)[-1] + (1,2)[0],$$
$$\varphi = [a, -a, a, -a, a, -a, 0,0], \tag{56}$$
$$\alpha_4 = \tfrac{1}{6}, \quad \alpha_6 = \tfrac{1}{36}, \quad \alpha_8 = \tfrac{1}{216};$$

$SU_4 \times U_1$:

$$8 = 4[1] + \bar{4}[-1],$$
$$\varphi = [a, -a, a, -a, a, -a, a, -a], \tag{57}$$
$$\alpha_4 = \tfrac{1}{8}, \quad \alpha_6 = \tfrac{1}{64}, \quad \alpha_8 = \tfrac{1}{512};$$

$SU_3 \times U_1 \times U_1$:

$$8 = 1[0,1] + 1[0,-1] + 3[1,0] + \bar{3}[-1,0], \tag{58}$$
$$\varphi = [a, -a, a, -a, a, -a, b, -b];$$

$SU_2 \times SU_2 \times U_1 \times U_1(A)$:

$$8 = (1,2)[1,0] + (1,2)[-1,0] + (2,1)[0,1] + (2,1)[0,-1], \tag{59}$$
$$\varphi = [a, -a, a, -a, b, -b, b, -b];$$

$SU_2 \times SU_2 \times U_1 \times U_1(B)$:

$$8 = (1,1)[0,1] + (1,1)[0,-1] + (1,2)[0,0] + (2,1)[1,0] + (2,1)[-1,0], \tag{60}$$
$$\varphi = [a, -a, a, -a, b, -b, 0,0];$$

$Sp_4 \times U_1 \times U_1$:

$$8 = 1[1,0] + 1[-1,0] + 1[0,1] + 1[0,-1] + 4[0,0], \tag{61}$$
$$\varphi = [a, -a, b, -b, 0,0,0,0];$$

$SU_2 \times U_1 \times U_1 \times U_1(A)$:

$$8 = [0,1,0] + 1[0,-1,0] + 1[0,0,1] + 1[0,0,-1] + 2[1,0,0] + 2[-1,0,0], \tag{62}$$
$$\varphi = [a, -a, a, -a, b, -b, c, -c];$$

$SU_2 \times U_1 \times U_1 \times U_1(B)$:

$$8 = 1[1,0,0] + 1[-1,0,0] + 1[0,1,0] + 1[0,-1,0] + 1[0,0,1] + 1[0,0,-1] + 2[0,0,0], \tag{63}$$
$$\varphi = [a, -a, b, -b, c, -c, 0,0].$$

The orbit space of the Sp_8 adjoint is identical to the SO_9 case except for the labeling of the little groups.

SO(8)

We choose the vector representation for the basis. The generic stratum and the orbit parameters are represented as follows:

$U_1 \times U_1 \times U_1 \times U_1$:

$$\varphi = [a, -a, b, -b, c, -c, d, -d],$$
$$\alpha_4 = \frac{2a^4 + 2b^4 + 2c^4 + 2d^4}{(2a^2 + 2b^2 + 2c^2 + 2d^2)^2}, \tag{64}$$
$$\alpha_6 = \frac{2a^6 + 2b^6 + 2c^6 + 2d^6}{(2a^2 + 2b^2 + 2c^2 + 2d^2)^3},$$
$$\alpha_4' = \frac{2^4 abcd}{(2a^2 + 2b^2 + 2c^2 + 2d^2)^2}.$$

The stratum of each little group is represented as follows:

$SO_6 \times U_1$:

$$8 = 1[1] + 1[-1] + 6[0],$$
$$\varphi = [a, -a, 0,0,0,0,0,0], \tag{65}$$
$$\alpha_4 = \tfrac{1}{2}, \quad \alpha_6 = \tfrac{1}{4}, \quad \alpha_4' = 0;$$

$SO_4 \times SU_2 \times U_1$:

$$8 = (2,1,1)[1] + (2,1,1)[-1] + (1,2,2)[0],$$
$$\varphi = [a, -a, a, -a, 0,0,0,0], \tag{66}$$
$$\alpha_4 = \tfrac{1}{4}, \quad \alpha_6 = \tfrac{1}{16}, \quad \alpha_4' = 0;$$

$SU_4 \times U_1$:

$$8 = 4[1] + \bar{4}[-1],$$
$$\varphi = [a, -a, a, -a, a, -a, a, -a], \tag{67}$$
$$\alpha_4 = \tfrac{1}{8}, \quad \alpha_6 = \tfrac{1}{64}, \quad \alpha_4' = \pm\tfrac{1}{4};$$

$SU_3 \times U_1 \times U_1$:

$$8 = 1[0,1] + 1[0,-1] + 3[1,0] + \bar{3}[-1,0], \tag{68}$$
$$\varphi = [a, -a, a, -a, a, -a, b, -b];$$

$SU_2 \times SU_2 \times U_1 \times U_1$:

$$8 = (2,1)[1,0] + (2,1)[-1,0] + (1,2)[0,1] + (1,2)[0,-1], \tag{69}$$
$$\varphi = [a, -a, a, -a, b, -b, b, -b];$$

$SO_4 \times U_1 \times U_1$:

$$8 = (1,1)[1,0] + (1,1)[-1,0] + (1,1)[0,1] + (1,1)[0,-1] + (2,2)[0,0], \tag{70}$$
$$\varphi = [a, -a, b, -b, 0,0,0,0];$$

$SU_2 \times U_1 \times U_1 \times U_1$:

$$8 = 1[0,1,0] + 1[0,-1,0] + 1[0,0,1] + 1[0,0,-1] + 2[1,0,0] + 2[-1,0,0], \tag{71}$$
$$\varphi = [a, -a, a, -a, b, -b, c, -c].$$

The orbit space is shown in Fig. 6. It is a warped tetrahedron. Cusp **P1** of $[SO_6 \times U_1]$ and cusp **P2** of $[SO_4 \times SU_2 \times U_1]$ are connected by line **L1** of $[SO_4 \times U_1 \times U_1]$. Cusps $\pm$ **P3** of $[SU_4 \times U_1]$ and **P2** are connected by line **L2** of $[SU_2 \times SU_2 \times U_1 \times U_1]$. **P1** and $\pm$ **P3** are connected by curve **C3** of $[SU_3 \times U_1 \times U_1]$. The stratum of $[SU_2 \times U_1 \times U_1 \times U_1]$ closes the boundary of the generic stratum $[U_1 \times U_1 \times U_1 \times U_1]$ which occupies the interior.

The projected orbit space $\alpha_4 - \alpha_6$ is not closed by the

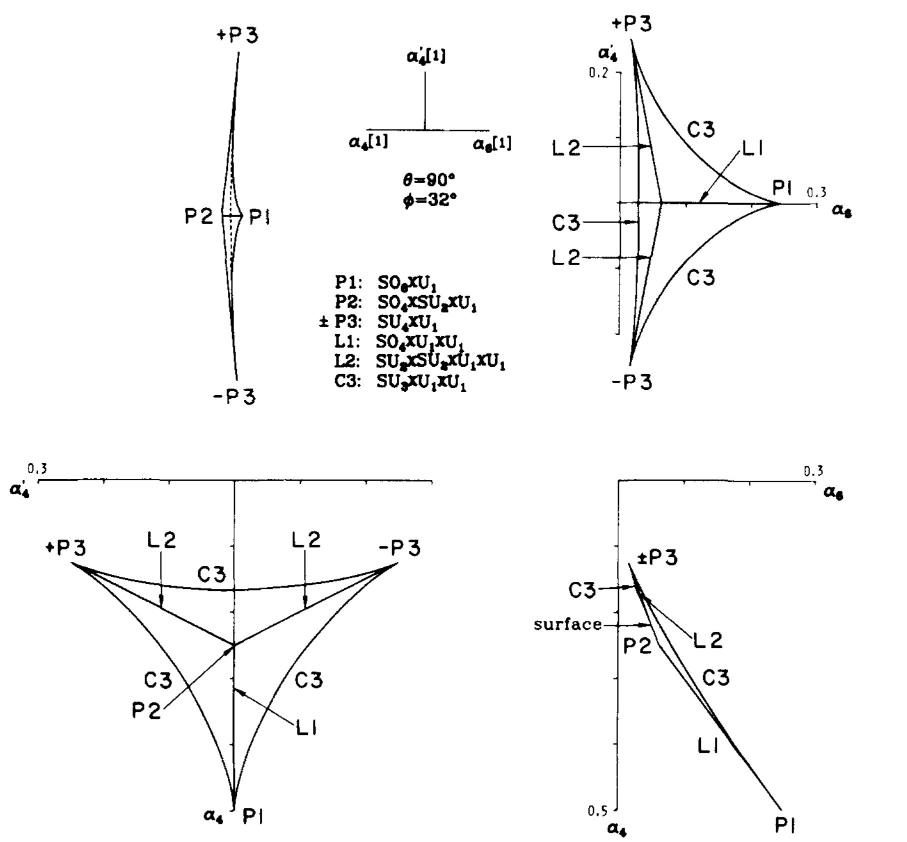


FIG. 6. The complete orbit space of the SO_8 adjoint representation. Shown at the upper left corner is a view from the direction oriented 32° from the α_4 axis and 90° from the α_4' axis. The numbers in the square brackets are the relative ratios of scale. Each projection is a view from the positive direction of the axis not shown in the picture except for the one viewed from the $-\alpha_6$ axis. The dotted curve represents the edge on the back (hidden) side of the orbit space.

one-dimensional strata **L1**, **L2**, and **C3**. The concave punctured portion belongs to the two-dimensional stratum. This is related to the fact that the triangular surface **P2**– + **P3**– – **P3** is convex in the direction + **P3**→ – **P3** but concave in the direction normal to it. All the surfaces that contain cusp **P2** are saddle-shaped. Surface **P1**– + **P3**– – **P3** is totally concave.

F(4)

We choose the 26-dimensional representation for the basis. The generic stratum and the orbit parameters are represented as follows:

$U_1 \times U_1 \times U_1 \times U_1$:

$$\begin{aligned}\varphi = [&2a,0,-2a,2c,0,-2c,b+d,b-d,-b+d,\\ &-b-d,a+b+c,a+b-c,a-b+c,-a+b+c,\\ &-a-b-c,-a-b+c,-a+b-c,a-b-c,\\ &a+c+d,a+c-d,a-c+d,-a+c+d,\\ &-a-c-d,-a-c+d,-a+c-d,a-c-d\,],\end{aligned} \tag{72}$$

$$\alpha_6 = \frac{\sum_{i=1}^{26} \varphi_i^6}{\left[\sum_{i=1}^{26} \varphi_i^2\right]^3}, \quad \alpha_8 = \frac{\sum_{i=1}^{26} \varphi_i^8}{\left[\sum_{i=1}^{26} \varphi_i^2\right]^4}, \quad \alpha_{12} = \frac{\sum_{i=1}^{26} \varphi_i^{12}}{\left[\sum_{i=1}^{26} \varphi_i^2\right]^6}.$$

The stratum of each little group is represented as follows:

$SO_7 \times U_1$:

$$26 = 1[0] + 1[2] + 1[-2] + 7[0] + 8[1] + 8[-1],$$

$$\varphi = [2a, -2a, 8\ a\text{'s}, 8\ (-a)\text{'s}, 8\ 0\text{'s}], \tag{73}$$

$$\alpha_6 = \tfrac{1}{96}, \quad \alpha_8 = \tfrac{11}{6912}, \quad \alpha_{12} = \tfrac{19}{442\,368};$$

$Sp_6 \times U_1$:

$$26 = 6[1] + 6[-1] + 14[0],$$

$$\varphi = [6\,a\text{'s}, 6\ (-a)\text{'s}, 14\ 0\text{'s}], \tag{74}$$

$$\alpha_6 = \tfrac{1}{144}, \quad \alpha_8 = \tfrac{1}{864}, \quad \alpha_{12} = \tfrac{1}{124\,416};$$

$SU_3 \times SU_2 \times U_1(A)$:

$$26 = (8,1)[0] + (3,2)[1] + (3,1)[-2] + (\bar{3},2)[-1] + (\bar{3},1)[2],$$

$$\varphi = [6\ a\text{'s}, 6\ (-a)\text{'s}, 3\ (2a)\text{'s}, 3\ (-2a)\text{'s}, 8\ 0\text{'s}], \tag{75}$$

$$\alpha_6 = 11/1296, \quad \alpha_8 = 43/46\,656, \quad \alpha_{12} = 683/60\,466\,176;$$

$SU_3 \times SU_2 \times U_1(B)$:

$$26 = (1,1)[0] + (1,2)[3] + (1,2)[-3] + (1,3)[0] + (3,1)[-2] + (3,2)[1] + (\bar{3},1)[2] + (\bar{3},2)[-1],$$

$$\varphi = [2\ (3a)\text{'s}, 2\ (-3a)\text{'s}, 6\ a\text{'s}, 6\ (-a)\text{'s}, 3\ (2a)\text{'s}, 3\ (-2a)\text{'s}, 4\ 0\text{'s}], \tag{76}$$

$$\alpha_6 = 23/2592, \quad \alpha_8 = 193/186\,624, \quad \alpha_{12} = 14\,933/967\,458\,816;$$

$SO_5\times U_1\times U_1$:

$$26=1[0,0]+1[2,0]+1[-2,0]+1[0,2]+1[0,-2]+5[0,0]+4[1,1]+4[1,-1]+4[-1,1]+4[-1,-1]\,, \tag{77}$$

$$\varphi=[2a,-2a,2b,-2b,6\ 0\text{'s},\ 4\ (a+b)\text{'s},\ 4\ (a-b)\text{'s},\ 4\ (-a+b)\text{'s},\ 4\ (-a-b)\text{'s}]\,;$$

$SU_2\times SU_2\times U_1\times U_1$:

$$26=(1,1)[2,0]+(1,1)[0,0]+(1,1)[-2,0]+(1,3)[0,0]+(2,1)[0,1]+(2,1)[0,-1]+(1,2)[1,1]+(1,2)[1,-1]+(2,2)[1,0]+(1,2)[-1,1]+(1,2)[-1,-1]+(2,2)[-1,0]\,, \tag{78}$$

$$\varphi=[2a,-2a,b,b,-b,-b,(a+b),(a+b),(a-b),(a-b),(-a+b),(-a+b),(-a-b),(-a-b),4\ a\text{'s},\ 4\ (-a)\text{'s},\ 4\ 0\text{'s}]\,;$$

$SU_3\times U_1\times U_1(A)$:

$$26=3[1,1]+\bar{3}[1,-1]+3[-1,1]+\bar{3}[-1,-1]+3[0,-2]+\bar{3}[0,2]+8[0,0]\,,$$
$$\varphi=[3\ (a+b)\text{'s},\ 3\ (a-b)\text{'s},\ 3\ (-a+b)\text{'s},\ 3\ (-a-b)\text{'s},\ 3\ (2a)\text{'s},\ 3\ (-2a)\text{'s},\ 8\ 0\text{'s}]\,; \tag{79}$$

$SU_3\times U_1\times U_1(B)$:

$$26=1[2,0]+1[0,0]+1[-2,0]+1[0,0]+3[0,2]+\bar{3}[0,-2]+1[1,3]+3[1,-1]+1[1,-3]+\bar{3}[1,1]+1[-1,3]+3[-1,-1]+1[-1,-3]+\bar{3}[-1,1]\,, \tag{80}$$

$$\varphi=[2a,0,-2a,0,3\ (2b)\text{'s},\ 3\ (-2b)\text{'s},\ (a+3b),\ 3\ (a-b)\text{'s},(a-3b),\ 3\ (a+b)\text{'s},\ (-a+3b),\ 3\ (-a-b)\text{'s},(-a-3b),\ 3\ (-a+b)\text{'s}]\,,$$

$SU_2\times U_1\times U_1\times U_1(A)$:

$$26=1[2,0,0]+1[0,0,0]+1[-2,0,0]+3[0,0,0]+1[0,1,1]+1[0,1,-1]+1[0,-1,1]+1[0,-1,-1]+2[1,1,0]+2[1,-1,0]+2[1,0,1]+2[1,0,-1]+2[-1,1,0]+2[-1,-1,0]+2[-1,0,1]+2[-1,0,-1]\,; \tag{81}$$

$SU_2\times U_1\times U_1\times U_1(B)$:

$$26=1[2,0,0]+1[0,0,0]+1[-2,0,0]+1[0,0,2]+1[0,0,0]+1[0,0,-2]+2[0,1,0]+2[0,-1,0]+1[1,1,1]+1[1,1,-1]+1[1,-1,1]+1[1,-1,-1]+2[1,0,1]+2[1,0,-1]+1[-1,1,1]+1[-1,1,-1]+1[-1,-1,1]+1[-1,-1,-1]+2[-1,0,1]+2[-1,0,-1]\,. \tag{82}$$

The unspecified components of φ in Eqs. (81) and (82) can be obtained as follows: in order to get φ_i, multiply the first number in the ith square bracket by a, the second by b, the third by c, and sum all three.

The orbit space is shown in Fig. 7. It is a very thin warped tetrahedron. Cusp **P1** of $[SO_7\times U_1]$ and cusp **P2** of $[Sp_6\times U_1]$ are connected by curve **C1** of $[SO_5\times U_1\times U_1]$. Cusp **P2** and cusp **P3** of $[SU_3\times SU_2\times U_1(A)]$ are connected by curve **C2** of $[SU_3\times U_1\times U_1(A)]$. Cusp **P1** and cusp **P4** of $[SU_3\times SU_2\times U_1(B)]$ are connected by curve **C3** of $[SU_3\times U_1\times U_1(B)]$. All the cusps are connected by curve **C4** of $[SU_2\times SU_2\times U_1\times U_1]$. Surface **S1** (**P1**–**P2**–**P3**) and surface **S2** (**P2**–**P3**–**P4**) belong to the stratum of $[SU_2\times U_1\times U_1\times U_1(A)]$. Surface **S3** (**P1**–**P3**–**P4**) and surface **S4** (**P1**–**P2**–**P4**) belong to the stratum of $[SU_2\times U_1\times U_1\times U_1(B)]$. The interior is occupied by the generic stratum of $[U_1\times U_1\times U_1\times U_1]$.

C2 and **C3** are convex plane-curves. **C1** and the portion of **C4** between **P3** and **P4** are concave space-curves. The other portions of **C4** are convex space-curves. Surface **P1**–**P3**–**P4** is totally concave. All the other surfaces are saddle-shaped.

4. SINGLE IRREDUCIBLE REPRESENTATIONS WITH LOW-DIMENSIONAL ORBIT SPACES

In this section we tabulate orbit spaces of all the single coregular irreducible representations which allow less than (or equal to)[41] five independent basic invariant polynomials. Since all the generators that do not leave a generic orbit invariant are consumed in simplifying the scalar fields through a global gauge transformation, the number I of independent basic invariants is given by

$$I=D-(\dim G-\dim G_g)\,, \tag{83}$$

where D is the dimension of the representation, $\dim G$ the number of generators of the symmetry group G, and $\dim G_g$ the number of generators of the little group of the generic orbit. The little group of the generic orbit is trivial for most irreps. Hsiang and Hsiang[31] listed the nontrivial little groups of generic strata of all the single irreps of compact connected Lie groups.

It is a nontrivial job to construct the invariant polynomials in the integrity basis for a given representation. Although there exists a systematic method[42] for constructing them, it is excessively laborious to build high degree invariant polynomials. However, we do have practical methods

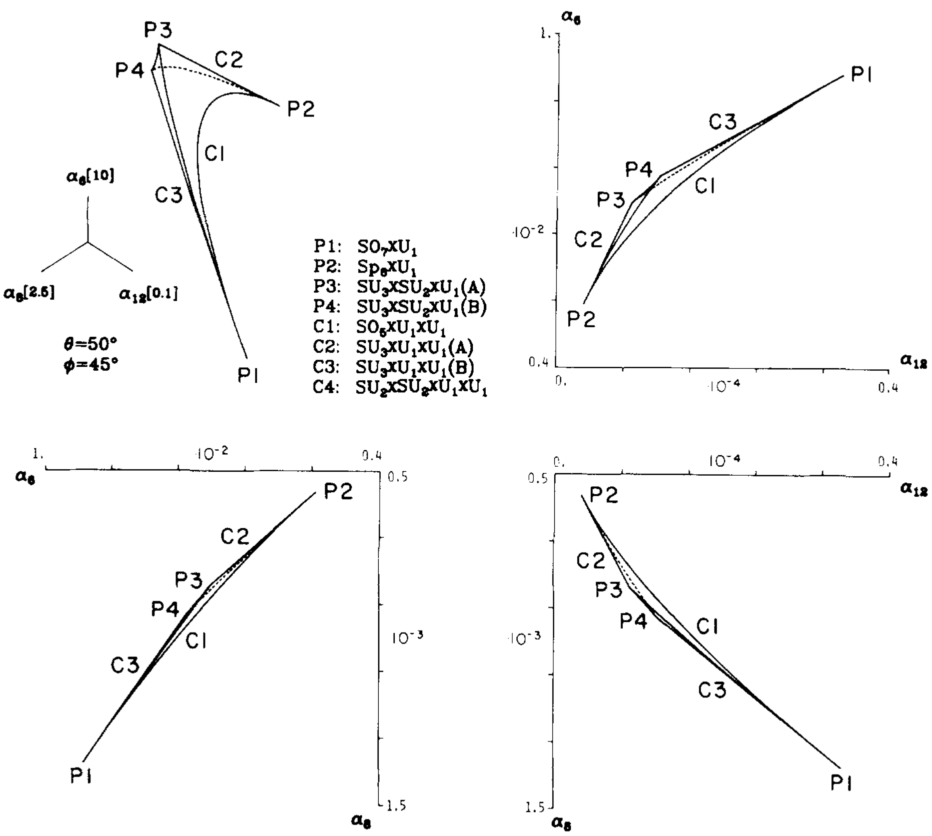


FIG. 7. The complete orbit space of the F_4 adjoint representation. Shown at the upper left corner is a view from the direction oriented 45° from the α_8 axis and 50° from the α_6 axis. The numbers in the square brackets are the relative ratios of scale. Each projection is a view from the positive direction of the axis not shown in the picture. The dotted curves represent edges on the back (hidden) side of the orbit space. The unlabeled curves are all portions of $C4$.

for some representations such as the examples considered in this paper.

Some valuable hints are available in the mathematical literature. Reference 32 lists the degrees and symmetry properties of the polynomials for coregular representations. Noncoregular representations admit polynomial identities (syzygies) among the members of the integrity basis. Patera and Sharp[43] developed a powerful method for finding character generating functions of finite group representations, which can be used for finding the degrees of polynomials in an integrity basis and the degrees of the syzygies among them.

It is convenient to have tables of maximal little groups[44] in carrying out classification of little groups.

A. Symmetric tensor representations

1. Symmetric tensors of SU(N)

Symmetric tensors ψ_{ij} of SU_N can be diagonalized through a group transformation, $\psi'_{ij} = U_{ik}(g)U_{jl}(g)\psi_{kl}$, where $U_{ij}(g)$ is a unitary matrix representing a group element. We abbreviate the diagonal elements as $\psi_{ij} = \mathrm{diag}(\psi_1,\psi_2,\ldots,\psi_N)\exp(i\delta)$ with ψ_i real. Thus there are $N+1$ independent basic invariants. They are given by

$$\begin{aligned} &I_2 = \psi_{ij}\psi^{ij}, \quad I_4 = \psi_{ij}\psi^{jk}\psi_{kl}\psi^{li},\ldots, \\ &I'_N = \epsilon^{ij\cdots kl}\psi_{1i}\psi_{2j}\cdots\psi_{(N-1)k}\psi_{Nl}, \\ &I'^{*}_N = \epsilon_{ij\cdots kl}\psi^{1i}\psi^{2j}\cdots\psi^{(N-1)k}\psi^{Nl}. \end{aligned} \tag{84}$$

We shall see that the cross section of the orbit space of an SU_N symmetric tensor at any phase angle is identical, except for different scale factors, to the orbit space of the SO_{2N} adjoint representation.

SU (3) symmetric tensor $6+\bar{6}$: The generic stratum is invariant under a finite group, $[Z_2\times Z_2]$, (Z_2: a finite group of order 2) and is represented by $\psi_{ij} = \mathrm{diag}(a,b,c)\exp(i\delta)$. Orbit parameters are defined as follows:

$$\begin{aligned} &\alpha_4 = (a^4+b^4+c^4)/(a^2+b^2+c^2)^2, \\ &\alpha'_3 = \exp(3i\delta)abc/(a^2+b^2+c^2)^{3/2}, \\ &\alpha'^{*}_3 = \exp(-3i\delta)abc/(a^2+b^2+c^2)^{3/2}. \end{aligned} \tag{85}$$

Each stratum and its little group are represented as follows:

SO_3: $3=3$:

$$\psi = \mathrm{diag}(a,a,a)\exp(i\delta), \tag{86}$$

$\alpha_4 = \frac{1}{3}$, $\alpha'_3 = \exp(3i\delta)/\sqrt{3}$;

$SU_2\times Z_2$: $3 = 1+2$:

$$\psi = \mathrm{diag}(a,0,0)\exp(i\delta), \tag{87}$$

$\alpha_4 = 1, \quad \alpha'_3 = 0$;

$U_1\times Z_2$:

$$\psi = \mathrm{diag}(a,b,b)\exp(i\delta). \tag{88}$$

The cross section of the orbit space at any angle δ is identical, except for different scale factors and locations, to the orbit space of the SO_6 adjoint representation (Fig. 1) with $[SO_3]$ at $\pm$ **P1**, $[SU_2\times Z_2]$ at **P2**, and $[U_1\times Z_2]$ at **C1**.

SU (4) symmetric tensor $10+\overline{10}$: The generic stratum is invariant under a finite group, $[Z_2\times Z_2\times Z_2]$, and is represented by $\psi_{ij} = \mathrm{diag}(a,b,c,d)\exp(i\delta)$. Orbit parameters are defined as follows:

$$\alpha_4 = (a^4 + b^4 + c^4 + d^4)/(a^2 + b^2 + c^2 + d^2)^2 ,$$
$$\alpha_6 = (a^6 + b^6 + c^6 + d^6)/(a^2 + b^2 + c^2 + d^2)^3 ,$$
$$\alpha_4' = \exp(4i\delta)abcd/(a^2 + b^2 + c^2 + d^2)^2 , \quad (89)$$
$$\alpha_4'^* = \exp(-4i\delta)abcd/(a^2 + b^2 + c^2 + d^2)^2 .$$

Each stratum and its little group are represented as follows:

$SU_3 \times Z_2$:
$$4 = 1 + 3 ,$$
$$\psi = \mathrm{diag}(a,0,0,0)\exp(i\delta) , \quad (90)$$
$$\alpha_4 = 1, \alpha_6 = 1, \alpha_4' = 0 ;$$

$SU_2 \times U_1 \times Z_2$:
$$4 = 1[1] + 1[-1] + 2[0] ,$$
$$\psi = \mathrm{diag}(a,a,0,0)\exp(i\delta) , \quad (91)$$
$$\alpha_4 = \tfrac{1}{2}, \quad \alpha_6 = \tfrac{1}{4}, \quad \alpha_4' = 0 ;$$

$SU_2 \times SU_2$:
$$4 = (2,2),$$
$$\psi = \mathrm{diag}(a,a,a,a)\exp(i\delta) , \quad (92)$$
$$\alpha_4 = \tfrac{1}{4}, \quad \alpha_6 = \tfrac{1}{16}, \quad \alpha_4' = \exp(4i\delta)/16 ;$$

$SU_2 \times Z_2 \times Z_2$:
$$4 = 1 + 1 + 2 ,$$
$$\psi = \mathrm{diag}(a,b,0,0)\exp(i\delta) ; \quad (93)$$

$U_1 \times U_1 \times Z_2$:
$$4 = 1[1,0] + 1[-1,0] + 1[0,1] + 1[0,-1] ,$$
$$\psi = \mathrm{diag}(a,a,b,b)\exp(i\delta) ; \quad (94)$$

$SO_3 \times Z_2$:
$$4 = 1 + 3 ,$$
$$\psi = \mathrm{diag}(a,b,b,b)\exp(i\delta) ; \quad (95)$$

$U_1 \times Z_2 \times Z_2$:
$$4 = 1[0] + 1[0] + 1[1] + 1[-1] ,$$
$$\psi = \mathrm{diag}(a,b,c,c)\exp(i\delta) . \quad (96)$$

The cross section of the orbit space at any angle δ is identical, except for different scale factors, to the orbit space of the SO_8 adjoint representation (Fig. 6) with $[SU_3 \times Z_2]$ at **P1**, $[SU_2 \times U_1 \times Z_2]$ at **P2**, $[SU_2 \times SU_2]$ at $\pm$**P3**, $[SU_2 \times Z_2 \times Z_2]$ at **L1**, $[U_1 \times U_1 \times Z_2]$ at **L2**, $[SO_3 \times Z_2]$ at **C3**, and $[U_1 \times Z_2 \times Z_2]$ on the surfaces.

2. Symmetric traceless tensors of SO(N)

Symmetric traceless tensors ψ_{ij} of SO_N can be diagonalized through a group transformation, $\psi'_{ij} = O_{ik}(g)O_{jl}(g)\psi_{kl}$, where $O_{ij}(g)$ is a real orthogonal matrix representing a group element. We abbreviate the diagonal elements as $\psi_{ij} = \mathrm{diag}(\psi_1,\psi_2,...,\psi_N)$ with ψ_i real. The traceless condition is given by $\delta_{ij}\psi_{ij} = \psi_1 + \psi_2 + \cdots + \psi_N = 0$. Thus there are $N-1$ independent basic invariants. They are given by

$$I_2 = \psi_{ij}\psi_{ij} , \quad I_3 = \psi_{ij}\psi_{jk}\psi_{ki} , \quad I_4 = \psi_{ij}\psi_{jk}\psi_{kl}\psi_{li},\cdots . \quad (97)$$

We immediately see that the orbit space of an SO_N symmetric traceless tensor is identical to that of the SU_N adjoint representation.

SO(3) symmetric traceless tensor 5: The orbit space is identical to that of SU_3 adjoint:

$U_1 \times Z_2$ $[3 = 1(0) + 1(1) + 1(-1)]$, $\alpha_3 = \pm 1/\sqrt{6}$,

$Z_2 \times Z_2$ occupying the interior.

SO(5) symmetric traceless tensor 14: Without going into details we identify various portions of the orbit space (Fig. 4) as follows:

$SU_2 \times SU_2 \times Z_2$ $[5 = (1,1) + (2,2)]$ at $\pm$**P1**,

$SO_3 \times U_1 \times Z_2$ $[5 = 1(2) + 1(-2) + 3(0)]$ at $\pm$**P2**,

$SO_3 \times Z_2 \times Z_2$ $[5 = 1 + 1 + 3]$ at **C1**,

$U_1 \times U_1 \times Z_2 \times Z_2$ at **C2**,

$U_1 \times Z_2 \times Z_2 \times Z_2$ on the surfaces, and

$Z_2 \times Z_2 \times Z_2 \times Z_2$ occupying the interior.

Embedding of each subgroup is indicated by the branching rule given in the square bracket.

3. Symmetric tensors of Sp(2N)

Symmetric tensors ψ_{ij} of Sp_{2N} are adjoint representations.

B. Antisymmetric tensor representations

1. Antisymmetric tensors of SU(N)

Antisymmetric tensors, φ_{ij}, of SU_N can be skew-diagonalized through a group transformation, $\varphi'_{ij} = U_{ik}(g)U_{jl}(g)\varphi_{kl}$. Each diagonal element consists of a real number for odd N and it comes with an overall phase factor for even N. Thus there are $(N-1)/2$ for odd N ($N/2+1$ for even N) independent basic invariants. They are given by

$$I_2 = \varphi_{ij}\varphi^{ij} , \quad I_4 = \varphi_{ij}\varphi^{jk}\varphi_{kl}\varphi^{li},\cdots, \quad \text{for odd } N, \quad (98a)$$
$$I_2 = \varphi_{ij}\varphi^{ij} , \quad I_4 = \varphi_{ij}\varphi^{jk}\varphi_{kl}\varphi^{li},\cdots, \quad (98b)$$
$$I'_{N/2} = \epsilon^{ij\cdots kl}\varphi_{ij}\cdots\varphi_{kl} ,$$
$$I'^*_{N/2} = \epsilon_{ij\cdots kl}\varphi^{ij}\cdots\varphi^{kl} , \quad \text{for even } N .$$

We immediately see that (the cross section at any angle δ of) the orbit space of an SU_N antisymmetric tensor is identical to that of the SO_N adjoint for $N > 4$. In the following we match various portions of each pair of orbit spaces.

SU(5) antisymmetric tensor 10 + $\overline{10}$: The orbit space is identical to that of the SO_5 adjoint, with

$SU_2 \times SU_3$ $[5 = (2,1) + (1,3)]$ replacing $SO_3 \times U_1$,

Sp_4 $[5 = 1 + 4]$ replacing $SU_2 \times U_1$, and

$SU_2 \times SU_2$ $[5 = (1,1) + (2,1) + (1,2)]$ replacing $U_1 \times U_1$.

SU(6) antisymmetric tensor 15 + $\overline{15}$: The cross section of the orbit space at any angle δ is identical to that of the SO_6 adjoint, with

$SP_6\,[6=6]$ at $\pm\mathbf{P}_1$,

$SU_2\times SU_4\,[6=(2,1)+(1,4)]$ at **P2**,

$SU_2\times Sp_4\,[6=(2,1)+(1,4)]$ at **C1**, and

$SU_2\times SU_2\times SU_2\,[6=(2,1,1)+(1,2,1)+(1,1,2)]$ occupying the interior.

SU(7) antisymmetric tensor $21+\overline{21}$: The orbit space is identical to that of the SO_7 adjoint (Fig. 2), with

$SU_2\times SU_5\,[7=(2,1)+(1,5)]$ at **P1**,

$SU_3\times Sp_4\,[7=(3,1)+(1,4)]$ at **P2**,

$Sp_6\,[7=1+6]$ at **P3**,

$SU_3\times SU_2\times SU_2\,[7=(3,1,1)+(1,2,1)+(1,1,2)]$ at **L1**,

$SU_2\times Sp_4\,[7=(1,1)+(2,1)+(1,4)]$ at **C2**, and

$SU_2\times SU_2\times SU_2\,[7=(1,1,1)+(2,1,1)+(1,2,1)+(1,1,2)]$ occupying the interior.

SU(8) antisymmetric tensor $28+\overline{28}$: The cross section of the orbit space at any angle δ is identical to that of the SO_8 adjoint (Fig. 6), with

$SU_2\times SU_6\,[8=(2,1)+(1,6)]$ at **P1**,

$Sp_4\times SU_4\,[8=(4,1)+(1,4)]$ at **P2**,

$Sp_8\,[8=8]$ at $\pm$ **P3**,

$SU_2\times SU_2\times SU_4\,[8=(2,1,1)+(1,2,1)+(1,1,4)]$ at **L1**,

$Sp_4\times Sp_4\,[8=(4,1)+(1,4)]$ at **L2**,

$SU_2\times Sp_6\,[8=(2,1)+(1,6)]$ at **C3**,

$SU_2\times SU_2\times Sp_4\,[8=(2,1,1)+(1,2,1)+(1,1,4)]$ on the surfaces, and

$SU_2\times SU_2\times SU_2\times SU_2\,[8=(2,1,1,1)+(1,2,1,1)+(1,1,2,1)+(1,1,1,2)]$ occupying the interior.

SU(9) antisymmetric tensor $36+\overline{36}$: The orbit space is identical to that of the SO_9 adjoint (Fig. 5), with

$SU_2\times SU_7\,[9=(2,1)+(1,7)]$ at **P1**,

$Sp_4\times SU_5\,[9=(4,1)+(1,5)]$ at **P2**,

$Sp_6\times SU_3\,[9=(6,1)+(1,3)]$ at **P3**,

$Sp_8\,[9=1+8]$ at **P4**,

$SU_2\times SU_2\times SU_5\,[9=(2,1,1)+(1,2,1)+(1,1,5)]$ at **C1**,

$Sp_6\times SU_2\,[9=(6,1)+(1,2)+(1,1)]$ at **C2**,

$Sp_4\times Sp_4\,[9=(4,1)+(1,4)+(1,1)]$ at **C3**,

$Sp_4\times SU_2\times SU_3\,[9=(4,1,1)+(1,2,1)+(1,1,3)]$ at **C4**,

$SU_2\times SU_2\times SU_2\times SU_3\,[9=(2,1,1,1)+(1,2,1,1)+(1,1,2,1)+(1,1,1,3)]$ occupying the warped triangular surface **P1–P2–P3** bounded by **C1** and **C2**,

$Sp_4\times SU_2\times SU_2\,[9=(4,1,1)+(1,2,1)+(1,1,2)]$ closing the rest of the boundary of the generic stratum,

$SU_2\times SU_2\times SU_2\times SU_2\,[9=(1,1,1,1)+(2,1,1,1)+(1,2,1,1)+(1,1,2,1)+(1,1,1,2)]$ occupying the interior.

2. Antisymmetric tensors of SO(N)

Antisymmetric tensors φ_{ij} of SO_N are adjoint representations.

3. Antisymmetric traceless tensors of Sp(2N)

Antisymmetric traceless tensors φ_{ij} of Sp_{2N} are skew-diagonalized through a group transformation, $\varphi'_{ij}=S_{ik}(g)S_{jl}(g)\varphi_{kl}$, where $S_{ij}(g)$ is a simplectic matrix satisfying $S_{ki}S_{lj}f_{kl}=f_{ij}$. We abbreviate the skew-diagonal elements as $\varphi_{ij}=$ skew-diag$(\varphi_1,\varphi_2,\ldots,\varphi_N)$ with φ_i real. In this notation $f_{ij}=$ skew-diag$(1,1,\ldots,1)$. The traceless condition is given by $f_{ij}\varphi_{ij}=2(\varphi_1+\varphi_2+\cdots+\varphi_N)=0$. Thus there are $N-1$ independent basic invariants:

$$I_2=f_{ik}\,f_{jl}\varphi_{ij}\varphi_{kl}\,,\quad I_3=f_{il}\,f_{jm}\,f_{kn}\,\varphi_{im}\varphi_{jn}\varphi_{kl},\cdots. \tag{99}$$

We shall see that the orbit space of an antisymmetric traceless tensor of Sp_{2N} is identical, except for different scale factors, to that of the SU_N adjoint representation.

Sp(6) antisymmetric tensor 14: The generic stratum is invariant under $[Sp_2\times Sp_2\times Sp_2]$. The orbit parameter is defined as follows:

$$\alpha_3=(2a^3+2b^3-2(a+b)^3)/(2a^2+2b^2+2(a+b)^2)^{3/2}. \tag{100}$$

Each stratum and its little group are represented as follows:

$Sp_2\times Sp_4$:

$$6=(2,1)+(1,4),\quad \varphi=\text{skew-diag}(a,a,-2a),\quad \alpha_3=\pm 1/2\sqrt{3}. \tag{101}$$

The orbit space is identical, except for different scale factors, to that of the SU_3 adjoint.

Sp(8) antisymmetric traceless tensor 27: The generic stratum is invariant under $[Sp_2\times Sp_2\times Sp_2\times Sp_2]$. The orbit parameters are defined as follows:

$$\begin{aligned}\alpha_3&=(2a^3+2b^3+2c^3-2(a+b+c)^3)/(2a^2+2b^2+2c^2+2(a+b+c)^2)^{3/2},\\ \alpha_4&=(2a^4+2b^4+2c^4+2(a+b+c)^4)/(2a^2+2b^2+2c^2+2(a+b+c)^2)^2.\end{aligned} \tag{102}$$

Each stratum and its little group are represented as follows:

$Sp_2\times Sp_6$:

$$8=(2,1)+(1,6),\quad \varphi=\text{skew-diag}(a,a,a,-3a),\quad \alpha_3=\pm 1/\sqrt{6},\quad \alpha_4=7/24; \tag{103}$$

$Sp_4\times Sp_4$:

$$8=(4,1)+(1,4),\quad \varphi=\text{skew-diag}(a,a,-a,-a),\quad \alpha_3=0,\quad \alpha_4=1/8; \tag{104}$$

$Sp_2\times Sp_2\times Sp_4$:

$$8=(2,1,1)+(1,2,1)+(1,1,4),\quad \varphi=\text{skew-diag}(a,a,b,-2a-b). \tag{105}$$

The orbit space is identical, except for different scale factors, to that of the SU_4 adjoint (Fig. 1). Identifications are:

$[Sp_2\times Sp_6]$ at $\pm$ **P**1, $[Sp_4\times Sp_4]$ at **P**2, and $[Sp_2\times Sp_2\times Sp_4]$ at **C**1.

Sp(10) antisymmetric traceless tensor 44: The generic stratum is invariant under $[Sp_2\times Sp_2\times Sp_2\times Sp_2\times Sp_2]$. The orbit parameters are defined as follows:

$$\alpha_3=\frac{(2a^3+2b^3+2c^3+2d^3-2(a+b+c+d)^3)}{(2a^2+2b^2+2c^2+2d^2+2(a+b+c+d)^2)^{3/2}},$$
$$\alpha_4=\frac{(2a^4+2b^4+2c^4+2d^4+2(a+b+c+d)^4)}{(2a^2+2b^2+2c^2+2d^2+2(a+b+c+d)^2)^2}, \tag{106}$$
$$\alpha_5=\frac{(2a^5+2b^5+2c^5+2d^5-2(a+b+c+d)^5)}{(2a^2+2b^2+2c^2+2d^2+2(a+b+c+d)^2)^{5/2}}.$$

Each stratum and its little group are represented as follows:

$Sp_2\times Sp_8$:
$$10=(2,1)+(1,8),$$
$$\varphi=\text{skew-diag}(a,a,a,a,-4a), \tag{107}$$
$$\alpha_3=\pm 3/2\sqrt{10},\quad \alpha_4=13/40,\quad \alpha_5=\pm 51/80\sqrt{10};$$

$Sp_4\times Sp_6$:
$$10=(4,1)+(1,6),$$
$$\varphi=\text{skew-diag}(2a,2a,2a,-3a,-3a), \tag{108}$$
$$\alpha_3=\pm 1/2\sqrt{15},\quad \alpha_4=7/60,\quad \alpha_5=\pm 13/120\sqrt{15};$$

$Sp_2\times Sp_2\times Sp_6$:
$$10=(2,1,1)+(1,2,1)+(1,1,6),$$
$$\varphi=\text{skew-diag}(a,a,a,b,-3a-b); \tag{109}$$

$Sp_2\times Sp_4\times Sp_4$:
$$10=(2,1,1)+(1,4,1)+(1,1,4),$$
$$\varphi=\text{skew-diag}(a,a,b,b,-2a-2b); \tag{110}$$

$Sp_2\times Sp_2\times Sp_2\times Sp_4$:
$$10=(2,1,1,1)+(1,2,1,1)+(1,1,2,1)+(1,1,1,4),$$
$$\varphi=\text{skew-diag}(a,a,b,c,-2a-b-c). \tag{111}$$

The orbit space is identical, except for different scale factors, to that of the SU_5 adjoint (Fig. 4). Identifications are: $[Sp_2\times Sp_8]$ at $\pm$ **P**1, $[Sp_4\times Sp_6]$ at $\pm$ **P**2, $[Sp_2\times Sp_2\times Sp_6]$ at **C**1, $[Sp_2\times Sp_4\times Sp_4]$ at **C**2, and $[Sp_2\times Sp_2\times Sp_2\times Sp_4]$ on the surfaces.

C. Other low-dimensional irreducible representations

The remaining irreps that allow less than four-dimensional orbit spaces (or the cross sections at arbitrary phase angles) are the defining representations of various groups, spinor representations of SO_N, and SO_3 representations.

The defining representations of classical Lie groups and G_2 yield single quadratic invariants only and their orbit spaces are trivial. Their little groups are

SU_{N-1} $(N\geqslant 2)$ for $N+\bar N$ of SU_N,

SO_{N-1} $(N\geqslant 3)$ for N of SO_N,

Sp_{2N-2} $(N\geqslant 3)$ for $2N+2N$ of Sp_{2N}, and

SU_3 for 7 of G_2.

The spinor representations of SO_N for low N (<10) also yield single quadratic invariants only. Their little groups are

Sp_2 for $4+4$ of SO_5,

SU_3 for $4+\bar 4$ of SO_6,

G_2 for 8 of SO_7,

SO_7 for 8 of SO_8,

SO_7 for 16 of SO_9.

The spinor representation of SO_{10}, the defining representations of F_4 and E_6, and low-dimensional (less than 8) SO_3 representations yield nontrivial low-dimensional orbit spaces.

Spinor representation of SO(10) 16 + $\overline{16}$

The 16 component complex spinor of SO_{10} is left invariant under SU_4 and is reduced, through an SO_{10} transformation generated by the 30 non-SU_4 generators made out of 45 σ-matrices σ_{ij},[23] to two real components, say, ψ_4 and ψ_6. Two independent basic invariant polynomials exist:

$$I_2=\bar\psi\psi=2(\psi_4^*\psi_4+\psi_6^*\psi_6),$$
$$I_4=\bar\psi\gamma_i\psi\bar\psi\gamma_i\psi=\bar\psi\gamma_5\psi\bar\psi\gamma_5\psi+\bar\psi\gamma_{10}\psi\bar\psi\gamma_{10}\psi \tag{112}$$
$$=16\psi_4^*\psi_4\psi_6^*\psi_6,$$

where we left the complex conjugate intact to show the contraction of ψ's between γ matrices.

Each stratum and its little group are represented as follows:

SU_5:
$$16=1+\bar 5+10, \tag{113}$$
$$\psi_4=a,\quad \psi_6=0,\quad \alpha_4=0;$$

SO_7:
$$16=1+7+8, \tag{114}$$
$$\psi_4=a,\quad \psi_6=a,\quad \alpha_4=1.$$

Defining representation of F(4) 26

The representation spaces of exceptional groups are naturally described on the octonionic basis. We refer the reader to Refs. 45 and 46 for further details.

The 26-dimensional defining representation of F_4 is represented by a 3×3 real, symmetric, and traceless matrix over octonions:

$$\varphi=\begin{pmatrix} a & \gamma & \bar\beta\\ \bar\gamma & b & \alpha\\ \beta & \bar\alpha & c\end{pmatrix}, \tag{115}$$

where a,b,c are real numbers satisfying $a+b+c=0$ and α,β,γ are real octonions (the bar denotes octonionic conjugation). It is left invariant under an SO_8 transformation $(26=1+1+8_v+8_s+8_c)$. The dimension formula (83) yields $2=26-52+28$, which is the number of independent basic invariants. They are given by

$$I_2=\tfrac12\operatorname{Tr}(\varphi^\dagger\cdot\varphi),\quad I_3=\tfrac13\operatorname{Tr}[(\varphi\times\varphi)\cdot\varphi], \tag{116}$$

where the dot represents the Jordan product (half the anticommutator) and $\varphi \times \varphi$ is the Freudenthal product:

$$\varphi \times \varphi = \varphi(\varphi - \mathrm{Tr}\,\varphi) - \tfrac{1}{2}\mathrm{Tr}[\varphi(\varphi - \mathrm{Tr}\,\varphi)] .$$

An F_4 transformation involving those generators not included in SO_8 reduces φ to a diagonal matrix containing two real parameters:

$$\varphi = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & -(a+b) \end{pmatrix} \equiv [a,b,-(a+b)] . \quad (117)$$

Written in terms of these two real parameters, the invariant polynomials are

$$I_2 = \tfrac{1}{2}(a^2 + b^2 + (a+b)^2), \quad I_3 = \pm ab(a+b) . \quad (118)$$

Each stratum and its little group are represented as follows:

SO_9:

$$26 = 1 + 9 + 16 ,$$
$$\psi = [a,a,-2a] , \quad \alpha_3 = \pm 2/3\sqrt{3} . \quad (119)$$

Defining representation of E(6) 27 + $\overline{27}$

The 27-dimensional complex defining representation of E_6 is represented by a 3×3 complex, Hermitian, octonionic matrix

$$\varphi = \begin{pmatrix} a & \gamma & \bar{\beta} \\ \bar{\gamma} & b & \alpha \\ \beta & \bar{\alpha} & c \end{pmatrix}, \quad (120)$$

where a,b,c are complex numbers and α, β, γ are complex octonions. It is left invariant under an SO_8 transformation $(27 = 1 + 1 + 1 + 8_v + 8_s + 8_c)$. The dimension formula (83) yields $4 = 27 + 27 - 78 + 28$, which is the number of independent basic invariants. They are given by[46]

$$I_2 = \tfrac{1}{2}\mathrm{Tr}(\varphi^\dagger \cdot \varphi), \quad I_4 = \tfrac{1}{2}\mathrm{Tr}[(\varphi\times\varphi)^\dagger \cdot (\varphi\times\varphi)] , \quad (121)$$
$$I_3 = \tfrac{1}{3}\mathrm{Tr}[(\varphi\times\varphi)\cdot\varphi] , \quad I_3^* = \text{complex conjugate of } I_3 .$$

An E_6 transformation involving those generators not included in SO_8 reduces φ to a diagonal matrix containing four real parameters:

$$\varphi = \exp(i\delta)\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \equiv \exp(i\delta)[a,b,c] . \quad (122)$$

Written in terms of these four real parameters the invariant polynomials are

$$I_2 = \tfrac{1}{2}(a^2 + b^2 + c^2), \quad I_4 = \tfrac{1}{2}(a^2b^2 + b^2c^2 + c^2a^2),$$
$$I_3 = \exp(3i\delta)abc . \quad (123)$$

Each stratum and its little group are represented as follows:

F_4:

$$27 = 1 + 26 ,$$
$$\varphi = \exp(i\delta)[a,a,a] , \quad (124)$$
$$\alpha_4 = \tfrac{2}{3}, \quad \alpha_3 = \exp(3i\delta)(2/3)^{3/2} ;$$

SO_{10}:

$$27 = 1 + 10 + 16 ,$$
$$\varphi = \exp(i\delta)[a,0,0] , \quad (125)$$
$$\alpha_4 = 0, \quad \alpha_3 = 0 ;$$

SO_9:

$$27 = 1 + 1 + 9 + 8 + 8 ,$$
$$\varphi = \exp(i\delta)[a,b,b] . \quad (126)$$

The cross section of the orbit space at an arbitrary phase angle δ is shown in Fig. 8. It is again a warped triangle.

The generic strata of SU_2 representations 4 + 4 and 6 + 6 have trivial little groups and thus their orbit spaces are four- and eight-dimensional. The SO_3 seven-dimensional representation (totally symmetric traceless third-rank tensor) is non-coregular and has five invariants of degree 2, 4, 6, 10, and 15 in the integrity basis and a syzygy. They are listed in Ref. 47. Its maximal little groups[12,48] are U_1, T, D_3, and D_2.

D. Comments

Our observations for single irreps are summarized as follows:

(1) The orbit spaces for the adjoint representations of Lie groups of the same rank all have similar geometrical shapes, namely, straight line for groups of rank two, triangle for groups of rank three, tetrahedron for groups of rank four, and so on. (This pattern was evident in the examples of Ref. 16.)

This implies that there is an interesting relationship between the degrees of polynomial invariants, the number of maximal little groups, and the shape of the orbit space. For example, the SU_5 adjoint has only two maximal little groups but odd degree invariants such as I_3 and I_5 duplicate the cusps providing the third and fourth cusps needed to build a tetrahedron. For the adjoint representations of all the other groups of rank four there are four maximal little groups and their invariants are of even degree yielding only four cusps, again just enough to build a tetrahedron.

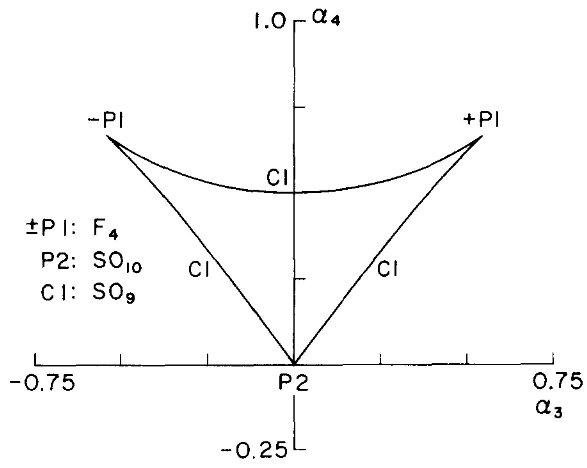


FIG. 8. The orbit space of 27 + $\overline{27}$ of E_6.

However, this relationship does not seem to hold in the case of non-coregular representations. The SO_3 seven-dimensional representation has four maximal little groups. If its orbit space is built out of polynomials of degree 2, 4, 6, and 10, then it is a warped tetrahedron. If it is built out of 2nd, 4th, 6th, and 15th degree polynomials, then some cusps may be duplicated and the orbit space may even become a warped octahedron. It will be interesting to see if the orbit space built out of all five polynomials and the syzygy has indeed a tetragonal shape.

(2) The orbit spaces (or the cross sections at an arbitrary phase angle) of symmetric and antisymmetric tensor representations are identical, up to scale factors, to those of adjoint representations.

Gürsey suggested that this similarity among the orbit spaces of adjoints, symmetric and antisymmetric second-rank tensors results from deeper mathematical roots.[49] These representations all have definite exchange symmetries among the tensor indices. Even the $27 + \overline{27}$ of E_6 has such symmetry: it is a Hermitian matrix (over octonions).

Michel[12] pointed out that the orbit spaces of the vector representations of Weyl groups are, up to scale factors, identical to those of corresponding adjoint representations. We quote two examples given in Ref. 50. The orbit spaces for the vector representations of the tetrahedral groups T and T_d are identical to that of SO_6 adjoint. The orbit spaces for the vector representations of the tetrahedral group T_h and the octahedral groups O and O_h is identical to that of SO_7 adjoint.

(3) Lower-dimensional strata of higher symmetries form the boundaries of higher-dimensional strata of lower symmetries in an orderly way. The hierarchy of protrusiveness on the orbit space boundary is not a global property (a poorly defined concept in any case) but a local property, which is shown by the saddle-shaped surfaces in most three-dimensional orbit spaces.

The last observation is not what we like because it may lead to a counterexample to the minimal symmetry breaking principle. However, none of the cases we have considered makes a counterexample.

In our formalism we take the singlet form for a given subgroup as the definition of a stratum. Its equation is obtained by putting the singlet form into the invariant polynomials and is thus parametric. In order to obtain the singlet form, which is the minimum information needed to specify an extremum point in any case, we have to find the matrix elements of group generators over the given representation and require that the subgroup generators annihilate the representation vector.

It is convenient to have nonparametric equations for strata. Since there are fewer independent parameters than basic invariants for all the strata except for the generic stratum, we should have some identities among the basic invariants on these strata. Like syzygies they are polynomials. It is not easy, though possible in principle, to derive these identities from our parametric equations. Abud and Sartori[29] devised a general method for finding nonparametric equations of orbits. It is a good tractable method usable also for the projected orbit space associated with a Higgs potential. It requires only the knowledge of invariant polynomials. One can obtain the singlet forms (though not the little groups) from the nonparametric orbit equations by solving high degree algebraic equations. This is as difficult as minimizing a Higgs potential using a conventional method. The method outlined in the previous paragraph is the only tractable way for finding the little groups and singlet forms, as far as we know. Jarić[50] devised another elegant nonparametric method for representations of finite groups which can be used for adjoint representations of low rank compact Lie groups. He provided both singlet forms and nonparametric equations for orbits. However, its applicability seems to be limited to only a small number of representations. It will be interesting to see if his method can be extended to more complicated cases.

In a Higgs potential there appear invariant polynomials only up to fourth degree. Thus we deal with a projected orbit space. Due to the projection some cusps are buried inside the projected space, as shown in SO_N and Sp_{2N} examples. These buried cusps cannot yield the absolute minimum, though they correspond to maximal little groups. A similar phenomenon was noticed earlier in the examples of SO_N adjoint + vector representations.[17] As a matter of fact, it was observed much earlier by Li.[23] This implies that, in unification theories, simple-minded classification of possible symmetry-breaking directions based on maximal or maximaximal little groups is not enough. One should check if the symmetry breaking really occurs in the desired direction.

5. TWO IRREDUCIBLE REPRESENTATIONS

The orbit spaces of two irreducible representations are normally high-dimensional because after one of the representations is simplified only a small number of group parameters are left for further simplification of the other representation. We have found two cases where the orbit space is three-dimensional, SU_3 adjoint + vector and SO_5 adjoint + vector.

A. SU(3) adjoint + vector representations

Using the same notation as in Ref. 15, the orbit parameters are

$$\alpha_3 = \frac{\Sigma\varphi_i^3}{(\Sigma\varphi_i^2)^{3/2}}, \tag{127}$$

$$\beta_1 = \frac{\Sigma\chi_i^*\varphi_i\chi_i}{(\Sigma\varphi_i^2)^{1/2}(\Sigma\chi_i^*\chi_i)}, \quad \beta_2 = \frac{\Sigma\chi_i^*\varphi_i^2\chi_i}{(\Sigma\varphi_i^2)(\Sigma\chi_i^*\chi_i)}. \tag{128}$$

The stratum of each little group is represented as follows:

SU_2:

$$\begin{aligned} &8 = 1 + 2 + 2 + 3, \quad 3 = 1 + 2, \\ &\varphi = [a,a,-2a], \quad \chi = [0,0,c], \\ &\alpha_3 = \pm 1/\sqrt{6}, \quad \beta_1 = \pm 2/\sqrt{6}, \quad \beta_2 = 2/3; \end{aligned} \tag{129}$$

$$U_1:$$
$$\varphi=[a,b,-a-b]\,,\quad \chi=[0,0,c]\,,$$
$$\alpha_3=(a^3+b^3-(a+b)^3)/(a^2+b^2+(a+b)^2)^{3/2}\,,$$
$$\beta_1=-(a+b)/(a^2+b^2+(a+b)^2)^{1/2}\,, \tag{130}$$
$$\beta_2=(a+b)^2/(a^2+b^2+(a+b)^2)\,.$$

The generic stratum is represented by Eqs. (127)–(128) and its little group is the null group. Can a curve confine a three-dimensional volume? The answer is no, and thus the stratum of the null group must confine itself. The volume is extremized when either χ_1 or χ_2 is equal to zero with all the other components nonzero. The orbit space is shown in Fig. 9. The strata of SU_2, namely, the cusps, are the most protrudent as we might guess from the fact that they satisfy the most singular boundary conditions. The stratum of U_1, namely, the curve, is the next most singular. This may lead us to expect that such a hierarchical relationship would be a prominent feature of the orbit space of two irreps. But, as we shall see in the next example, the strata of a lower level little group can be as singular as the higher level ones.

B. SO(5) adjoint + vector representations

Using the same notation as in Ref. 17, the orbit parameters are

$$\alpha_4=\frac{\Sigma\varphi_i^4}{2(\Sigma\varphi_i^2)^2}\,, \tag{131}$$

$$\beta_2=\frac{\Sigma\chi_i\varphi_i^2\chi_i}{(2\Sigma\varphi_i^2)(\Sigma\chi_i\chi_i+\chi_3\chi_3)}\,, \tag{132}$$

$$\beta_4=\frac{\Sigma\chi_i\varphi_i^4\chi_i}{(2\Sigma\varphi_i^2)^2(\Sigma\chi_i\chi_i+\chi_3\chi_3)}\,,$$

where i runs from 1 to 2. The stratum of each little group is represented as follows:

$$SO_3:$$
$$10=1+3+3+3\,,\quad 5=1+1+3\,,$$
$$\varphi=[a,0]\,,\quad \chi=[c,0,0]\,, \tag{133}$$
$$\alpha_4=\tfrac12\,,\quad \beta_2=\tfrac12\,,\quad \beta_4=\tfrac14\,;$$

$$SU_2\times U_1:$$
$$10=1(0)+1(2)+1(-2)+3(0)+2(1)+2(-1)\,,$$
$$5=1(0)+2(1)+2(-1)\,,$$
$$\varphi=[a,a]\,,\quad \chi=[0,0,c]\,, \tag{134}$$
$$\alpha_4=\tfrac14\,,\quad \beta_2=0\,,\quad \beta_4=0\,;$$

$$U_1\times U_1:$$
$$\varphi=[a,b]\,,\quad \chi=[0,0,c]\,,$$
$$\alpha_4=(2a^4+2b^4)/(2a^2+2b^2)^2\,,\quad \beta_2=0\,,\quad \beta_4=0\,; \tag{135}$$

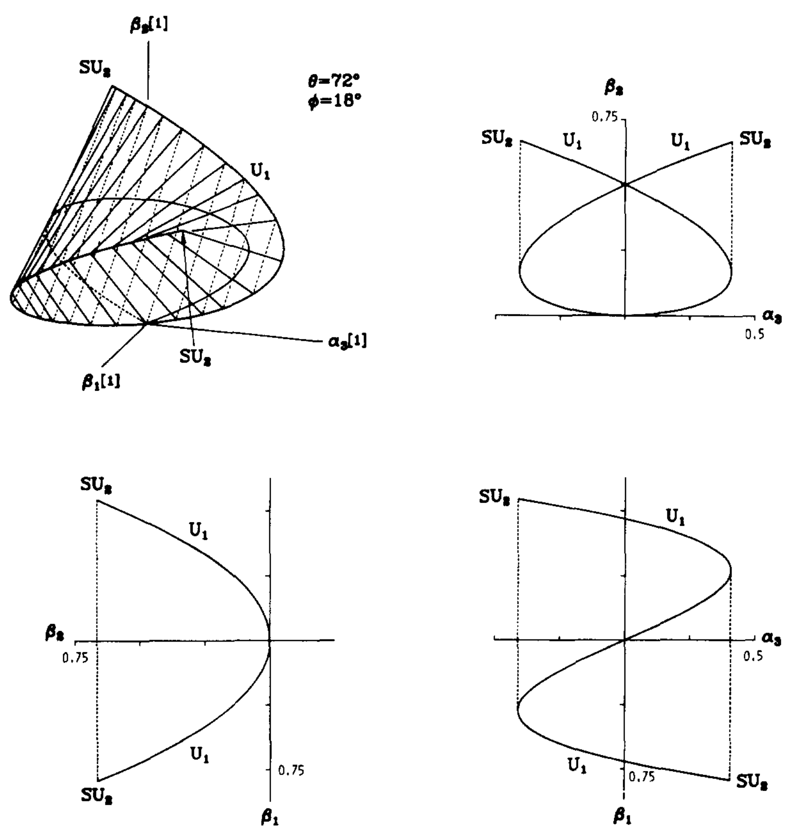


FIG. 9. The complete orbit space of the SU_3 adjoint + vector. Shown at the upper left corner is a view from the direction oriented 18° from the β_1 axis and 72° from the β_2 axis. The dotted lines are hidden lines. The numbers in the square brackets are the relative ratios of scale. Each projection is a view from the positive direction of the axis not shown in the picture. Here the dotted lines are portions of the boundary belonging to the null group stratum. Thus the hidden curves are drawn solidly.

$$
\begin{aligned}
&U_1: \\
&\varphi = [a,b]\,, \quad \chi = [0,c,d]\,, \\
&\alpha_4 = (2a^4 + 2b^4)/(2a^2 + 2b^2)^2\,, \\
&\beta_2 = (b^2c^2)/(2a^2 + 2b^2)(c^2 + d^2)\,, \\
&\beta_4 = (b^4c^2)/(2a^2 + 2b^2)^2(c^2 + d^2)\,.
\end{aligned} \tag{136}
$$

The generic stratum is represented by Eqs. (131)–(132) and its little group is the null group. The stratum of U_1 is two-dimensional and thus has a chance to enclose the whole volume. The U_1 stratum occupies the surfaces represented by dotted lines in Fig. 10, but the surface represented by solid lines is a part of the generic stratum. This is in contrast to the case of one irrep where there was no mixture of this kind. That is, equally singular surfaces consist of both the stratum of a maximaximal little group and a lower level one. Though the portion of the surface belonging to the null group is more singular than the interior, there is no way to distinguish them because there is no more subgroup left. The volume is extremized when either χ_2 is equal to zero (U_1) or χ_3 is zero (the null group) with all the other components nonzero.

C. Comments

Contrary to the case of one irrep where the strata of successively lower level little groups occupy successively higher-dimensional and less singular (locally less protrudent) surfaces on the orbit space boundary, the orbit space boundary of two irreps is more complex and things are pretty much mixed. Whereas orbit parameters associated with each irrep tend to form warped concave boundary surfaces, orbit parameters associated with both irreps tend to destroy such behavior. With the representation vector of one irrep fixed (consequently, orbit parameters associated with that irrep fixed), one can rotate the vector of the other irrep creating a volume traced by pencils.

It is notable that the generic strata in both examples are not totally open as in an irrep case. They close themselves partially. The same is true for lower-dimensional strata.

In the case of SU_3 adjoint + vector (Fig. 9) we find that the maximaximal little groups, SU_2 and U_1, occupy the most protrudent portions of the boundary. But in the case of SO_5 adjoint + vector (Fig. 10) we find that the U_1 stratum occupies the boundary planes indicated by the dotted lines and the stratum of the null group occupies the boundary plane indicated by the solid lines. That is, there is no sharp distinction between the maximaximal little group U_1 and the lower level little group, the null group, in terms of dimensionality and concavity.

Another interesting point is that the little groups alone cannot distinguish the fine structure of the orbit space. In both of the above-mentioned examples we see that the null group strata consist of two-dimensional surface *and* three-dimensional volume. In the SO_5 case the strata of U_1 consists of an edge curve *and* two-dimensional surfaces. This indistinguishability comes from the fact that, whereas for a given group there are only a finite number of subgroups, there is no limit to the dimension of a representation. As we see from Eq. (83) the orbit space dimension can be arbitrarily high. On

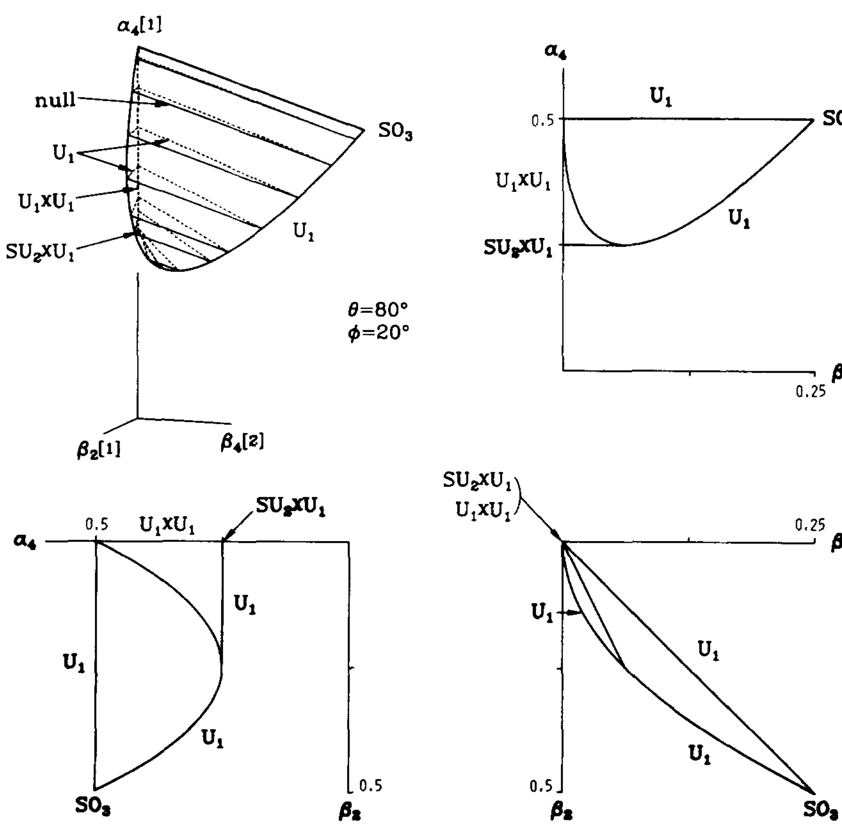


FIG. 10. The complete orbit space of the SO_5 adjoint + vector. Shown at the upper left corner is a view from the direction oriented 20° from the β_2 axis and 80° from the α_4 axis. The dotted lines are hidden lines. The numbers in the square brackets are the relative ratios of scale. Each projection is a view from the positive direction of the axis not shown in the picture. Here the hidden curves and lines are drawn solidly.

the other hand, the number of subgroups is too small to classify all the dimensions of the orbit space. Thus the indistinguishability is inevitable in both cases.

ACKNOWLEDGMENTS

I would like to thank Professor S. Frautschi for carefully reading the manuscripts. I am grateful to Dr. G. Sartori and Dr. M. Jarić for sending me their papers and illustrating their methods. I would like to thank Professor F. Gürsey for illuminating discussions and extensive references and for introducing to me mathematicians, Professor G. Zuckerman of Yale University and Professor Jacob Towber of De Paul University, who gave me an excellent lecture on classical invariant theory. I would like to thank Professor L. Michel for informing me of counterexamples to the Michel–Radicati conjecture and for encouraging me. I would like to thank Professor K. S. Kang for constructive criticism. I would like to thank Professor G. Schwarz of Brandeis University for explaining his works and Professor W.-C. Hsiang of Princeton University for helpful discussions and relevant reprints. I would like to thank Dr. M. Roček and the theoretical physics group at Stony Brook and Professor C. M. Bender for encouragement.

APPENDIX A

Some people might doubt that the minimization can be achieved so cheaply. To remove a possible doubt, we interpret our method of minimization in the orbit space in terms of conventional language. We show how to find all the extrema (in the representation space) of a smooth group-invariant function which is monotonic in the orbit parameters. We also explain how to tell the type of an extremum, i.e., a local minimum or an inflection point.

1. Single irreducible representation

Let us consider a group invariant smooth function, $P(\varphi)\equiv F(I_2,\lambda_1,\lambda_2,\cdots)$, which is a monotonic function of orbit parameters λ_a in the projected orbit space. In order to find an extremum of P in the representation space, we need to find the solution of the equation

$$\frac{\partial P}{\partial \varphi_i}=\frac{\partial r}{\partial \varphi_i}\frac{\partial F}{\partial r}+\frac{\partial \lambda_1}{\partial \varphi_i}\frac{\partial F}{\partial \lambda_1}+\frac{\partial \lambda_2}{\partial \varphi_i}\frac{\partial F}{\partial \lambda_2}+\cdots\equiv 0, \qquad \text{(A1)}$$

with $r=I_2^{1/2}$. Due to the assumed monotonicity, all $\partial F/\partial \lambda_a$ are nonzero in the projected orbit space.

Case (i): $\partial F/\partial r=0$

There are two ways to satisfy Eq. (A1):

$$\partial \lambda_a/\partial \varphi_i=0 \quad \text{for all } a \text{ and } i. \qquad \text{(A2)}$$

This is satisfied at all the cusps (including buried ones) corresponding to maximal little groups[28,29,16] with phase angles of complex invariants not counted

$$\left(\frac{\partial \lambda_1}{\partial \varphi_i},\frac{\partial \lambda_2}{\partial \varphi_i},\cdots\right)\perp\left(\frac{\partial F}{\partial \lambda_1},\frac{\partial F}{\partial \lambda_2},\cdots\right) \quad \text{for all } i. \qquad \text{(A3)}$$

This is satisfied when the contour of directional extrema contacts the orbit space boundary tangentially. Thus extrema of P with respect to φ_i may occur at points on the curves, two dimensional surfaces, etc. Notice that the tangential contact can occur.

It is laborious to check signs of a Hessian matrix in order to find the extremum type. However, once the orbit space is constructed, we can easily tell the type of an extremum from the way the contour of directional extremum meets the orbit space. For example, in the case of SO_8 adjoint (Fig. 11) there are only five extrema consisting of four cusps and one tangential contact point on the boundary curve. Two of the cusps are local minima because the contour has the lowest values at the cusps in their neighborhoods. (The contour touching the upper cusp passes through the lower right portion of the orbit space. However, the cusp is isolated from that portion.) The lower right one is the absolute minimum. On the other hand, the lower left one is a saddle point and the remaining two extrema are inflection points.

As we showed in Ref. 16, the contour of directional minima for a most general fourth-degree Higgs potential is flat or concave in the direction of increasing equipotential. Thus unless higher-dimensional strata are more protrudent than cusps, the absolute minimum will occur at cusps on the boundary of the projected orbit space. For a general smooth group-invariant function the contour may be convex in the direction of increasing equipotential. The cusps will still be the most likely points for the absolute minimum to occur. However, higher-dimensional strata will now have a better chance for becoming an absolute minimum.

Case (ii): $\partial F/\partial r\neq 0$

$$\left(\frac{\partial r}{\partial \varphi_i},\frac{\partial \lambda_1}{\partial \varphi_i},\frac{\partial \lambda_2}{\partial \varphi_i},\cdots\right)\perp\left(\frac{\partial F}{\partial r},\frac{\partial F}{\partial \lambda_1},\frac{\partial F}{\partial \lambda_2},\cdots\right) \quad \text{for all } i. \qquad \text{(A4)}$$

One of the equations cannot be satisfied, namely,

$$\epsilon_{ab\cdots}\frac{\partial \lambda_a}{\partial \varphi_i}\frac{\partial \lambda_b}{\partial \varphi_j}\cdots=\text{const (of } a \text{ and } i)\,\frac{\partial F}{\partial r}. \qquad \text{(A5)}$$

On the boundary the lhs of Eq. (A5) is identically zero. Thus Eq. (A5) can only be satisfied inside the projected orbit space, where the vectors, $(\partial \lambda_a/\partial \varphi_i, \partial \lambda_b/\partial \varphi_i,\cdots)$, are independent.

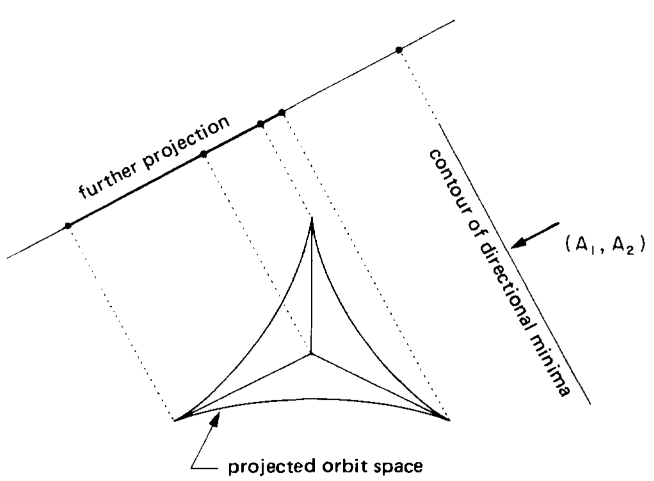


FIG. 11. The projected orbit space of the complete orbit space of the SO_8 adjoint is further projected onto a line with the contour of directional minima projected onto a point. Some cusps are projected onto the extreme boundary points; others into the interior.

The smallest number of independent vectors is the dimension of the projected orbit space. Now too many independent vectors have to be perpendicular to a fixed vector, the rhs of (A4). Therefore, Eq. (A4) cannot be satisfied at any point of the orbit space.

In fact, Michel[12] showed that a fourth degree Higgs potential cannot have an extremum in the generic stratum. We have extended his result to include more general cases. We have shown that a group-invariant function of a single irreducible representation monotonic in the projected orbit space can have an extremum only on the boundary of the projected orbit space with $\partial F/\partial r = 0$.

2. Two irreducible representations

Let us consider a group-invariant smooth function, $P(\varphi,\chi) \equiv F(I_2,\alpha_1,\alpha_2; J_2,\gamma_1,\gamma_2; \beta_1,\beta_2)$, which is a monotonic function of orbit parameters, $(\alpha_a,\gamma_c,\beta_b)$, in the projected orbit space. We have omitted further orbit parameters for the sake of saving space. It will not affect the generality of the following argument. In order to find an extremum of P in the representation space, we need to find the solution of the equation:

$$\frac{\partial P}{\partial \varphi_i} = \frac{\partial r}{\partial \varphi_i}\frac{\partial F}{\partial r} + \frac{\partial \alpha_1}{\partial \varphi_i}\frac{\partial F}{\partial \alpha_1} + \frac{\partial \alpha_2}{\partial \varphi_i}\frac{\partial F}{\partial \alpha_2} + \frac{\partial \beta_1}{\partial \varphi_i}\frac{\partial F}{\partial \beta_1} + \frac{\partial \beta_2}{\partial \varphi_i}\frac{\partial F}{\partial \beta_2} \equiv 0, \tag{A6a}$$

$$\frac{\partial P}{\partial \chi_j} = \frac{\partial s}{\partial \chi_j}\frac{\partial F}{\partial s} + \frac{\partial \gamma_1}{\partial \chi_j}\frac{\partial F}{\partial \gamma_1} + \frac{\partial \gamma_2}{\partial \chi_j}\frac{\partial F}{\partial \gamma_2} + \frac{\partial \beta_1}{\partial \chi_j}\frac{\partial F}{\partial \beta_1} + \frac{\partial \beta_2}{\partial \chi_j}\frac{\partial F}{\partial \beta_2} \equiv 0, \tag{A6b}$$

with $r = I_2^{1/2}$, $s = J_2^{1/2}$. Due to the assumed monotonicity, all $\partial F/\partial \alpha_a$, $\partial F/\partial \gamma_c$, $\partial F/\partial \beta_b$ are nonzero in the projected orbit space.

Case (i): $\partial F/\partial r = 0$ and $\partial F/\partial s = 0$

There are many ways to satisfy Eqs. (A6):

$$\frac{\partial \alpha_a}{\partial \varphi_i} = 0, \quad \frac{\partial \beta_b}{\partial \varphi_i} = 0, \quad \text{for all } a,b,i, \tag{A7a}$$

$$\frac{\partial \gamma_c}{\partial \chi_j} = 0, \quad \frac{\partial \beta_b}{\partial \chi_j} = 0, \quad \text{for all } c,b,j, \tag{A7b}$$

$$\left(\frac{\partial \alpha_1}{\partial \varphi_i}, \frac{\partial \alpha_2}{\partial \varphi_i}, \frac{\partial \beta_1}{\partial \varphi_i}, \frac{\partial \beta_2}{\partial \varphi_i}, 0,0\right) \perp \left(\frac{\partial F}{\partial \alpha_1}, \frac{\partial F}{\partial \alpha_2}, \frac{\partial F}{\partial \beta_1}, \frac{\partial F}{\partial \beta_2}, \frac{\partial F}{\partial \gamma_1}, \frac{\partial F}{\partial \gamma_2}\right) \quad \text{for all } i, \tag{A8a}$$

$$\left(0,0, \frac{\partial \beta_1}{\partial \chi_j}, \frac{\partial \beta_2}{\partial \chi_j}, \frac{\partial \gamma_1}{\partial \chi_j}, \frac{\partial \gamma_2}{\partial \chi_j}\right) \perp \left(\frac{\partial F}{\partial \alpha_1}, \frac{\partial F}{\partial \alpha_2}, \frac{\partial F}{\partial \beta_1}, \frac{\partial F}{\partial \beta_2}, \frac{\partial F}{\partial \gamma_1}, \frac{\partial F}{\partial \gamma_2}\right) \quad \text{for all } j. \tag{A8b}$$

Any combination of (A*I*a) and (A*J*b) with $I,J = 7,8$ will yield a solution to Eqs. (A6a) and (A6b). However, Eqs. (A7) are less frequently satisfied than in a single irrep case. The extremum conditions (A7) can also be satisfied partially in contrast to the single irrep case where all the orbit parameters are extremized simultaneously (with phase angles of complex invariants not counted). The points satisfying these conditions may be at cusps, on curves, on two dimensional surfaces, etc., on the orbit space boundary. They are all tangential contact points of the contour of directional extrema with the orbit space boundary. Again there are only finitely many extrema.

Since an orbit space for two irreps is formed from two independent spaces through the joint invariants, the Jacobian determinant[12,16] contains many zero elements. The dimension of a boundary portion is still given by the rank of the Jacobian.

Case (ii): $\partial F/\partial r \neq 0$ and/or $\partial F/\partial s \neq 0$

This condition again takes us into the projected orbit space and yields too many vectors to be perpendicular to a fixed vector.

Again we have shown that a group invariant function of two irreducible representations monotonic in the projected orbit space can have an extremum point only on the boundary of the projected orbit space with $\partial F/\partial r = 0$ and $\partial F/\partial s = 0$.

APPENDIX B

When the Higgs potential contains more than four independent invariant polynomials, it seems difficult to visually minimize the potential. We show how to find the absolute minima of these potentials. Let us consider a Higgs potential for a single irrep containing two third-degree invariants and three fourth-degree invariants. Call the associated orbit parameters, $\beta_1, \beta_2, \alpha_1, \alpha_2, \alpha_3$. Consider the following two-dimensional projected space of the five-dimensional orbit space:

$$\beta \equiv B_1\beta_1 + B_2\beta_2, \quad \alpha \equiv A_1\alpha_1 + A_2\alpha_2 + A_3\alpha_3, \tag{B1}$$

where B's and A's are the coupling coefficients of the corresponding invariant polynomials in the Higgs potential. Notice that β is proportional to the distance of the point (β_1, β_2) from the $\beta = 0$ line perpendicular to the vector (B_1,B_2), and that α is proportional to the distance of the point $(\alpha_1,\alpha_2,\alpha_3)$ from the $\alpha = 0$ plane perpendicular to the vector (A_1,A_2,A_3).

In the β–α space the absolute minimum of the potential can be found using the formula previously derived in Ref. 16. Now one necessarily asks whether we can uniquely determine (β_1, β_2) and $(\alpha_1,\alpha_2,\alpha_3)$ from a given set of (β,α). From the geometrical meaning of β and α, we see that there is a continuous range of orbit parameters satisfying Eq. (B1) for a given set of (β,α). However, the absolute minimum occurs at a unique point on the orbit space boundary, most protrudent to the direction of decreasing directional minimum. Thus we have a unique solution to Eq. (B1) at the absolute minimum. If the absolute minimum occurs on a concave portion of the orbit space boundary,[51] then there is a continuum of points satisfying Eq. (B1) and we have to stay in the five-dimensional orbit space. We have illustrated the mechanism in Fig. 11.

This raises the question: Is it safe to work in the projected orbit space which is dimensionally smaller than the representation vector space? Let us reconsider the significance of the boundary conditions:

$$\frac{\partial \lambda_a}{\partial \varphi_i} = 0 \quad \text{for all } a \text{ and } i\,, \tag{B2a}$$

$$\epsilon_{ab}\,\frac{\partial \lambda_a}{\partial \varphi_i}\,\frac{\partial \lambda_b}{\partial \varphi_j} = 0 \quad \text{for all } (a,b) \text{ and } (i,j),\cdots. \tag{B2b}$$

Equation (B2a) implies that, at a cusp corresponding to a null dimensional stratum, if we specify one orbit parameter, then all the other orbit parameters are determined. Equation (B2b) implies that, on a singular curve corresponding to a one-dimensional stratum, if we specify two orbit parameters, then all the other orbit parameters are determined, and so on.

The boundary conditions are strong enough to let us determine all the components of the scalar field (a vector in the representation space) at the absolute minimum from the knowledge of the norm and a small number of orbit parameters. The absolute minimum condition prompts the boundary condition, which in turn determines the whole vector.

Without further arguments we state that for an even degree Higgs potential of two irreps, we can safely work in the projected space, (α, β, γ), of the possibly high-dimensional projected orbit space of the complete orbit space.

However, in the process of further projection we lose the detailed extremum structure of the invariant function. This is evident in Fig. 11: we see five extrema before the projection and only two extrema afterwards. As far as we are looking for the absolute minimum only, the projection is harmless.

[1]R. Brout and F. Englert, Phys. Rev. Lett. **13**, 321 (1964); P. Higgs, Phys. Lett. **12**, 132 (1964); G. Guralnik, C. Hagen, and T. Kibble, Phys. Rev. Lett. **13**, 585 (1964); P. Higgs, Phys. Rev. **145**, 1156 (1966); T. Kibble, Phys. Rev. **155**, 1554 (1967).

[2]G. 't Hooft, Nucl. Phys. B **35**, 167 (1971); B. W. Lee and J. Zinn-Justin, Phys. Rev. D **5**, 3121, 3137, 3155 (1972); S. Coleman, "Secret Symmetry," Ettore Majorana Lecture Notes, 1973.

[3]R. Jackiw and K. Johnson, Phys. Rev. D **8**, 2386 (1973); J. M. Cornwall and R. E. Norton, Phys. Rev. D **8**, 3338 (1973); J. M. Cornwall, Phys. Rev. D **10**, 500 (1974); E. Eichten and F. Feinberg, Phys. Rev. D **10**, 3254 (1974).

[4]L. Susskind, Phys. Rev. D **20**, 2619 (1979); S. Dimopoulos and L. Susskind, Nucl. Phys. B **155**, 237 (1979); E. Eichten and K. D. Lane, Phys. Lett. B **90**, 125 (1980); E. Fahri and L. Susskind, Phys. Rep. **74**, 279 (1981).

[5]S. Weinberg, Phys. Rev. Lett. **19**, 1264 (1967); A. Salam, *Elementary Particle Theory*, edited by N. Svartholm (Almquist and Wiksell, Stockholm, 1968), p. 367.

[6]J. C. Pati and A. Salam, Phys. Rev. D **8**, 1240 (1973); H. Georgi and S. L. Glashow, Phys. Rev. Lett. **32**, 438 (1974); H. Fritzsch and P. Minkowski, Ann. Phys. (N.Y.) **93**, 193 (1975); F. Gürsey, P. Ramond, and P. Sikivie, Phys. Lett. B **60**, 177 (1976); P. Langacker, Phys. Rep. **72**, 185 (1981).

[7]H. Georgi, H. R. Quinn, and S. Weinberg, Phys. Rev. Lett. **33**, 451 (1974); E. Gildner, Phys. Rev. D **14**, 1667 (1976); S. Weinberg, Phys. Lett. B **82**, 387 (1979); J. Ellis, M. K. Gaillard, A. Petermann, and C. Sachrajda, Nucl. Phys. B **164**, 253 (1980).

[8]P. Fayet, Phys. Lett. B **58**, 67 (1975); L. O'Raifeartaigh, Nucl. Phys. B **96**, 331 (1975); B. deWitt and H. Nicolai, Nucl. Phys. B **208**, 323 (1982); N. Warner, "Some Properties of the Scalar Potential in Gauged Supergravity Theories," CALT-68-992, Apr. 1983.

[9]L. D. Landau, Phys. Z. Sov. **11**, 26 (1937); L. D. Landau and E. M. Lifschitz, *Statistical Physics* (Pergamon, London, 1958).

[10]P. W. Anderson, *Concepts in Solids* (Benjamin, New York, 1963); P. Kadanoff *et al.*, Rev. Mod. Phys. **39**, 395 (1967); R. Blinc and B. Žekš, *Soft Modes in Ferroelectrics and Antiferroelectrics* (Elsevier, New York, 1974); A. G. McLellan, *The Classical Thermodynamics of Deformable Materials* (Cambridge U. P., Cambridge, 1980); N. Boccara, *Symmetries and Broken Symmetries* (IDSET, Paris, 1981).

[11]L. O'Raifeartaigh, Rep. Prog. Phys. **42**, 159 (1979).

[12]L. Michel, CERN TH2716, 1979; Rev. Mod. Phys. **52**, 617 (1980); See also Ref. 27.

[13]F. Gürsey and M. Serdaroğlu, Nuovo Cimento A **65**, 337 (1981).

[14]J. S. Kim, Nucl. Phys. B **196**, 285 (1982).

[15]S. Frautschi and J. S. Kim, Nucl. Phys. B **196**, 301 (1982).

[16]J. S. Kim, Nucl. Phys. B **197**, 174 (1982).

[17]J. S. Kim, Nucl. Phys. B **207**, 374 (1982).

[18]R. Brout, Nuovo Cimento A **47**, 932 (1967).

[19]D. Hilbert, Math. Ann. **36**, 473 (1890); **42**, 313 (1893).

[20]S. Aronhold, J. Reine angew. Math. **62**, 281 (1863).

[21]L. Michel and L. A. Radicati, Ann. Phys. (N.Y.) **66**, 758 (1971).

[22]Nonlinear cases have been treated in Ref. 17. But the linear case is the ordinary one, and we confine our argument to it here for the sake of simplicity. The Landau potential is expanded normally up to the fourth or sixth degree in order parameters and is not always linear in the orbit parameters.

[23]L.-F. Li, Phys. Rev. D **9**, 1723 (1974).

[24]M. Magg and Q. Shafi, Z. Phys. C **4**, 63 (1980); F. Buccella, H. Ruegg, and C. A. Savoy, Nucl. Phys. B **169**, 68 (1980); F. Buccella, H. Ruegg, and C. A. Savoy, Phys. Lett. B **94**, 491 (1980); H. Ruegg, Phys. Rev. D **22**, 2040 (1980).

[25]L. Michel and L. A. Radicati, *Evolution of Particle Physics*, edited by M. Conversi (Academic, New York, 1970), p. 191.

[26]By an irreducible representation we mean a representation irreducible over the real. For example, $3 + \bar{3}$ or 8 of SU_3 is an irrep.

[27]M. Gell-Mann, private communication, 1980; R. Slansky, Phys. Rep. **79**, 1 (1981).

[28]L. Michel, C. R. Acad. Sci. Paris **272**, 433 (1971).

[29]M. Abud and G. Sartori, Phys. Lett. B **104**, 147 (1981); G. Sartori, J. Math. Phys. **24**, 765 (1983); M. Abud and G. Sartori, "The Geometry of Spontaneous Symmetry Breaking," preprint, Padova University, Jan. 1983 (to be published in Ann. Phys.).

[30]W. McKay and J. Patera, *Tables of Dimensions, Second and Fourth Indices and Branching Rules of Simple Algebras* (Dekker, New York, 1980).

[31]G. E. Bredon, *Introduction to Compact Transformation Groups* (Academic, New York, 1972); B. Kostant, Am. J. Math. **85**, 327 (1963); D. Mumford, "Geometric Invariant Theory," Ergeb. Math. **34** (1965); W. C. Hsiang and W. Y. Hsiang, Ann. Math. **92**, 189 (1970).

[32]V. G. Kac, V. L. Popov, and E. B. Vinberg, C. R. Acad. Sci. Paris **283**, 875 (1976); G. W. Schwarz, Inv. Math. **49**, 167 (1978).

[33]We observe that, in general, this is true only locally but not globally.

[34]The matrix T is called the representation of the group. However, particle theorists commonly call the vector space φ the representation space.

[35]We think Michel and Radicati[21] were the first physicists who constructed orbit spaces explicitly.

[36]M. V. Jarić and D. Mukamel, "Breakdown of Maximality Conjecture in Continuous Phase Transitions," preprint Weizman Institute, March 1983; M. V. Jarić, "Non-maximality and Successive Phase Transition," preprint, Montana State University, April 1983.

[37]O'Raifeartaigh, S. Y. Park, and K. C. Wali, *Symmetries in Science*, edited by B. Gruber (Plenum, New York, 1980); P. Mitra, J. Math. Phys. **22**, 2069 (1981); D.-d. Wu, Nucl. Phys. B **199**, 523 (1982).

[38]J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory* (Springer-Verlag, New York, 1972); B. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974); R. Gilmore, *Lie Groups, Lie Algebras, and Some of Their Applications* (Wiley, New York, 1974); M. L. Mehta, J. Math. Phys. **7**, 1824 (1966).

[39]G. Racah, Ergeb. Exakt. Naturwiss. **37**, 28 (1965).

[40]B. Gruber and L. O'Raifeartaigh, J. Math. Phys. **5**, 1796 (1964).

[41]For complex representations some invariants are complex. Then their complex conjugates are also invariant. In Ref. 14 we showed that when a Higgs potential contains such a pair of invariants, their phase at the potential minimum is determined by the phase of the complex coupling coefficients. Thus we will be working with the cross section of the orbit space cut at this phase angle.

[42]J. H. Grace and A. Young, *Algebra of Invariants* (Cambridge U. P., London, 1903); H. Weyl, *Classical Groups* (Princeton U. P., Princeton, NJ, 1946), 2nd ed.; J. A. Dieudonné and J. B. Carrell, *Invariant Theory, Old and New* (Academic, New York, 1971).

[43]J. Patera and R. T. Sharp, *Lecture Notes in Physics* **94**, edited by W. Beiglböck *et al.* (Springer-Verlag, New York, 1979), p. 175.

[44]M. Gell-Mann and R. Slansky, unpublished, 1979.

[45]H. Freudenthal, Adv. Math. **1**, 143 (1964); P. Ramond, "Introduction to Exceptional Lie Groups and Algebras," Preprint CALT-68-577, Cal-Tech, Dec. 1976; F. Gürsey, *Lecture Notes in Physics* **94**, edited by W. Beiglböck *et al.* (Springer-Verlag, New York, 1979), p. 508.
[46]F. Gürsey, invited talk at the New Hampshire Workshop, Apr. 1980; F. Gürsey, *Group Theoretical Methods in Physics*, edited by R. T. Sharp (Academic, New York, 1977), p. 213.
[47]G. Salmon, *Modern Higher Algebra* (Hodges, Forster, and Co., Dublin, 1876), 3rd ed., pp. 231–238.
[48]B. A. Ovrut, J. Math. Phys. **19**, 418 (1978).
[49]H. S. M. Coxeter, Ann. Math. **35**, 588 (1934); H. Freudenthal and H. de Vries, *Linear Lie Groups* (Academic, New York, 1969); N. Bourbaki, *Elements de Mathematique, Groupes et Algèbres de Lie* (Herman, Paris, 1971); S. Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces* (Academic, New York, 1978).
[50]M. V. Jarić, J. Math. Phys. **24**, 917 (1983); M. V. Jarić, L. Michel, and R. T. Sharp, "Zeros of Covariant Vector Fields for the Point Groups: Invariant Formulation," IHES/P/83/23, Mar. 1983.
[51]This may happen when the contour of directional minima is convex in the direction of increasing equipotential. A general group invariant function, other than the two types of Higgs potentials we are considering in this appendix, may allow a convex contour.

On the Hankel transform of a generalized Laguerre polynomial and on the convolution involving special Bessel functions

H. Navelet
Institut de Recherche Fondamentale, Departement de Physique Generale, CEN-Saclay, 91191 Gif-sur-Yvette Cedex, France

(Received 29 July 1983; accepted for publication 5 October 1983)

In this paper we calculate two of three simple integrals which are present neither in Bateman nor in Gradshtein and Ryzhik. These integrals are useful to treat some specific physics problems.

PACS numbers: 02.30.Bi, 02.30.Qy

1. INTRODUCTION

In the heyday of the Regge + absorption phenomenology we have been led[1] to consider the integral corresponding to the absorption of the form $I_{nn'} = \int J_\nu(xy) x^{2n+\nu+1} L_{n'}^{\nu+n}(\frac{1}{2}x^2) e^{-x^2/2}$ with $n \neq n'$. The integral with $n = n'$ is already known[2] but not the general one for $n \neq n'$. The purpose of our second part is to give the full proof of the result. On the other hand, quite recently we have been obliged to calculate the convolution integral involving two $I_{\nu,\mu}(x)$ functions. The convolution with a weight $x^{\nu,\mu}$ with ν integer and positive is already known[3] (with an error in Ref. 3) but we needed in our work[4] to calculate the analogous convolution when ν,μ are negative integers. This is the purpose of the third part of this paper.

2. CALCULATION OF $I_{nn'} = \int J_\nu(xy) x^{2n+\nu+1} L_{n'}^{\nu+n}(\frac{1}{2}x^2) e^{-x^2/2} dx,\ n \neq n'$

This integral in the special case $n = n'$ is well known[2] namely $I_{nn} = y^{2n+\nu} e^{-y^2/2} L_n^{\nu+n}(\frac{1}{2}y^2)$. In this work we show that it is possible to get a similar result for $n \neq n'$.

Indeed

$$L_{n'}^{\nu+n}(\tfrac{1}{2}x^2) = \frac{\Gamma(n+n'+\nu+1)}{\Gamma(n'+1)\Gamma(n+\nu+1)} \times {}_1F_1\left(-n', n+\nu+1, \frac{x^2}{2}\right), \tag{1}$$ [5]

where ${}_1F_1(a,b,c,x)$ is the confluent hypergeometric function

$$= \frac{\Gamma(n+n'+\nu+1)}{\Gamma(n'+1)\Gamma(n+\nu+1)} e^{x^2/2} \times {}_1F_1\left(n+\nu+1+n', n+\nu+1, -\frac{x^2}{2}\right),$$ [6]

so that

$$e^{-x^2/2} L_{n'}^{\nu+n}(\tfrac{1}{2}x^2) = \frac{\Gamma(n+n'+\nu+1)}{\Gamma(n'+1)\Gamma(n+\nu+1)} \times {}_1F_1\left(n+n'+\nu+1, n+\nu+1, -\frac{x^2}{2}\right),$$

$$I_{nn'} = \frac{1}{\sqrt{y}} \frac{\Gamma(n+n'+\nu+1)}{\Gamma(n'+1)\Gamma(n+\nu+1)} \times \int x^{2n+\nu+1/2} \times {}_1F_1\left(n+n'+\nu+1, n+\nu+1, -\frac{x^2}{2}\right) \times \sqrt{xy} J_\nu(xy) dx.$$

From Bateman,[7] the Hankel transform of $x^{2\beta-\nu-3/2}\,{}_1F_1(\alpha,\beta,-\lambda x^2)$ is

$$\frac{2^{2\beta-2\alpha-\nu-1}}{\Gamma(\alpha-\beta+\nu+1)\lambda^\alpha} y^{2\alpha-2\beta+\nu+1/2} \times {}_1F_1\left(\alpha, 1+\alpha-\beta+\nu, -\frac{y^2}{4\lambda}\right),$$

$$0 < \mathrm{Re}\,\beta < \tfrac{3}{4} + \mathrm{Re}(\alpha + \tfrac{1}{2}), \quad \mathrm{Re}\,\lambda > 0.$$

By identification,

$$\alpha = n + n' + \nu + 1,$$
$$\beta = n + \nu + 1,$$
$$\lambda = \tfrac{1}{2}$$

so that,

$$2\beta - \nu - \tfrac{3}{2} = 2n + \nu + \tfrac{1}{2},$$
$$2\beta - 2\alpha - \nu - 1 = -2n' - \nu - 1,$$
$$2\alpha - 2\beta + \nu + \tfrac{1}{2} = +2n' + \nu + \tfrac{1}{2},$$

and

$$I_{nn'} = \frac{1}{\sqrt{y}} \frac{\Gamma(n+n'+\nu+1)}{\Gamma(n'+1)\Gamma(n+\nu+1)} \times \frac{\Gamma(n+\nu+1)}{\Gamma(n'+\nu+1)} 2^{n-n'} y^{2n'+\nu+1/2} \times {}_1F_1\left(n+n'+\nu+1, n'+\nu+1, -\frac{y^2}{2}\right),$$

$$I_{nn'} = y^{2n'+\nu} 2^{n-n'} \frac{\Gamma(n+1)}{\Gamma(n'+1)} e^{-y^2/2} L_n^{\nu+n'}\left(\frac{1}{2}y^2\right). \tag{2}$$

By putting $n = n'$ we recover at once the well-known result. We note that there is one inversion of the indices n and n' between formula (1) and (2).

3. FINITE INTEGRALS INVOLVING BESSEL FUNCTIONS OF THE CONVOLUTION TYPE

These definite integrals may be evaluated by means of the convolution formula of the Laplace transform.

For instance let

$$F*G = \int_0^t F(v)G(t-v)dv,$$

$$f(s) = \int_0^\infty e^{-st}F(t)dt = L(F),$$

the Laplace transform of F,

$$g(s) = \int_0^\infty e^{-st}G(t)dt = L(G).$$

thus $L(F*G) = f(s)\cdot g(s)$. Let

$$F_\mu(a,t) = a^{-\mu}t(\mu/2)\,J_\mu(a\,t^{1/2})\mathrm{Re}\,\mu > -1,$$

$$f_\mu(a,s) = 2^{-\mu}s^{-\mu-1}\exp\left(-\frac{a^2}{4s}\right),$$ [3]

$$f_\mu(a,s)f_\nu(b,s) = 2\,2^{-(\mu+\nu+1)}s^{-(\mu+\nu+1)-1}$$

$$\times\exp\left(-\frac{a^2+b^2}{4s}\right)\mathrm{Re}\,\nu,\mu > -1$$

$$= 2f_{\mu+\nu+1}[(a^2+b^2)^{1/2},s].$$

As a straightforward consequence

$$\int_0^\tau dt\,t^{\mu/2}J_\mu(at^{1/2})(\tau-t)^{\nu/2}$$

$$J_\nu(b(\tau-t)^{1/2}) = 2a^\mu b^\nu\left\{\frac{\tau}{a^2+b^2}\right\}^{(\mu+\nu+1)/2}$$

$$\times J_{\mu+\nu+1}((a^2+b^2)^{1/2}\sqrt{\tau}),$$

$$\mathrm{Re}\,\mu,\nu > -1.\tag{3}$$ [8]

In this letter we want to find a similar convolution integral

$$I = \int_0^\tau t^{-\mu/2}I_\mu(at^{1/2})(\tau-t)^{-\nu/2}I_\nu(b(\tau-t)^{1/2})dt,$$

where ν and μ are real positive integers.

Let

$$F_\nu^a = t^{-\nu/2}I_\nu(2a^{1/2}t^{1/2}),$$

$$L(F_\nu^a) = p^{\nu-1}e^{a/p}\frac{\gamma(\nu,a/p)}{\Gamma(\nu)}a^{-\nu/2},$$ [9]

where

$$\gamma(\nu,x) = \int_0^x t^{\nu-1}e^{-t}dt$$

$$= \left(1 - e^{-x}\sum_{i=0}^{\nu-1}\frac{x^i}{i!}\right)\frac{1}{\Gamma(\nu)}$$

for ν integer $\geqslant 1$,

$$L(F_\nu^a * F_\mu^b) = p^{\nu+\mu-2}e^{(a+b)/p}\frac{\gamma(\nu,a/p)}{\Gamma(\nu)}\frac{\gamma(\mu,b/p)}{\Gamma(\mu)}.$$

A little bit of algebra yields

$$L(F_\nu^a \times F_\mu^b) = p^{\nu+\mu-2}e^{(a+b)/p}\left\{1 - e^{-a/p}\sum_{n=0}^{\nu-1}\left(\frac{a}{p}\right)^n\frac{1}{n!}\right\}\left\{1 - e^{-b/p}\sum_{n'=0}^{\mu-1}\left(\frac{b}{p}\right)^{n'}\frac{1}{n'!}\right\}$$

$$= p^{\nu+\mu-2}e^{(a+b)/p}\left[1 - e^{-(a+b)/p}\sum_{n''=0}^{\nu+\mu-2}\left(\frac{a+b}{p}\right)^{n''}\frac{1}{n''!}\right.$$

$$+ e^{-(a+b)/p}\left\{\sum_{n''=0}^{\nu+\mu-2}\left(\frac{a+b}{p}\right)^{n''}\frac{1}{n''!}\right.$$

$$\left.\left. + \sum_{n=0}^{\nu-1}\left(\frac{a}{p}\right)^n\frac{1}{n!}\sum_{n'=0}^{\mu-1}\left(\frac{b}{p}\right)^{n'}\frac{1}{n'!}\right\}\right] - e^{b/p}\sum_{n=0}^{\nu-1}\left(\frac{a}{p}\right)^n\frac{1}{n!}p^{\nu+\mu-2}$$

$$- e^{a/p}\sum_{n'=0}^{\mu-1}\left(\frac{b}{p}\right)^{n'}p^{\nu+\mu-2},$$

$$L(F_\nu^a \times F_\mu^b) = p^{\nu+\mu-2}e^{(a+b)/p}\frac{\gamma(\nu+\mu-1,(a+b)/p))}{\Gamma(\nu+\mu-1)} + p^{\nu+\mu-2}\left\{\sum_{n''=0}^{\nu+\mu-2}\left(\frac{a+b}{p}\right)^{n''}\frac{1}{n''!}\right.$$

$$\left. + \sum_{n=0}^{\nu-1}\left(\frac{a}{p}\right)^n\frac{1}{n!}\sum_{n'=0}^{\mu-1}\left(\frac{b}{p}\right)^{n'}\frac{1}{n'!}\right\}$$

$$- \sum_{n=0}^{\nu-1}\frac{a^n}{n!}p^{\nu+\mu-2-n}e^{b/p}\left[1 - e^{-b/p}\sum_{\alpha=0}^{\nu+\mu-2-n}\left(\frac{b}{p}\right)^\alpha\frac{1}{\alpha!}\right]$$

$$- \left\{\sum_{n=0}^{\nu-1}\left(\frac{a}{p}\right)^n\frac{1}{n!}\sum_{\alpha=0}^{\nu+\mu-2-n}\left(\frac{b}{p}\right)^\alpha\frac{1}{\alpha!}\right\}p^{\nu+\mu-2}$$

$$- \sum_{n'=0}^{\mu-1}\frac{b^{n'}}{n'!}p^{\nu+\mu-2+n'}e^{a/p}\left[1 - e^{-a/p}\sum_{\alpha'=0}^{\nu+\mu-2-n}\left(\frac{a}{p}\right)^{\alpha'}\frac{1}{\alpha'!}\right]$$

$$- \left\{\sum_{n'=0}^{\mu-1}\left(\frac{b}{p}\right)^{n'}\frac{1}{n'!}\sum_{\alpha'=0}^{\nu+\mu-2-n'}\left(\frac{a}{p}\right)^{\alpha'}\frac{1}{\alpha'!}\right\}p^{\nu+\mu-2}$$

$$= p^{\nu+\mu-2}e^{(a+p)/p}\frac{\gamma(\nu+\mu-1,(a+b)/p)}{\Gamma(\nu+\mu-1)}$$

$$
- \sum_{n=0}^{\nu-1} \frac{a^n}{n!} p^{\nu+\mu-2-n} e^{b/p} \frac{\gamma(\nu+\mu-1-n, b/p)}{\Gamma(\nu+\mu-1-n)}
- \sum_{n'=0}^{\mu-1} \frac{b^{n'}}{n'!} p^{\nu+\mu-2-n'} e^{a/p} \frac{\gamma(\nu+\mu-1-n', a/p)}{\Gamma(\nu+\mu-1-n')} + p^{\nu+\mu-2} S_{\mu\nu},
$$

where $S_{\mu\nu} = S^1_{\mu\nu} + S^2_{\mu\nu} - S^3_{\mu\nu} - S^4_{\mu\nu}$ and

$$
S^1_{\mu\nu} = \sum_{n''=0}^{\nu+\mu-2} \left(\frac{a+p}{p}\right)^{n''} \frac{1}{n''} = \sum_{n=0}^{\nu+\mu-2} \frac{1}{p^n} \sum_{\gamma=0}^{n} \frac{a^\gamma}{\gamma!} \frac{b^{n-\gamma}}{(n-\gamma)!},
$$

$$
S^2_{\mu\nu} = \sum_{n=0}^{\nu-1} \left(\frac{a}{p}\right)^n \frac{1}{n!} \sum_{n'=0}^{\mu-1} \left(\frac{b}{p}\right)^{n'} \frac{1}{n'!},
$$

$$
S^3_{\mu\nu} = \sum_{n=0}^{\nu-1} \left(\frac{a}{p}\right)^n \frac{1}{n!} \sum_{\alpha=0}^{\nu+\mu-2-n} \left(\frac{b}{p}\right)^{\alpha} \frac{1}{\alpha!},
$$

$$
S^4_{\mu\nu} = \sum_{n'=0}^{\mu-1} \left(\frac{b}{p}\right)^{n'} \frac{1}{n'!} \sum_{\beta=0}^{\nu+\mu-2-n'} \left(\frac{a}{p}\right)^{\alpha'} \frac{1}{\alpha'!}.
$$

Without loss of generality we may assume $\nu - 1 \geqslant \mu - 1 \geqslant 0$.

The different domains D_i of summation read (see Fig. 1). From inspection,

$$
D_1 = A + B + C, \quad D_2 = A \quad D_1 + D_2 = 2A + B + C,
$$

$$
D_3 = A + B, \quad D_4 = A + C \quad D_3 + D_4 = 2A + B + C.
$$

Thus trivially $S_{\mu\nu} = 0$ and

$$
L(F^a_\nu \times F^b_\mu) = p^{\nu+\mu-2} e^{(a+b)/p} \frac{\gamma(\nu+\mu-1, (a+b)/p)}{\Gamma(\nu+\mu-1)}
- \sum_{n=0}^{\nu-1} \frac{a^n}{n!} p^{\nu+\mu-2-n} e^{b/p} \frac{\gamma(\nu+\mu-1-n, (b/p))}{\Gamma(\nu+\mu-1-n)}
- \sum_{n'=0}^{\mu-1} \frac{b^{n'}}{n'!} p^{\nu+\mu-2-n'} e^{a/p} \frac{\gamma(\nu+\mu-1-n', (a/p))}{\Gamma(\nu+\mu-1-n')}. \tag{4}
$$

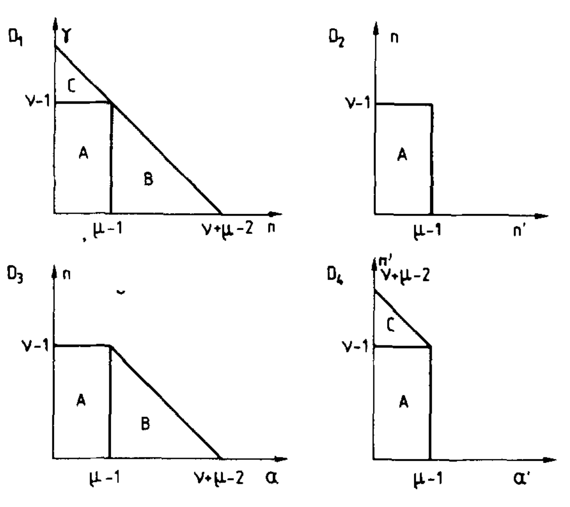


FIG. 1. Description of the domain D_i corresponding to the double summation $S^i_{\mu\nu}$. Domain A corresponds to $0 \leqslant \alpha_i \leqslant \mu - 1$, $0 \leqslant \beta_i \leqslant \nu - 1$, and domain B to $\mu - 1 \leqslant \alpha_i \leqslant \nu + \mu - 2$.

By performing the Laplace inverse transform we get

$$
I = \int_0^t x^{-\mu/2} I_\mu(\alpha\sqrt{x})(t-x)^{-\nu/2} I_\nu(\beta\sqrt{(t-x)})\,dx
$$

$$
= \left(\frac{2}{\alpha}\right)^\mu \left(\frac{2}{\beta}\right)^\nu \left\{ \left[\frac{\alpha^2+\beta^2}{4t}\right]^{(\nu+\mu-1)/2} \times I_{\nu+\mu-1}((\alpha^2+\beta^2)^{1/2}\sqrt{t}) \right.
$$

$$
- \sum_{i=0}^{\mu-1} \left(\frac{\alpha^2}{4}\right)^i \frac{1}{i!} \left[\frac{\beta^2}{4t}\right]^{(\nu+\mu-1-i)/2} I_{\nu+\mu-1-i}(\beta\sqrt{t})
$$

$$
\left. - \sum_{j=0}^{\nu-1} \left(\frac{\beta^2}{4}\right)^j \frac{1}{j!} \left[\frac{\alpha^2}{4t}\right]^{(\nu+\mu-1-j)/2} I_{\mu+\nu-1-j}(\alpha\sqrt{t}) \right\}. \tag{5}
$$

ACKNOWLEDGMENTS

I want to acknowledge fruitful discussion with A. Gervois.

[1] J. M. Drouffe and H. Navelet, Nuovo Cimento **39**, (1971).
[2] W. Magnus and F. Oberhetlinger, *Formeln und Sätze für die Specicelen Functionen der Mathematischen Physik* (Springer-Verlag, Berlin, 1948); *Gradshteyn Ryshik Table of Integral Series and Products* (Academic, New York, 1965), p. 847, Formula 7.421 (3).
[3] Bateman Manuscript Project, California Institute of Technology, *Higher Transcendental Functions*, edited by A. Erdelyi (McGraw-Hill, New York, 1953), Vol. 2, p. 46.
[4] G. Cohen-Tannoudji, A. Mantrach, H. Navelet, and R. Peschanski, "QCD and the rise of the hadronic total cross section," Phys. Rev. D **28**, 1628 (1983).
[5] Bateman Manuscript Project, California Institute of Technology, *Higher Transcendental Functions*, edited by A. Erdelyi (McGraw-Hill, New York, 1953), Vol. 2, p. 189, Formula 14.
[6] Bateman Manuscript Project, California Institute of Technology, *Higher Transcendental Functions*, edited by A. Erdelyi (McGraw-Hill, New York, 1953), Vol. 1, p. 253, Formula 7.
[7] Bateman Manuscript Project, California Institute of Technology, *Tables of Internal Transforms*, edited by A. Erdelyi (McGraw-Hill, New York, 1954), Vol. 2, p. 88, Formula 5.
[8] There is a misprint in Ref. 3.
[9] Bateman Manuscript Project, California Institute of Technology, *Tables of Integral Transforms*, edited by A. Erdelyi (McGraw-Hill, New York, 1954), Vol. 2, p. 185, Formula 31.

Explicit expressions of the nth gradient of $1/r$

B. Blaive and J. Metzger
Institute of Petroleschemistry and Industrial Organic Synthesis, Faculty of Sciences of Aix-Marseilles III, Saint-Jérome, 13013-Marseilles, France

(Received 20 September 1983; accepted for publication 19 January 1984)

Explicit expressions of the nth gradient tensor and partial derivatives of the function $1/r$ are given, in the absolute and in the local bases associated, respectively, with the Cartesian and the spherical coordinates of space. They may be used in various multipolar developments of r^{-1} interaction energies.

PACS numbers: 02.30.Gp

I. INTRODUCTION

The nth gradient of the function $1/r$ is useful notably for the calculation of electric $1/r$ interactions, when multipolar developments of the systems of charges are considered. The nth gradient tensor $\nabla^n(1/r)$ occurs (in two places sometimes) within the nth term of the corresponding series development of the interaction energy.[1–3] In many cases, the series converges slowly, which means that terms up to high values of n must be taken into account.

Because of their relation to the spherical harmonics, several properties of the nth gradients have been studied,[4–7] but explicit expressions of the nth gradient and derivatives of r^{-1} are not easily deduced from them, nor from Hobson's theorem.[8,9] A simple formula [see (15) further on] is known for the three diagonal terms of $\nabla^n(1/r)$, and an explicit expression of any nth partial derivative of $1/r$ may be obtained as a particular case of Maxwell's Theory of Poles.[10,11] But in some physical calculations, a tensorial expression of the nth gradient as a whole is needed, which is not provided simply by known results. In addition, an extrapolation from the first computed gradients is difficult, because of a great number of permuted terms which appear.

Notations: throughout, the tensorial product is written without a multiplicative sign, and $\overset{(q)}{\cdot}$ means a tensorial contraction of multiplicity q.

II. TENSORIAL EXPRESSION OF $\nabla^n(1/r)$ VERSUS VECTOR r

First it is proposed to establish the following explicit expression of the nth gradient tensor of r^{-1}:

$$\nabla^n \frac{1}{r} = \frac{(-1)^n}{r^{2n+1}} \mathscr{S}_n [P_n(\xi)] , \tag{1}$$

where, in the Legendre polynomial $P_n(\xi)$, the $(n-2q)$th power ξ^{n-2q} of ξ (ξ^0 included) symbolically stands for the 3^n-tensor $\xi^{n-2q} = (\overset{(q)}{\cdot}\mathbf{r}^n)\mathbf{I}_2^q$. Here $\mathbf{I}_2$ is the second-order identity tensor, and $\mathscr{S}_n: \mathbf{T} \mapsto \Sigma_\sigma \sigma\mathbf{T}$ the nth-order symmetrization operator,[12] sum of the $n!$ permutations.

The way in which (1) has been obtained is indicated in a shortened form in the Appendix. We will now give a faster derivation suggested by one referee.

A Taylor's development of the function $f(\mathbf{r}) = r^{-1}$ at point $\mathbf{r}$ may be written, using for instance the exponential mapping[13]

$$(\exp \boldsymbol{\zeta} \cdot \nabla) \frac{1}{r}\Big|_{\mathbf{r}} \equiv \sum_{n=0}^{\infty} \frac{1}{n!} \boldsymbol{\zeta}^n \overset{(n)}{\cdot} \nabla^n \frac{1}{r}\Big|_{\mathbf{r}} = \frac{1}{r}\Big|_{\mathbf{r}+\boldsymbol{\zeta}} , \tag{2}$$

that is, for any increase $-\boldsymbol{\zeta}$ of $\mathbf{r}$,

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \boldsymbol{\zeta}^n \overset{(n)}{\cdot} \nabla^n \frac{1}{r} = (r^2 - 2\mathbf{r}\cdot\boldsymbol{\zeta} + \zeta^2)^{-1/2} . \tag{3}$$

Let us denote the integer part of a number x by $E(x)$, and by

$$a_{n,q} = 2^{-n} \frac{(-1)^q (2n-2q)!}{(n-2q)!\,q!\,(n-q)!} , \tag{4}$$

the coefficients of the Legendre polynomial P_n. When transformed into the generating function

$$(1 - 2xy + y^2)^{-1/2} = \sum_{n=0}^{\infty} P_n(x) y^n \quad (0 < y < 1) \tag{5}$$

of the P_n's, the right-hand side of (3) becomes, for $\zeta < r$,

$$\frac{1}{r} \sum_{n=0}^{\infty} P_n\left(\frac{\mathbf{r}\cdot\boldsymbol{\zeta}}{r\zeta}\right)\left(\frac{\zeta}{r}\right)^n \tag{6}$$

$$= \frac{1}{r} \sum_{n=0}^{\infty} r^{-2n} \sum_{q=0}^{E(n/2)} a_{n,q} (\mathbf{r}\cdot\boldsymbol{\zeta})^{n-2q} r^{2q} \zeta^{2q} . \tag{7}$$

In (7),

$$(\mathbf{r}\cdot\boldsymbol{\zeta})^{n-2q} r^{2q} \zeta^{2q} = (\overset{(q)}{\cdot}\mathbf{r}^n)\mathbf{I}_2^q \overset{(n)}{\cdot} \boldsymbol{\zeta}^n . \tag{8}$$

Thus, Eq. (3) may be expressed

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \nabla^n \frac{1}{r} \overset{(n)}{\cdot} \boldsymbol{\zeta}^n = \sum_{n=0}^{\infty} \frac{1}{r^{2n+1}} P_n(\xi) \overset{(n)}{\cdot} \boldsymbol{\zeta}^n , \tag{9}$$

with the same meaning of ξ as in (1). When $\boldsymbol{\zeta}$ varies, $\boldsymbol{\zeta}^n$ linearly generates the space of symmetric 3^n-tensors; identification in (9) gives

$$\mathscr{S}_n\left(\frac{(-1)^n}{n!} \nabla^n \frac{1}{r} - \frac{1}{r^{2n+1}} P_n(\xi)\right) = \mathbf{0} , \tag{10}$$

which leads to (1) since $\nabla^n(1/r)$ is symmetric.

III. CARTESIAN PARTIAL DERIVATIVES OF $1/r$

It is easy to deduce from (1) the following expression of the Cartesian nth partial derivatives: for any α, β, γ $(\alpha + \beta + \gamma = n)$,

 0022-2488/84/061721-04$02.50

$$\frac{\partial^n}{\partial x^\alpha \partial y^\beta \partial z^\gamma}\frac{1}{r} = \frac{(-1)^n}{2^n r^{2n+1}}\alpha!\beta!\gamma! \sum_{q=0}^{E(n/2)}(-1)^q \frac{(2n-2q)!}{(n-q)!} r^{2q} \times \sum_{\xi=0}^{E(\alpha/2)}\sum_{\zeta=0}^{E(\beta/2)}\sum_{\mu=0}^{E(\gamma/2)} \frac{1}{\xi!\zeta!\mu!(\alpha-2\xi)!(\beta-2\zeta)!(\gamma-2\mu)!} x^{\alpha-2\xi}y^{\beta-2\zeta}z^{\gamma-2\mu} \quad (11)$$

with $\xi+\zeta+\mu+0$.

Let $Y_{x_1\cdots x_n}$ be any coordinate of $\nabla^n(1/r)$, where every $x_1,\ldots,x_n$ is either x or y or z. Its value drawn from (1),

$$Y_{x_1\cdots x_n} = \frac{(-1)^n}{2^n r^{2n+1}}\sum_{q=0}^{E(n/2)}(-1)^q\frac{(2n-2q)!}{(n-2q)!q!(n-q)!}r^{2q} \times \sum_{\text{perm }\sigma}(\delta_{x_1x_2}\cdots\delta_{x_{2q-1}x_{2q}}x_{2q+1}\cdots x_n), \quad (12)$$

depends only on the numbers α,β,γ of times that x, y, z, respectively, occur among $x_1,\ldots,x_n$.

Here, $A=\delta_{x_1x_2}\cdots\delta_{x_{2q-1}x_{2q}}x_{2q+1}\cdots x_n$ is different from zero for any permutation σ of $x_1\cdots x_n$ such that $\sigma(x_1)=\sigma(x_2),\ldots,\sigma(x_{2q-1})=\sigma(x_{2q})$, and then A takes the value $\sigma(x_{2q+1})\ldots\sigma(x_n)$. Let us count $2\xi,2\zeta,2\mu$ as the respective numbers of x, y, z among $\sigma(x_1),\ldots,\sigma(x_{2q})$.

$$0\leqslant\xi\leqslant E(\alpha/2),\quad 0\leqslant\zeta\leqslant E(\beta/2),\quad 0\leqslant\mu\leqslant E(\gamma/2),\qquad \xi+\zeta+\mu=q. \quad (13)$$

Then A is equal to $x^{\alpha-2\xi}y^{\beta-2\zeta}z^{\gamma-2\mu}$, and the number of permutations σ which correspond to the same triplet (ξ,ζ,μ) is

$$B=\binom{q}{\xi}\binom{q-\xi}{\zeta}\binom{n-2q}{\alpha-2\xi}\binom{n-2q-\alpha+2\xi}{\beta-2\zeta}\alpha!\beta!\gamma! = \frac{q!(n-2q)!\alpha!\beta!\gamma!}{\xi!\zeta!\mu!(\alpha-2\xi)!(\beta-2\zeta)!(\gamma-2\mu)!}. \quad (14)$$

This justifies the conversion of (12) into (11).

It is also possible to verify that (11) is in full agreement with Maxwell's general expression for a surface harmonic with given poles [Ref. 10, Chap. IX, (43)].

The diagonal term[2,10] $\alpha=\beta=0$, $\gamma=n$,

$$\frac{\partial^n}{\partial z^n}\frac{1}{r}=(-1)^n n! r^{-n-1}P_n(\cos\theta) \quad (15)$$

(θ is the angle of $\mathbf{r}$ with the z axis) is easily recovered from (1), since (1) implies

$$\frac{\partial^n}{\partial z^n}\frac{1}{r}=\frac{(-1)^n}{r^{2n+1}}n!P_n(\xi),$$

where $\xi^{n-2q}=r^n(\cos\theta)^{n-2q}$.

IV. TENSORIAL EXPRESSION OF $\nabla^n(1/r)$ IN THE LOCAL BASIS ASSOCIATED WITH THE SPHERICAL COORDINATES

Let $(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3)$ be the orthonormal local basis associated with the spherical coordinates (r,θ,φ). Let us show that the tensor $\nabla^n(1/r)$ can be explicitly expressed as

$$\nabla^n\frac{1}{r}=\frac{(-1)^n n!}{r^{n+1}}\mathscr{S}_n\sum_{p=0}^{E(n/2)}\frac{(-1)^p}{2^{2p}(n-2p)!p!^2} \times \mathbf{e}_1^{n-2p}(\mathbf{e}_2^2+\mathbf{e}_3^2)^p, \quad (16)$$

where $\mathscr{S}_n$ still represents the nth-order symmetrization operator.

The gradient operator in spherical coordinates

$$\nabla=\mathbf{e}_1\frac{\partial}{\partial r}+\mathbf{e}_2\frac{1}{r}\frac{\partial}{\partial\theta}+\mathbf{e}_3\frac{1}{r\sin\theta}\frac{\partial}{\partial\varphi}$$

may be applied to $\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3$ to give the tensors

$$\nabla\mathbf{e}_1=(1/r)(\mathbf{e}_2^2+\mathbf{e}_3^2)=\mathbf{v}/r, \quad (17a)$$

$$\nabla\mathbf{e}_2=(1/r)(-\mathbf{e}_2\mathbf{e}_1+\mathbf{e}_3^2\cot\theta), \quad (17b)$$

$$\nabla\mathbf{e}_3=-(1/r)(\mathbf{e}_3\mathbf{e}_1+\mathbf{e}_3\mathbf{e}_2\cot\theta). \quad (17c)$$

On using (17), the gradient of $\mathbf{v}=\mathbf{e}_2^2+\mathbf{e}_3^2$ verifies

$$\mathscr{S}_3\nabla\mathbf{v}=-(2/r)\mathscr{S}_3(\mathbf{e}_1\mathbf{v}). \quad (18)$$

For the first values of n, $\nabla^n(1/r)$ may be easily obtained as a function of r and $(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3)$, by repeated application of (17a) and (18). Consideration of the values obtained up to $n=4$ leads to the search of $\nabla^n(1/r)$ in the form

$$\nabla^n\frac{1}{r}=r^{-n-1}(-1)^n\mathscr{S}_n\sum_{p=0}^{E(n/2)}(-1)^p b_{n,p}\mathbf{e}_1^{n-2p}\mathbf{v}^p, \quad (19)$$

where $b_{n,p}$ is a coefficient to be determined.

With the help of relations (17a) and (18), let us calculate $\nabla^{n+1}(1/r)=\nabla\nabla^n(1/r)$ as a function of the coefficients $b_{n,p}$s, which will lead to a recursive relation between the $b_{n+1,p}$s and the $b_{n,p}$s.

Here, $\nabla(\mathbf{e}_1^{n-2p}\mathbf{v}^p)$ has the same symmetric transform through $\mathscr{S}_{n+1}$ as

$$(1/r)[(n-2p)\mathbf{e}_1^{n-2p-1}\mathbf{v}^{p+1}-2p\mathbf{e}_1^{n-2p+1}\mathbf{v}^p].$$

This and (19) give, after some rearrangements,

$$\nabla^{n+1}\frac{1}{r}=\frac{(-1)^{n+1}}{r^{n+2}}\mathscr{S}_{n+1}\left[\sum_{p=0}^{E(n/2)}(-1)^p\left(1+\frac{2p}{n+1}\right)\times b_{n,p}\mathbf{e}_1^{n+1-2p}\mathbf{v}^p+\sum_{q=1}^{E(n/2+1)}(-1)^q \times\frac{n-2q+2}{n+1}b_{n,q-1}\mathbf{e}_1^{n+1-2q}\mathbf{v}^q\right]. \quad (20)$$

Comparison of (20) with (19) written to the order $n+1$, implies the recursive relation

$$b_{n+1,p}=\begin{cases} b_{n,p}\left(1+\dfrac{2p}{n+1}\right)+b_{n,p-1}\dfrac{n-2p+2}{n+1}, & \text{for } p=1,\ldots,E(n/2),\\ b_{n,p}, & \text{for } p=0,\\ b_{n,p-1}\dfrac{n-2p+2}{n+1}, & \text{for } p=E(n/2+1)\neq E(n/2).\end{cases} \quad (21)$$

The resolution of (21) is not obvious, but the solution

$$b_{n,p} = \frac{n!}{(n-2p)!}\frac{1}{2^{2p}}\frac{1}{p!^2} \tag{22}$$

arises after some attempts. Its substitution into (19) gives (16).

In this derivation, the existence of (19) has been supposed, and the uniqueness of (22) is not shown. However, once (16) is found, the precedent reasoning can serve, almost without change, as an inductive demonstration.

Note that substitution of

$$\mathbf{I}_2^q = (\mathbf{e}_1^2 + \mathbf{e}_2^2 + \mathbf{e}_3^2)^q \equiv \sum_{m=0}^{q} (\mathbf{e}_2^2 + \mathbf{e}_3^2)^m \mathbf{e}_1^{2q-2m} \frac{q!}{m!(q-m)!}$$

(last equality modulo a symmetrization by $\mathscr{S}_{2q}$) in (1), leads to the expression

$$\nabla^n \frac{1}{r} = \frac{(-1)^n}{r^{n+1}2^n} \mathscr{S}_n \sum_{m=0}^{E(n/2)} \frac{1}{m!} \mathbf{e}_1^{n-2m}(\mathbf{e}_2^2 + \mathbf{e}_3^2)^m \times \sum_{q=m}^{E(n/2)} \frac{(-1)^q(2n-2q)!}{(n-2q)!(n-q)!(q-m)!}. \tag{23}$$

The equality

$$\sum_{q=m}^{E(n/2)} \frac{(-1)^q(2n-2q)!}{(n-2q)!(n-q)!(q-m)!} = \frac{(-1)^m n! 2^{n-2m}}{(n-2m)!m!} \tag{24}$$

results from (16) and (23), but no other very much simpler derivation of (24) seems to exist.

V. nth PARTIAL DERIVATIVES ALONG THE LOCAL AXES

Let x', y', z' be the three coordinates of space in the local basis $(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3)$. The nth partial derivatives of r^{-1} along the axes $\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3$ are the coordinates $Z_{x'_1\cdots x'_n}$ of the tensor $\nabla^n(1/r)$ in the product basis $\{\mathbf{e}_{i_1}\cdots\mathbf{e}_{i_n}\}$, in which $\mathbf{e}_{i_1},\ldots,\mathbf{e}_{i_n}$ run independently over the three vectors $\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3$. Since $\nabla^n(1/r)$ is a symmetric tensor, $Z_{x'_1\cdots x'_n}$ only depends on the numbers α,β,γ of times that x', y', z', respectively, occur among $x'_1,\ldots,x'_n$: $Z_{x'_1\cdots x'_n} = Z_{x'^\alpha y'^\beta z'^\gamma} = Z_{\alpha\beta\gamma}$.

Let us prove that for any $\alpha, \beta, \gamma, \alpha+\beta+\gamma = n$,

$$\frac{\partial^n}{\partial x'^\alpha \partial y'^\beta \partial z'^\gamma}\frac{1}{r} = \begin{cases} 0, & \text{if } \beta \text{ or } \gamma \text{ is odd}, \\ \dfrac{n!}{r^{n+1}} \dfrac{(-1)^{(3n-\alpha)/2}\beta!\gamma!}{2^{n-\alpha}(\beta/2)!(\gamma/2)![(\beta+\gamma)/2]!}, & \text{if } \beta \text{ and } \gamma \text{ are even}. \end{cases} \tag{25}$$

Being symmetric, $\nabla^n(1/r)$ is equal to

$$\nabla^n \frac{1}{r} = \sum_{\substack{\alpha,\beta,\gamma \\ \alpha+\beta+\gamma=n}} \frac{Z_{\alpha\beta\gamma}}{\alpha!\beta!\gamma!} \mathscr{S}_n(\mathbf{e}_1^\alpha \mathbf{e}_2^\beta \mathbf{e}_3^\gamma). \tag{26}$$

The comparison of (26) with the development of (16),

$$\nabla^n \frac{1}{r} = \frac{(-1)^n n!}{r^{n+1}} \sum_{p=0}^{E(n/2)} \frac{(-1)^p}{2^{2p}(n-2p)!p!} \times \sum_{t=0}^{p} \frac{1}{t!(p-t)!} \mathscr{S}_n(\mathbf{e}_1^{n-2p}\mathbf{e}_2^{2t}\mathbf{e}_3^{2p-2t}), \tag{27}$$

shows that $Z_{\alpha\beta\gamma}/\alpha!\beta!\gamma!$ is the coefficient appearing in (27) corresponding to the values $p = (n-\alpha)/2$ and $t = \beta/2$. This gives the value (25) of $Z_{\alpha\beta\gamma}$. As expected, (25) is symmetric with respect to β and γ, not α.

VI. CONCLUSION

The explicit expressions (1), (11), (16), and (25) may be used for the nth gradient and partial derivatives of the function r^{-1}, in the absolute and the local bases associated, respectively, with the Cartesian and the spherical coordinates of space. The linear decomposition of $\nabla^n(1/r)$ on fixed bases of $2n+1$ tensors [that generate an irreducible linear representation of the SO(3) group[14–16]] has not been considered here, but will be elsewhere.

These calculations have been performed for dealing with electric interactions at the molecular scale [as an example, (1) may be used to express Buckingham's multipolar moments[1] as functions of the ordinary Cartesian moments]. However, it is possible that they could be applied also in other parts of physics concerned with r^{-1} interactions.

APPENDIX: ORIGINAL DERIVATION OF EXPRESSION (1)

Internal contraction of $\nabla^n(1/r)$ leads to a null tensor:

$$\cdot\nabla^n(1/r) = \mathbf{0}. \tag{A1}$$

Apart from this, examination of the first orders n show that $(-1)^n[2^n n!/(2n)!]r^{2n+1}\nabla^n(1/r)$ and $\mathbf{r}^n$ are equivalent in the following sense: for any harmonic function Φ,

$$(-1)^n \frac{2^n n!}{(2n)!} r^{2n+1}\nabla^n \frac{1}{r} \overset{(n)}{\cdot} \nabla^n \Phi = \mathbf{r}^n \overset{(n)}{\cdot} \nabla^n \Phi. \tag{A2}$$

Expression (1) may be derived by constructing a tensor $\mathbf{A}$ which could be substituted to $\nabla^n(1/r)$ both in (A1) and (A2).

It is not hard to see first that the contracted product $\mathbf{r}^n \overset{(n)}{\cdot} \nabla^n \Phi$ is unchanged when adding to $\mathbf{r}^n$ any linear combination of terms of the type $\mathscr{S}_n(\mathbf{B}^{(n-2)}\mathbf{I}_2)$, where $\mathbf{B}^{(n-2)}$ is any tensor of order $n-2$. So it is considered to subtract from $\mathbf{r}^n$ a series $\mathbf{S}$ of the type

$$\mathbf{S} = \sum_q c_{n,q} \mathscr{S}_n \left[\mathbf{I}_2^q \left(\overset{(q)}{\cdot} \mathbf{r}^n\right)\right].$$

A recursive construction of the $c_{n,q}$s, to satisfy to $\cdot\mathbf{S} = \mathbf{0}$, leads to (1). For a rigorous proof of (1), induction may then be used.

[1] A. D. Buckingham, Q. Rev. **13**, 183 (1959).
[2] C. J. F. Böttcher, O. C. Van Belle, P. Bordewijk, and A. Rip, *Theory of Electric Polarization* (Elsevier, Amsterdam, 1973).
[3] B. Blaive and J. Metzger, J. Phys. (Paris) **42**, 1533 (1981).
[4] A. Clebsch, J. Math. **60**, 343 (1862).
[5] E. Heine, *Handbuch der Kugelfunctionen, Theorie und Anwendungen. Theorie der Kugelfunctionen und der verwandten Functionen* (Reimer, Berlin, 1878).
[6] F. Neumann, *Beiträge zur Theorie der Kugelfunctionen* (Teubner, Leipzig, 1878).
[7] E. W. Hobson, *The Theory of Spherical and Ellipsoidal Harmonics* (Chelsea, New York, 1965).

[8]E. W. Hobson, Proc. London Math. Soc. **24**, 55 (1892).
[9]E. W. Hobson, Messenger Math. **23**, 115 (1894).
[10]J. C. Maxwell, *A Treatise on Electricity and Magnetism* (Clarendon, Oxford, 1891).
[11]J. J. Sylvester, Phil. Mag. **2**, 291 (1876).
[12]L. Schwartz, *Les tenseurs* (Hermann, Paris, 1981).
[13]Y. Choquet-Bruhat, C. de Witt-Morette, and M. Dillard-Bleick, *Analysis, Manifolds and Physics* (North-Holland, Amsterdam, 1977).
[14]G. Pichon, *Groupes de Lie. Représentations linéaires et applications* (Hermann, Paris, 1973).
[15]A. J. Stone, Mol. Phys. **29**, 1461 (1975).
[16]J-M. Normand and J. Raynal, J. Phys. A **15**, 1437 (1982).

A recursive generation of local higher-order sine–Gordon equations and their Bäcklund transformation

Marco Boiti,[a] Jérôme J.-P. Leon,[b] and Flora Pempinelli[a]
Laboratoire de Physique Mathématique, Université des Sciences et Techniques du Languedoc, F 34060 Montpellier Cedex, France

(Received 2 December 1983; accepted for publication 10 February 1984)

A new hierarchy of local nonlinear evolution equations is generated by a recursion operator and its explicit inverse. It is shown that this hierarchy satisfies a canonical geometrical scheme and that it contains as special cases the sine–Gordon and Liouville equations in laboratory coordinates. A generalization of the well-known Bäcklund transformation and nonlinear superposition formula for the sine–Gordon equation is also obtained.

PACS numbers: 02.30.Jr, 03.40.Kf

I. INTRODUCTION

The inverse spectral transform (IST) has been successfully used to solve many interesting nonlinear evolution equations (NEE's for short).[1] These NEE's usually appear as members of an infinite hierarchy of partial differential equations generated by an integrodifferential operator, the so-called recursion operator, that is associated with an eigenvalue problem. Among all these IST-solvable NEE's there is one famous equation that does not belong to a hierarchy: the sine–Gordon equation (SG) in laboratory coordinates (also called the "physical sine–Gordon equation")

$$\omega_{tt} - \omega_{xx} + \sin \omega = 0, \quad \omega = \omega(x,t), \quad x\in\mathbb{R}, \quad t\in\mathbb{R} \tag{1.1}$$

(subscripts x and t mean partial derivative) that arises in many physical systems.[2] According to the AKNS method this equation has been obtained[3] as the integrability condition written in terms of the one-form Ω

$$d\Omega - \Omega \wedge \Omega = 0, \tag{1.2}$$

or as a Lax-pair representation written in terms of the spectral operators U and V

$$U_t - V_x + [U,V] = 0, \tag{1.3}$$

for the isospectral eigenvalue (λ) problem

$$d\Psi = \Omega\Psi, \quad \Omega = U(\omega;\lambda)dx + V(\omega;\lambda)dt, \tag{1.4a}$$

$$U(\omega;\lambda) = -\frac{i}{4}\sigma_3(\omega_x + \omega_t) + i\sigma_2\Big(\lambda - \frac{1}{16\lambda}\exp[i\sigma_3\omega]\Big), \tag{1.4b}$$

$$V(\omega;\lambda) = -\frac{i}{4}\sigma_3(\omega_x + \omega_t) + i\sigma_2\Big(\lambda + \frac{1}{16\lambda}\exp[i\sigma_3\omega]\Big), \tag{1.4c}$$

where the σ_j's are the Pauli matrices.

Usually an infinite hierarchy of IST-solvable NEE's related to a single operator U is obtained by choosing infinitely many different V satisfying the Lax-pair representation (1.3). However, it seems that the above given structure of $U(\omega;\lambda)$ is compatible only with the particular choice (1.4c) of V [no higher powers of λ and λ^{-1} are allowed for V in (1.3), see Sec. V].

In order to overcome this difficulty, we use the fact that a NEE in one field can be written as a system of two coupled NEE's in two fields and we search for a convenient spectral problem involving two independent potentials (fields).

Then we are able to exhibit the sine–Gordon equation as a member of a NEE hierarchy, to find the recursion operator for the hierarchy and moreover to provide the geometrical (Hamiltonian) structure for this new hierarchy. We also derive the Bäcklund transformation and the nonlinear superposition formula (double Bäcklund transformation) for the equations in the hierarchy. Both of them are a nontrivial generalization of the corresponding ones found by Bäcklund and by Bianchi for the sine–Gordon equation.

We use as a starting point the spectral problem recently introduced[4]:

$$\Psi_x = U\Psi, \quad U = -i\lambda\sigma_3 + u\sigma_1 + (i/\lambda)(s\sigma_3 + iv\sigma_2), \tag{1.5}$$

together with the so-called auxiliary spectral problem

$$\Psi_t = V\Psi, \quad V = \sum_{j=0}^{n} V_j\lambda^{n-j} + \sum_{j=0}^{p} W_j\lambda^{j-p-1}. \tag{1.6}$$

The main point is the fact that the three potentials $u(x,t)$, $v(x,t)$, and $s(x,t)$ will be shown to obey the following one-parameter family of compatible reductions[5]:

$$s^2 - v^2 = s_0^2, \quad s_{0,x} = s_{0,t} = 0. \tag{1.7}$$

Two main hierarchies of (reduced) NEE's then follow from (1.3) in which U and V are now given by (1.5) and (1.6): one for $s_0 \neq 0$ which includes the sine–Gordon equation and some generalizations, the other for $s_0 = 0$ which includes the Liouville equation (and generalizations) and has already been investigated, in the case $W_j = 0$ in (1.6), in a slightly different way.[6]

We do not deal here with the sine–Gordon equation in the light-cone coordinates ($\omega_{xt} = \sin \omega$) which is related to the Zakharov–Shabat spectral problem generalized by Ablowitz *et al.*[7] for which the recursion operator and geometrical structure have been found.[8] In that case the hierarchy of NEE's is integrodifferential (for a spectral operator V

[a] Permanent address: Dipartimento di Fisica dell'Università, I-73100 Lecce, Italy.
[b] Equipe de recherche associée au Centre National de la Recherche Scientifique, ERA 154, Montpellier, France.

 0022-2488/84/061725-10$02.50

with negative powers of λ) and there arise local and nonlocal conservation laws. Whereas for the "physical" case (1.1) an infinite sequence of local conservation laws exists,[9,10] and we exhibit here the hierarchy of NEE's which are shown to be purely differential.

We will make our study in the frame of the geometrical scheme developed by many authors,[11-17] and use the notations of Ref. 17. In this scheme it will be shown that the hierarchy corresponding to (1.5) and (1.6) is generated by an operator L^+ and its (explicit) inverse $(L^+)^{-1}$ which are both (i) coupled to a cosymplectic operator J, (ii) Nijenhuis operators (or hereditary symmetries), and (iii) recursion operators (or strong symmetries). We then use the machinery detailed in Ref. 17 to establish the eigenvalue problems for L, to derive the Hamiltonian character of the NEE's of the hierarchy and to find the recursion relations that allow us to compute the explicit forms of the infinitely many conserved quantities (which are in involution).

We recall hereafter some of the definitions and notations used in this paper: $q(x,t)$ denotes a vector valued field function with C^∞ components $q_j(x,t)$ decaying "rapidly enough" (to zero or to a constant) as $|x|\to\infty$ (see Ref. 18). As a function of the x coordinate only, q can be considered as a point in a linear normed space M. With each point we associate a tangent space T_q of smooth vector fields $\alpha(x,t)$ (contravariant fields) with the same asymptotic behavior as q. The cotangent space T_q^* of covariant fields $\beta(x,t)$ is defined through the symmetric bilinear form:

$$\langle \beta,\alpha\rangle = \sum_j \int_{-\infty}^{+\infty} dx\, \beta_j(x,t)\alpha_j(x,t). \tag{1.8}$$

The directional derivative (or Gâteaux derivative) of a functional $G(q,q_x,q_{xx},\ldots)$ is given by

$$G'(q)[\alpha] = \frac{d}{d\epsilon} G(q+\epsilon\alpha)|_{\epsilon=0}. \tag{1.9}$$

An operator γ: $M\to T_q^*$ is called a potential operator iff

$$\exists F: M\to C / F'(q)[\alpha] = \langle\gamma,\alpha\rangle, \tag{1.10}$$

or, equivalently, iff

$$\exists F: M\to C / \gamma = (\delta/\delta q)F(q), \tag{1.11}$$

where the variational derivative $\delta/\delta q$ is defined as follows:

$$\frac{\delta}{\delta q}\int_{-\infty}^{+\infty} dx\, f(q) = \sum_{j\geqslant 0} (-D)^j \partial_j f, \tag{1.12}$$

with

$$D \equiv \frac{d}{dx}, \quad \partial_j \equiv \frac{\partial}{\partial(D^j q)}. \tag{1.13}$$

As the inverse of the D operator we choose the integral operator I:

$$I(\cdot) = \frac{1}{2}\left(\int_{-\infty}^{x} - \int_{x}^{\infty}\right) dx(\cdot) \tag{1.14}$$

that satisfies

$$I^+ = -I. \tag{1.15}$$

[The notation A^+ means adjoint of operator A according to the bilinear form (1.8).]

II. HIERARCHY OF NONLINEAR EVOLUTION EQUATIONS

To derive the explicit form of the recursion operators L and L^{-1} we solve the equation (1.3) in which U and V are given in (1.5) and (1.6). Since U is traceless, we consider a traceless auxiliary spectral operator V and make the following two alternative choices for V:

$$\begin{aligned} &W_j(x,t) = a_j(x,t)\sigma_3 + \tfrac{1}{2}b_j(x,t)\sigma_1 - \tfrac{1}{2}c_j(x,t)\sigma_2,\\ &V_j(x,t)\equiv 0, \end{aligned} \tag{2.1a}$$

$$\begin{aligned} &V_j(x,t) = d_j(x,t)\sigma_3 + \tfrac{1}{2}e_j(x,t)\sigma_1 - \tfrac{1}{2}f_j(x,t)\sigma_2,\\ &W_j(x,t)\equiv 0. \end{aligned} \tag{2.1b}$$

By equating to zero the coefficients of given powers of λ in (1.3) for the first choice (2.1a) we get the following set of recursion relations:

$$b_0 = 0, \tag{2.2}$$

$$sc_0 = -2iva_0, \tag{2.3}$$

$$sb_1 = \tfrac{1}{2}c_{0,x} - 2iua_0, \tag{2.4}$$

$$vb_1 = -ia_{0,x} + uc_0, \tag{2.5}$$

$$\left.\begin{aligned} vb_{j+1} &= -ia_{j,x} + uc_j,\\ sb_{j+1} &= b_{j-1} + \tfrac{1}{2}c_{j,x} - 2iua_j,\\ sc_j &= c_{j-2} - \tfrac{1}{2}b_{j-1,x} - 2iva_j, \end{aligned}\right\}\ j=1,\ldots,p, \tag{2.6}$$

where we have introduced for convenience b_{p+1}. In the same way we get for the second choice (2.1b) the set of recursion relations for V_j:

$$e_0 = 0, \quad e_1 = 2iud_0, \tag{2.7}$$

$$d_{0,x} = 0, \tag{2.8}$$

$$f_0 = f_1 = 0, \tag{2.9}$$

$$\left.\begin{aligned} d_{j,x} &= -iuf_j + ive_{j-1}, && j=1,2,\ldots,n,\\ e_{j+1} &= se_{j-1} - \tfrac{1}{2}f_{j,x} + 2iud_j, && j=1,2,\ldots,n-1,\\ f_{j+1} &= sf_{j-1} + \tfrac{1}{2}e_{j,x} + 2ivd_{j-1}, && j=1,2,\ldots,n, \end{aligned}\right\} \tag{2.10}$$

where again we have introduced f_{n+1}. Now, by inspection of the coefficients of λ^0 and λ^{-1} one obtains the corresponding two evolutions:

$$u_t = f_{n+1}, \quad u_t = -c_p, \tag{2.11a}$$

$$v_t = -se_n, \quad v_t = sb_{p+1}, \tag{2.11b}$$

$$s_t = -ve_n, \quad s_t = vb_{p+1}. \tag{2.11c}$$

Note that both evolutions imply

$$s_0 s_{0,t} = ss_t - vv_t = 0. \tag{2.11d}$$

The method consists now in solving explicitly the recursion relations (2.2)÷6 and (2.7)÷10 for u, v, and s satisfying the constraint

$$s_0 s_{0,x} = ss_x - vv_x = 0, \tag{2.12}$$

and the following asymptotic behaviors[18]:

$$u\to 0, \quad v\to 0, \quad \text{as } |x|\to\infty, \tag{2.13}$$

$$s\to s_0, \quad \text{as } |x|\to\infty, \tag{2.14}$$

which are compatible with the reduction (2.12). Moreover we choose the particular solutions of (2.3) and (2.8) as

$$a_0 = -is \quad (\text{implies } c_0 = -2v), \tag{2.15}$$

$$d_0 = -i, \tag{2.16}$$

which are still compatible with (2.12).

By inserting now (2.15) and (2.16), respectively, into (2.6) and (2.10) we may consistently set

$$b_{2k} = d_{2k+1} = 0, \quad e_{2k} = 0,$$
$$a_{2k+1} = c_{2k+1} = 0, \quad f_{2k+1} = 0. \tag{2.17}$$

We postpone the study of the case $s_0 = 0$ to Sec. VII and first look at the case $s_0 \neq 0$ which allows us to deal with the sine–Gordon equation in laboratory coordinates.

The relations (2.2)–(2.5) and (2.7)–(2.9) together with the solutions (2.12) and (2.13) can be written

$$\begin{pmatrix} f_2 \\ -se_1 \end{pmatrix} = \begin{pmatrix} u_x + 2v \\ -2us \end{pmatrix}, \tag{2.18}$$

$$\begin{pmatrix} -c_0 \\ sb_1 \end{pmatrix} = \begin{pmatrix} 2v \\ -v_x - 2us \end{pmatrix}. \tag{2.19}$$

Moreover, the recursion relations (2.10) with a convenient choice of the constants of integration can be cast into

$$\begin{pmatrix} f_{j+1} \\ -se_j \end{pmatrix} = L^+ \begin{pmatrix} f_{j-1} \\ -se_{j-2} \end{pmatrix}, \quad j = 3,5,\ldots,n. \tag{2.20}$$

$$d_j = -iI(uf_j - ve_{j-1}), \quad j = 2,4,\ldots,n-1, \tag{2.21}$$

with

$$L^+ = \begin{pmatrix} -\frac{1}{4}D^2 + (u_x + 2v)Iu + u^2 + s & -\frac{1}{2}D + \frac{uv}{s} + (u_x + 2v)I\frac{v}{s} \\ \frac{1}{2}sD - 2suIu & s - 2suI(v/s) \end{pmatrix}, \tag{2.22}$$

and $n = 2m + 1$. For n even, one would have chosen $d_0 = 0$ and $d_1 \neq 0$.

In the same way, and using (1.7) for $s_0 \neq 0$, the system (2.6) can be written

$$\begin{pmatrix} -c_j \\ sb_{j+1} \end{pmatrix} = M^+ \begin{pmatrix} -c_{j-2} \\ sb_{j-1} \end{pmatrix}, \quad j = 2,4,\ldots,p, \tag{2.23}$$

$$a_j = i\frac{s}{s_0^2}I\left((v - w_x)b_{j-1} - \frac{1}{2}wc_{j-2}\right), \quad j = 2,4,\ldots,p, \tag{2.24}$$

with

$$M^+ = \frac{1}{s_0^2}\begin{pmatrix} s - vIw & \frac{1}{2}sD\frac{1}{s} - 2vI\left(\frac{v}{s} + \frac{1}{4}wD\frac{1}{s}\right) \\ \frac{1}{2}s(-D + wIw) & s - \frac{1}{4}sD^2\frac{1}{s} + swI\left(\frac{v}{s} + \frac{1}{4}wD\frac{1}{s}\right) \end{pmatrix}, \tag{2.25}$$

where

$$w = (v_x/s) + 2u, \tag{2.26}$$

and $p = 2k$. Before going further one should notice that (2.6) can also be written

$$L^+ \begin{pmatrix} -c_j \\ sb_{j+1} \end{pmatrix} = \begin{pmatrix} -c_{j-2} \\ sb_{j-1} \end{pmatrix}, \tag{2.27}$$

which naturally comes from the fact that recursion relations (2.6) are obtained for negative powers of λ whereas (2.10) are obtained for positive powers of λ. Equations (2.23) and (2.27) suggest the relevant fact that L^+ can be explicitly inverted and indeed one can check directly that

$$L^+M^+ = M^+L^+ = \mathbb{1}. \tag{2.28}$$

Therefore we shall write indifferently M^+ or $(L^+)^{-1}$. Using now (2.25) and defining

$$J = \begin{pmatrix} D & 2s \\ -2s & 0 \end{pmatrix}, \tag{2.29}$$

the relations (2.18) and (2.19) become

$$\begin{pmatrix} f_2 \\ -se_1 \end{pmatrix} = J\begin{pmatrix} u \\ v/s \end{pmatrix}, \tag{2.30}$$

$$\begin{pmatrix} -c_0 \\ sb_1 \end{pmatrix} = s_0JL^{-1}\begin{pmatrix} u \\ v/s \end{pmatrix}, \tag{2.31}$$

where L^{-1} is the adjoint of $(L^+)^{-1}$ with respect to the bilinear form defined by (1.8). Finally the evolution equations (1.11) for $n = 2m + 1$ and $p = 2k$ are

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = JL^m\begin{pmatrix} u \\ v/s \end{pmatrix}, \quad \begin{pmatrix} u_t \\ v_t \end{pmatrix} = s_0JL^{-(k+1)}\begin{pmatrix} u \\ v/s \end{pmatrix},$$
$$s_0 \neq 0. \tag{2.32}$$

The general IST-solvable NEE related to the spectral operator U is obtained by taking a linear combination of these equations with arbitrary time-dependent coefficients.

III. CANONICAL STRUCTURE

The tools described and developed in Ref. 17 can now be applied for the set of NEE's (2.32) to derive the properties of L and L^{-1} and to exhibit the canonical structure of these equations.

The operator J defined by (2.29) is skew symmetric with respect to the bilinear form (1.8)

$$J^+ = -J, \tag{3.1}$$

and the bracket (α being a contravariant field of T_q and β and γ covariant fields of T_q^*)

$$\{\alpha,\beta,\gamma\} = \langle\alpha, J'[J\beta]\gamma\rangle \tag{3.2}$$

satisfies the Jacobi identity

$$\{\alpha,\beta,\gamma\}+\{\beta,\gamma,\alpha\}+\{\gamma,\alpha,\beta\}=0. \tag{3.3}$$

Then J is called a cosymplectic operator and can be used to define a Poisson bracket for any two functionals F and G: $M \rightarrow C$ via

$$\{F,G\}=\left\langle \frac{\delta F}{\delta q}, J\frac{\delta G}{\delta q}\right\rangle, \tag{3.4}$$

which is skew symmetric and satisfies the Jacobi identity.

Now J and L satisfy the first coupling condition

$$JL=L^{+}J, \tag{3.5}$$

and consequently also J and L^{-1}.

Moreover we can prove (directly or by using results of Ref. 4) that L^{+} is a Nijenhuis operator (or hereditary symmetry), that is,

$$L^{+\prime}[L^{+}\alpha]\beta - L^{+}L^{+\prime}[\alpha]\beta \tag{3.6}$$

is symmetric with respect to any $\alpha,\beta \in T_q$. Now, starting from (3.6) and using the identity

$$L^{+\prime}[\alpha]=-L^{+}(M^{+\prime}[\alpha])L^{+}, \tag{3.7}$$

which immediately follows from (2.28) [for simplicity of notations we use again M^{+} instead of $(L^{+})^{-1}$] one obtains from (3.6) that

$$L^{+2}(M^{+}M^{+\prime}[\delta]\gamma - M^{+\prime}[M^{+}\delta]\gamma), \tag{3.8a}$$

with

$$\delta=L^{+}\alpha, \quad \gamma=L^{+}\beta, \tag{3.8b}$$

is symmetric in δ,γ, that is, M^{+} [or $(L^{+})^{-1}$] is a Nijenhuis operator.

Applying now the results of Ref. 4 to the reduced operator L^{+}, it is easy to prove that L^{+} and J also satisfy the second coupling condition:

$$\langle \alpha, L^{+\prime}[\gamma]J\beta\rangle - \langle \alpha, L^{+\prime}[J\beta]\gamma\rangle + \langle \beta, L^{+\prime}[J\alpha]\gamma\rangle + \langle \beta, L^{+}J^{\prime}[\gamma]\alpha\rangle - \langle \beta, J^{\prime}[L^{+}\gamma]\alpha\rangle = 0. \tag{3.9}$$

The above property together with identity (3.7) and the fact that $(L^{+})^{-1}$ is a Nijenhuis operator also give that $(L^{+})^{-1}$ and J satisfy the second coupling condition and consequently L^{+} and $(L^{+})^{-1}$ are strong symmetries[11,17] for all the equations of the hierarchies (2.32) and they generate infinitely many commuting symmetries.

Therefore, according to the general theorem by Magri,[11] since

$$f(q)=\begin{pmatrix} u \\ v/s \end{pmatrix}, \quad L^{+}f(q), \quad (L^{+})^{-1}f(q)$$

are potentials operators [see (1.10)] the NEE's in (2.32) are Hamiltonian systems with commuting flows. They can be written as follows:

$$q_t=J\frac{\delta H_m^{(+)}}{\delta q}, \quad q_t=J\frac{\delta H_k^{(-)}}{\delta q}, \tag{3.10}$$

where

$$q=\begin{pmatrix} u \\ v \end{pmatrix}, \quad \frac{\delta}{\delta q}=\begin{pmatrix} \dfrac{\delta}{\delta u} \\ \dfrac{\delta}{\delta v}+\dfrac{v}{s}\dfrac{\delta}{\delta s} \end{pmatrix}, \tag{3.11}$$

J is given in (2.29) and $H_{m,k}^{(\pm)}$ are convenient functionals to be determined.

However, by using the general method proposed in Ref. 17, it is possible to derive the Hamiltonian structure of the considered NEE's without using the hereditary symmetry property of L^{+} and $(L^{+})^{-1}$, and to relate explicitly the Hamiltonians in (3.9) to the infinitely many conserved quantities for the hierarchy.

In the next section we shall show that a convenient conserved quantity $H(q;\lambda)$ can be asymptotically expanded at the same time in powers of λ^{-1} and λ:

$$H(q;\lambda)=\sum_{n=0}^{\infty}\lambda^{\mp n}H_n^{(\pm)}(q), \quad \lambda^{\mp 1}\rightarrow 0. \tag{3.12}$$

Also, that its gradient $\delta H/\delta q$ satisfies the eigenvalue spectral equation for the operator L:

$$L\frac{\delta H}{\delta q}=\lambda^{2}\frac{\delta H}{\delta q}-i\lambda\begin{pmatrix} u \\ v/s \end{pmatrix}. \tag{3.13}$$

By inserting successively the two asymptotic expansions (3.12) into (3.13) and by equating to zero the coefficients of successive powers of λ^{-1} and λ we get the following recursion relations:

$$\frac{\delta H_0^{(\pm)}}{\delta q}=0, \quad \frac{\delta H_1^{(\pm)}}{\delta q}=\pm iL^{(-1\pm 1)/2}\begin{pmatrix} u \\ v/s \end{pmatrix}, \tag{3.14}$$

$$\frac{\delta H_{n+2}^{(\pm)}}{\delta q}=L^{\pm 1}\frac{\delta H_n^{(\pm)}}{\delta q}. \tag{3.15}$$

These can be explicitly solved furnishing

$$\frac{\delta H_{2n}^{(\pm)}}{\delta q}=0, \quad \frac{\delta H_{2m\pm 1}^{(\pm)}}{\delta q}=\pm iL^{\pm m}\begin{pmatrix} u \\ v/s \end{pmatrix}. \tag{3.16}$$

The above equations prove the Hamiltonian character of the considered NEE's and relate the Hamiltonians to the conserved quantities generated by $H(q;\lambda)$.

The general IST-solvable NEE related to the spectral operator U, in the reduced case $s_{0,t}=s_{0,x}=0$, can be written as follows:

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix}=\sum_{j=0}^{m}\mu_j^{(+)}JL^{j}\begin{pmatrix} u \\ v/s \end{pmatrix}+\sum_{j=0}^{k}\mu_j^{(-)}J \times L^{-(j+1)}\begin{pmatrix} u \\ v/s \end{pmatrix} = -i\sum_{j=0}^{m}\mu_j^{(+)}J\frac{\delta H_{2j+1}^{(+)}}{\delta q} + i\sum_{j=0}^{k}\mu_j^{(-)}J\frac{\delta H_{2j+1}^{(-)}}{\delta q}, \tag{3.17}$$

where the coefficients $\mu_j^{(\pm)}$ are arbitrary functions of t.

In the next section we shall show that the conserved densities related to the $H_n^{(\pm)}$ are local and, therefore, it follows that the NEE'S in (3.17) are purely differential. Notice that, for the Zakharov–Shabat case[7] within the reduction $r=-q$ (which leads to the sine–Gordon equation in light-cone coordinates), an explicit inverse of the recursion operator has been obtained[8] but the hierarchy of NEE's is a set of integrodifferential equations (except for the first one). The same remark can be made about the Kaup–Newell spectral problem for which an explicit inverse still exists but the NEE's are also integrodifferential.[19,20] Thus it seems that the case considered here is the first showing this nice property.

IV. CONSERVED DENSITIES

Applying the general procedure[16] in the way used in Ref. 17 we transform the spectral problem (1.5) into a Riccati equation for the projective variable

$$\mathscr{Z}(x,t) = \psi_2(x,t)/\psi_1(x,t), \tag{4.1}$$

where $(\psi_1,\psi_2)^T$ is a solution of (1.5). The equation for $\mathscr{Z}$ then reads

$$\mathscr{Z}_x = [u-(i/\lambda)v] + 2[i\lambda-(i/\lambda)s]\mathscr{Z} - [u+(i/\lambda)v]\mathscr{Z}^2. \tag{4.2}$$

Given a solution $\mathscr{Z}$ of (4.2), the quantity

$$\mathscr{H} = -i\lambda + (i/\lambda)s + [u+(i/\lambda)v]\mathscr{Z} \tag{4.3}$$

is a conserved density and, consequently,

$$H = \int_{-\infty}^{+\infty} dx\, \mathscr{H}(x,t) \tag{4.4}$$

is a constant of the motion for all NEE's obtained by solving the Lax-pair representation (1.3) for any choice of the spectral operator V including the nonreduced case.

To derive explicit λ-independent conserved densities we look for asymptotic expansions as $\lambda\to\infty$ and $\lambda\to 0$ for the solution $\mathscr{Z}^{(\pm)}$ of (4.2):

$$\mathscr{Z}^{(+)} = \sum_{n=0}^{\infty} \lambda^{-n}\mathscr{Z}_n^{(+)}, \quad \mathscr{Z}^{(-)} = \sum_{n=0}^{\infty} \lambda^{n}\mathscr{Z}_n^{(-)}, \tag{4.5}$$

which are both allowed by the structure of the Riccati equation (4.2).[9,21] Inserting (4.5) into (4.2) one gets the following recursion relations for $\mathscr{Z}^{(\pm)}$:

$$\mathscr{Z}_0^{(+)} = 0, \quad \mathscr{Z}_1^{(+)} = (i/2)u, \quad \mathscr{Z}_2^{(+)} = \tfrac{1}{4}(u_x+2v), \tag{4.6a}$$

$$\mathscr{Z}_n^{(+)} = -\frac{i}{2}\mathscr{Z}_{n-1,x}^{(+)} + s\mathscr{Z}_{n-2}^{(+)} - \frac{i}{2}u\sum_{k=0}^{n-1}\mathscr{Z}_k^{(+)}\mathscr{Z}_{n-k-1}^{(+)} + \frac{v}{2}\sum_{k=0}^{n-2}\mathscr{Z}_k^{(+)}\mathscr{Z}_{n-k-2}^{(+)}, \quad n\geqslant 3, \tag{4.6b}$$

$$\mathscr{Z}_0^{(-)} = \frac{-v}{s+s_0}, \quad \mathscr{Z}_1^{(-)} = \frac{i}{2s_0}[\mathscr{Z}_{0,x} + u(\mathscr{Z}_0^2-1)], \tag{4.7a}$$

$$\mathscr{Z}_{n+1}^{(-)} = \frac{1}{s_0}\left(\mathscr{Z}_{n-1}^{(-)} + \frac{iu}{2}\sum_{k=0}^{n}\mathscr{Z}_k^{(-)}\mathscr{Z}_{n-k}^{(-)} - \frac{v}{2}\sum_{k=1}^{n}\mathscr{Z}_k^{(-)}\mathscr{Z}_{n-k+1}^{(-)} + \frac{i}{2}\mathscr{Z}_{n,x}^{(-)}\right), \quad n\geqslant 2. \tag{4.7b}$$

Then, we define the sequences $\mathscr{H}_n^{(\pm)}$ of λ-independent conserved densities through

$$\mathscr{H}^{(+)} = -i\lambda + \sum_{n=0}^{\infty}\lambda^{-n}\mathscr{H}_n^{(+)}, \tag{4.8}$$

$$\mathscr{H}^{(-)} = \frac{is_0}{\lambda} + \sum_{n=0}^{\infty}\lambda^{n}\mathscr{H}_n^{(-)}, \tag{4.9}$$

and, using (4.3) and the recursion relations for the $\mathscr{Z}_i^{(\pm)}$,

$$\mathscr{H}_0^{(+)} = 0, \quad \mathscr{H}_1^{(+)} = is + (i/2)u^2, \tag{4.10a}$$

$$\mathscr{H}_n^{(+)} = u\mathscr{Z}_n^{(+)} + iv\mathscr{Z}_{n-1}^{(+)}, \quad n\geqslant 2, \tag{4.10b}$$

$$\mathscr{H}_0^{(-)} = -\tfrac{1}{2}[s_x/(s+s_0)], \tag{4.11a}$$

$$\mathscr{H}_n^{(-)} = u\mathscr{Z}_n^{(-)} + iv\mathscr{Z}_{n+1}^{(-)}, \quad n\geqslant 1. \tag{4.11b}$$

Note that, up to this point, we have not used the reduction condition $s_{0,t} = s_{0,x} = 0$. Therefore the two sets $\{\mathscr{H}_n^{(\pm)}\}$ furnish the conserved densities for reduced and nonreduced NEE's related to the spectral operator U. In order to relate the conserved quantities $\mathscr{H}_n^{(\pm)}$ to the Hamiltonians of the NEE's (3.17) we need to prove that both $\delta H^{(+)}/\delta q$ and $\delta H^{(-)}/\delta q$ are eigenfunctions of the operator L via Eq. (3.13).

For $\delta H^{(+)}/\delta q$ this can be easily obtained from the corresponding eigenvalue equation derived in Ref. 4 for the nonreduced case. One only needs to check that the asymptotic behavior

$$\frac{\delta H^{(+)}}{\delta r} \to i\lambda^{-1}\begin{pmatrix}0\\0\\1\end{pmatrix}, \quad \text{as } |x|\to\infty, \tag{4.12}$$

where

$$\frac{\delta}{\delta r} = \left(\frac{\delta}{\delta u},\frac{\delta}{\delta v},\frac{\delta}{\delta s}\right)^T, \tag{4.13}$$

derived in Ref. 4 [note the misprint in formula (5.8) of Ref. 4], does not change if one imposes on u, v, and s the asymptotic behaviors given in (2.13) and (2.14).

Then $\delta H^{(-)}/\delta q$ will also be eigenfunction of L if one can prove that

$$\frac{\delta H^{(+)}}{\delta q} = \frac{\delta H^{(-)}}{\delta q}, \tag{4.14}$$

or that $H^{(+)} = H^{(-)}$. To that purpose it is convenient to go back to the nonreduced case (where the fields u, v, and s are independent) and prove that

$$\frac{\delta H^{(+)}}{\delta r} = \frac{\delta H^{(-)}}{\delta r}. \tag{4.15}$$

Since it has been proved[4] that $\delta H^{(\pm)}/\delta r$ obeys to the first-order linear matrix differential equation

$$\left(\frac{\delta H^{(\pm)}}{\delta r}\right)_x = \begin{pmatrix} 0 & 2(\lambda^2-s) & -2v \\ -2\lambda^{-2}(\lambda^2-s) & 0 & 2u \\ -2\lambda^{-2}v & 2u & 0 \end{pmatrix} \times \frac{\delta H^{(\pm)}}{\delta r}, \tag{4.16}$$

a sufficient condition for (4.15) is that $\delta H^{(+)}/\delta r$ and $\delta H^{(-)}/\delta r$ possess the same behaviors as $|x|\to\infty$. From (4.9) and (4.4) one gets

$$\frac{\delta H^{(-)}}{\delta r} = \frac{i}{\lambda s_0}\begin{pmatrix}0\\-v\\s\end{pmatrix} + \sum_{n=0}^{\infty}\lambda^n\frac{\delta H_n^{(-)}}{\delta r}, \tag{4.17}$$

and one can easily verify that $\delta H_n^{(-)}/\delta r$ goes asymptotically to zero for u, v, and s obeying the boundary conditions (4.13), (4.14), and, consequently, derive that (4.15), and thus (4.14), hold.

To end this section we list below the first two nonzero Hamiltonians in the reduced case,

$$-iH_1^{(+)} = \int_{-\infty}^{+\infty} dx\left(s - s_0 + \frac{1}{2}u^2\right), \tag{4.18}$$

$$iH_1^{(-)} = \int_{-\infty}^{+\infty} dx \left[s - s_0 + \frac{1}{2}u^2 + \frac{1}{8}\frac{v_x}{s}\left(\frac{v_x}{s} + 4u\right)\right], \tag{4.19}$$

and two examples of IST-solvable NEE's (3.17): (i) $m = k = 0, \mu_0^{(+)} = 1, \mu_0^{(-)} = s_0, \mu_j^{(\pm)} = 0\ (j \geqslant 1)$,

$$u_t = u_x + 4v, \tag{4.20}$$

$$v_t = -v_x - 4us; \tag{4.21}$$

and (ii) $m = k = 1, \mu_1^{(+)} = 1, \mu_1^{(-)} = -s_0^3$, $\mu_j^{(\pm)} = 0\ (j = 0, j \geqslant 2)$,

$$u_t = -\tfrac{1}{4}u_{xxx} + \tfrac{3}{2}u^2 u_x + 3u_x s - \tfrac{3}{4}v_x(s_x/s), \tag{4.22}$$

$$v_t = -\frac{1}{4}s\left(\frac{v_x}{s}\right)_{xx} + \frac{3}{4}uv_x\left(\frac{v_x}{s} + 2u\right) + 3sv_x + \frac{1}{8}v_x\left(\frac{v_x}{s}\right)^2. \tag{4.23}$$

V. THE SINE–GORDON EQUATION AND GENERALIZATIONS

We are now able to examine the sine–Gordon equation (SG) (1.1) within the scheme developed in the preceeding sections. Indeed, introducing the new function $\omega(x,t)$ via

$$v = is_0 \sin\omega, \tag{5.1}$$

$$s = s_0 \cos\omega, \tag{5.2}$$

Eq. (4.21) gives $u(x,t)$ in terms of $\omega(x,t)$ through

$$u = -(i/4)(\omega_x + \omega_t), \tag{5.3}$$

and thus (4.20) reads

$$\omega_{tt} - \omega_{xx} + 16s_0 \sin\omega = 0 \tag{5.4}$$

[which is (1.1) for $s_0 = \frac{1}{16}$]. Note that (1.5) is then equivalent to (1.4) via a constant gauge transformation.

So we see that the sine–Gordon equation in laboratory coordinates appears as a member of the (purely differential) hierarchy of NEE's (3.17) under the choice (5.1) and (5.2) for the solution of the reduction (1.7). The first two conserved quantities $H_1^{(\pm)}$ can be connected to the physical conserved quantities for the sine–Gordon equation[9,10]: the momentum p and the energy E given by

$$p = \int_{-\infty}^{+\infty} \omega_x \omega_t \, dx, \tag{5.5}$$

$$E = \int_{-\infty}^{+\infty} \left(\frac{1}{2}\omega_t^2 + \frac{1}{2}\omega_x^2 + 1 - \cos\omega\right) dx. \tag{5.6}$$

In fact, (4.18) and (4.19) give for $s_0 = \frac{1}{16}$:

$$-iH_1^{(+)} = -\frac{1}{16}(E + p), \tag{5.7}$$

$$iH_1^{(-)} = -\frac{1}{16}(E - p) \tag{5.8}$$

(note that, as for E, see Ref. 10, $H_1^{(+)}$ has to be renormalized by defining $\mathscr{H}_{1R}^{(+)} = \mathscr{H}_1^{(+)} - is_0$).

A "second-order sine–Gordon equation" can be obtained by decoupling the two equations (4.22) and (4.23) by use of (5.1) and (5.2). Equation (4.23) then becomes

$$u^2 + \frac{i}{2}\omega_x u - \frac{1}{12}\omega_x^2 + 2\cos\omega - \frac{1}{6}\frac{\omega_{xxx}}{\omega_x} - \frac{2}{3}\frac{\omega_t}{\omega_x} = 0, \tag{5.9}$$

which can be explicitly solved for u in terms of $\omega(x,t)$. Inserting the solution in (4.22) we get a second-order sine–Gordon equation:

$$\begin{aligned} -\tfrac{1}{2}\omega_{xt} \pm \Delta_t = {} & \tfrac{1}{8}\omega_{xxxx} \mp \tfrac{1}{4}\Delta_{xxx} + \tfrac{3}{64}\omega_x^2\omega_{xx} \mp \tfrac{3}{32}\Delta_x\omega_x^2 \\ & - \tfrac{3}{8}\Delta(\Delta \mp \omega_x)[-(\omega_{xx}/2) \pm \Delta_x] \\ & - \tfrac{3}{2}\omega_{xx}\cos\omega \pm \Delta_x\cos\omega - \tfrac{3}{2}\omega_x^2\sin\omega, \end{aligned} \tag{5.10}$$

where

$$\Delta^2 = 8\cos\omega - \tfrac{1}{12}\omega_x^2 - \tfrac{2}{3}(\omega_{xxx}/\omega_x) - \tfrac{8}{3}(\omega_t/\omega_x). \tag{5.11}$$

Let us now make two remarks.

(i) Although higher-order SG equations [that is, any equation of the hierarchy (3.17) for v and s given by (5.1) and (5.2)] may not be an explicit one-field equation, the inverse spectral transform allows us to give explicit soliton solutions. However, one would need to solve the direct and inverse spectral problems for (1.5) which will be the purpose of a future work.

(ii) The transformation (5.1),(5.2) explicitly decouples both equations (4.20) and (4.21) and (4.22) and (4.23) but furnishes different relations between u and ω. In general it is not possible to decouple the equations explicitly. That explains why, in order to get a hierarchy containing the sine–Gordon equation, one needs to study NEE's with two coupled fields.

A generalized IST-solvable sine–Gordon equation is obtained by choosing $k = m = 0$ in (3.17) and

$$\mu_0^+(t) = \alpha(t), \quad \mu_0^{(-)}(t) = s_0\beta(t), \quad \mu_j^{(\pm)} = 0 \quad (j \geqslant 1), \tag{5.12}$$

and by setting

$$v = is_0\sin\omega, \quad s = s_0\cos\omega. \tag{5.13}$$

Then one gets the following relation that solves one of the two coupled NEE's (for $\alpha + \beta \neq 0$):

$$u = -[i/2(\alpha + \beta)](\omega_t + \beta\omega_x). \tag{5.14}$$

Also one gets the equation

$$\begin{aligned} \omega_{tt} - \alpha\beta\,\omega_{xx} - (\alpha - \beta)\omega_{xt} + 4s_0(\alpha + \beta)^2\sin\omega \\ = (\alpha + \beta)f(x,t), \end{aligned} \tag{5.15}$$

$$f(x,t) = -\left(\frac{1}{\alpha + \beta}\right)_t \omega_t - \left(\frac{\beta}{\alpha + \beta}\right)_t \omega_x, \tag{5.16}$$

which is an exactly solvable equation for any choice of the arbitrary functions $\alpha(t)$ and $\beta(t)$. The one-soliton solution of (5.15) can be obtained either by using the Bäcklund transformation (see Sec. VI) or else from the standard kink solution of the sine–Gordon equation by noticing that, under the change of coordinates,

$$(x,t) \to (y(x,t), \tau(x,t)), \tag{5.17a}$$

$$y = x + \int_0^t d\zeta \, \frac{1}{2}[\alpha(\zeta) - \beta(\zeta)], \tag{5.17b}$$

$$\tau = \int_0^t d\zeta \, \frac{1}{2}[\alpha(\zeta) + \beta(\zeta)], \tag{5.17c}$$

Eq. (5.15) reduces to

$$\omega_{\tau\tau} - \omega_{yy} + 16 s_0 \sin\omega = 0. \tag{5.18}$$

Then the one solition solution of (5.15) can be written

$$\omega(x,t) = 4\tan^{-1}\exp\Big[\frac{k}{2}\Big(1+\frac{16 s_0}{k^2}\Big)x + \frac{k}{2}\int_0^t d\xi\Big(\alpha(\xi) - \frac{16 s_0}{k^2}\beta(\xi)\Big)\Big], \tag{5.19}$$

where k is an arbitrary parameter (it is the parameter of the Bäcklund transformation).

Although the solutions of the generalized sine–Gordon equation (5.15) can be obtained from those of (5.18) through the mapping (5.17), it may be useful, depending on the choices of $\alpha(t)$ and $\beta(t)$, to obtain directly the solutions of (5.15) (for instance to study the asymptotic behaviors in t or x) or to look at the evolution of the spectral data associated with solutions of (5.15) instead of (5.18), more especially if the equation (5.15), understood as a perturbed sine–Gordon equation,[2,22,23] has a physical interest.

VI. BÄCKLUND TRANSFORMATION AND SOLITON SOLUTIONS

The Bäcklund transformation (BT) for the SG equation has been the first one discovered for a solition equation, just one century ago.[24]

Afterwards the BT has been found for many different hierarchies of IST-solvable NEE's related to many different spectral problems.

The resulting common feature is that the so-called x-component of the BT has a universal character, in the sense that it holds, unchanged in form, for all the equations in the hierarchy, while the so-called t-component has a specific form for any considered equation.

However, in the case considered in this paper, the x-component of the BT for the SG equation found by Bäcklund is *not* the x-component of the BT for all the equations in the hierarchy described in the previous sections, because the equations in the hierarchy couple two independent fields, while the SG equation is a one-field evolution equation.

Therefore, we need to find explicitly the BT for the hierarchy and to verify that in the specific case of SG equation it reduces to the already known BT.

Let us consider a nonsingular 2×2 matrix gauge transformation B of the eigenmatrix Ψ in the spectral equation (1.5) and (1.6),

$$\overline{\Psi} = B\Psi. \tag{6.1}$$

Here $B = B(\bar{q},q;\lambda)$ is considered as a functional of

$$q = \begin{pmatrix} u \\ v \end{pmatrix},$$

of the transformed field

$$\bar{q} = \begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix},$$

and of the spectral parameter λ. $\overline{\Psi}$ is required to satisfy the spectral equations

$$\overline{\Psi}_x = \overline{U}\,\overline{\Psi}, \tag{6.2}$$

$$\overline{\Psi}_t = \overline{V}\,\overline{\Psi}, \tag{6.3}$$

with spectral operators $\overline{U}$ and $\overline{V}$ obtained by substituting q with $\bar{q}$ in U and V.

It is easy to verify that B must satisfy the matrix differential equations

$$B_x = \overline{U}B - BU, \tag{6.4}$$

$$B_t = \overline{V}B - BV. \tag{6.5}$$

By cross differentiating one gets

$$B_{xt} - B_{tx} = (\overline{U}_t - \overline{V}_x + [\overline{U},\overline{V}])B - B(U_t - V_x + [U,V]), \tag{6.6}$$

and, consequently, if the matrix B is a solution of the set of equations (6.4) and (6.5) and q is a solution of the evolution equation defined by the Lax representation

$$U_t - V_x + [U,V] = 0, \tag{6.7}$$

then $\bar{q}$ is a solution of the evolution equation induced by

$$\overline{U}_t - \overline{V}_x + [\overline{U},\overline{V}] = 0. \tag{6.8}$$

The gauge B is called a Bäcklund gauge and it is said to generate the x-component of the BT via Eq. (6.4) and the t-component of the BT via Eq. (6.5).

To find the explicit form of B one must first solve Eq. (6.4) and successively determine the time evolution of the constants of integration by imposing that the resulting B asymptotically satisfy Eq. (6.5), say at $x = -\infty$,

$$B_t^{(-)} = \overline{V}^{(-)}B^{(-)} - B^{(-)}V^{(-)}. \tag{6.9}$$

In Ref. 20 it has been proved that this procedure uniquely determines the Bäcklund gauge B.

Let us choose B to have a simple pole at $\lambda = \infty$,

$$B(\bar{q},q;\lambda) = B_0(\bar{q},q)\lambda + B_1(\bar{q},q), \tag{6.10}$$

with

$$B_j = \alpha_j\sigma_3 + \beta_j\sigma_1 + \gamma_j\sigma_2 + \delta_j\mathbb{1} \quad (j=0,1). \tag{6.11}$$

By equating to zero the coefficients of λ and λ^0 in (6.4) and by introducing for convenience α_j,β_j for $j=-1,2$ and γ_j,δ_j for $j=-1,2,3$ we get the recursion relations

$$\alpha_{-1} = \beta_{-1} = \gamma_{-1} = \delta_{-1} = 0, \tag{6.12a}$$

$$\alpha_{jx} = i(\bar{u}+u)\gamma_j + i(\bar{s}-s)\delta_{j-1} + i(\bar{v}+v)\beta_{j-1}, \tag{6.12b}$$

$$\beta_{jx} = -2\gamma_{j+1} + (\bar{u}-u)\delta_j + (\bar{s}+s)\gamma_{j-1} - i(\bar{v}+v)\alpha_{j-1}, \tag{6.12c}$$

$$\gamma_{jx} = 2\beta_{j+1} - i(\bar{u}+u)\alpha_j - (\bar{s}+s)\beta_{j-1} - (\bar{v}-v)\delta_{j-1}, \tag{6.12d}$$

$$\delta_{jx} = (\bar{u}-u)\beta_j + i(\bar{s}-s)\alpha_{j-1} - (\bar{v}-v)\gamma_{j-1}, \quad (j=0,1,2), \tag{6.12e}$$

together with the constraints

$$\alpha_2 = \beta_2 = 0, \tag{6.13a}$$

$$\gamma_j = \delta_j = 0 \quad (j=2,3). \tag{6.13b}$$

The constraints $\alpha_2 = 0$, $\gamma_2 = 0$, $\beta_3 = 0$ via Eqs. (6.12b) and (6.12d) and the constraints $\delta_2 = 0$, $\beta_2 = 0$, $\gamma = 0$ via Eqs. (6.12c) and (6.12e) furnish, respectively, the equations

$$\delta_1 = -[(\bar v + v)/(\bar s - s)]\,\beta_1, \tag{6.14a}$$

$$\delta_1 = -[(\bar s + s)/(\bar v - v)]\,\beta_1, \tag{6.14b}$$

and the equations

$$\alpha_1 = -[i(\bar v - v)/(\bar s - s)]\,\gamma_1, \tag{6.15a}$$

$$\alpha_1 = -i[(\bar s + s)/(\bar v + v)]\,\gamma_1. \tag{6.15b}$$

If one excludes the uninteresting case $\alpha_1 = \beta_1 = \gamma_1 = \delta_1 = 0$, the two sets of equations are compatible iff $s_0^2 = \bar s_0^2$. We choose $s_0 = \bar s_0$.

Then, the recursion relations (6.12) can be explicitly solved and one gets the B_j's ($j = 0,1$) and four differential equations in the x-variable.

In order to get the right number of differential equations (one for any independent field variable) one must choose $\alpha_0 = 0$ or $\delta_0 = 0$.

In the first case ($\alpha_0 = 0$) one gets the following Bäcklund gauge:

$$B = \delta_0\Big(\lambda \mathbb{1} - \frac{i}{2}\frac{\bar v - v}{\bar s - s}(\bar u - u)\sigma_3 + \frac{1}{2}(\bar u - u)\sigma_2\Big), \tag{6.16}$$

with δ_0 an arbitrary function of t, and the two differential equations

$$[[(\bar v - v)/(\bar s - s)](\bar u - u)]_x = -(\bar u^2 - u^2) - 2(\bar s - s), \tag{6.17}$$

$$(\bar u - u)_x = -[(\bar v - v)/(\bar s - s)](\bar u^2 - u^2) - 2(\bar v - v), \tag{6.18}$$

which are to be interpreted as the x-component of the searched BT.

In the second case ($\delta_0 = 0$) one gets a BT that can be obtained by composing the previous BT with the trivial BT defined by $v \to -v$, $u \to -u$.

It is convenient to express v,s and $\bar v,\bar s$ in terms of only one field ω and $\bar\omega$, respectively,

$$v = is_0 \sin\omega, \tag{6.19a}$$

$$s = s_0 \cos\omega, \tag{6.19b}$$

$$\bar v = is_0 \sin\bar\omega, \tag{6.20a}$$

$$\bar s = s_0 \cos\bar\omega, \tag{6.20b}$$

and to rewrite Eqs. (6.17) and (6.18) as follows:

$$(\bar\omega_x + \omega_x)(\bar u - u) = 2i(\bar u^2 - u^2) - 8is_0 \sin\frac{\bar\omega + \omega}{2}\sin\frac{\bar\omega - \omega}{2}, \tag{6.21}$$

$$\bar u_x - u_x = i(\bar u^2 - u^2)\cot\frac{\bar\omega + \omega}{2} - 4is_0\cos\frac{\bar\omega + \omega}{2}\sin\frac{\bar\omega - \omega}{2}. \tag{6.22}$$

From these two equations it is easy to obtain

$$[(\bar u - u)/\sin[(\bar\omega + \omega)/2]]_x = 0. \tag{6.23}$$

Therefore, the x-component of the BT is given by the explicitly integrated equation

$$\bar u - u = -i(k/2)\sin[(\bar\omega + \omega)/2], \tag{6.24}$$

with k an arbitrary function of t, and by the differential equation in $\bar\omega$,

$$\bar\omega_x + \omega_x = 2i(\bar u + u) + 16(s_0/k)\sin[(\bar\omega - \omega)/2]. \tag{6.25}$$

The time behaviors of k and δ_0 are determined by inserting into (6.9) the asymptotic expression of B at $x = -\infty$

$$B^{(-)} = \delta_0\lambda\mathbb{1} + (i/4)\delta_0 k\sigma_3, \tag{6.26}$$

which is obtained by using the asymptotic behaviors $u,\bar u,\omega,\bar\omega \to 0$ as $|x| \to \infty$ and Eq. (6.24).

In order to compute the asymptotic behavior of V at $x = -\infty$, let us note that, by using (2.30), (2.20) and (2.31), (2.23), the d_j and a_j in Eqs. (2.21) and (2.24) can be cast into the form

$$d_j = -iI\left[(u \quad v/s)JL^{(j-1)/2}\begin{pmatrix} u \\ v/s\end{pmatrix}\right], \quad j = 2,4,\ldots,2m, \tag{6.27}$$

and

$$a_j = -isI\left[\left(L^{-1}\begin{pmatrix} u \\ v/s\end{pmatrix}\right)^T JL^{-j/2}\begin{pmatrix} u \\ v/s\end{pmatrix}\right], \quad j = 2,4,\ldots,p. \tag{6.28}$$

From (3.16) we get for d_j and a_j at $x = -\infty$,

$$d_j \to -\frac{i}{2}\left\langle\frac{\delta H_1^{(+)}}{\delta q}, J\frac{\delta H_j^{(+)}}{\delta q}\right\rangle = -\frac{i}{2}\{H_1^{(+)}, H_j^{(+)}\}, \quad x \to -\infty \quad (j = 2,4,\ldots,2m), \tag{6.29}$$

$$a_j \to -\frac{i}{2}s_0\left\langle\frac{\delta H_1^{(-)}}{\delta q}, J\frac{\delta H_{j-1}^{(-)}}{\delta q}\right\rangle = -\frac{i}{2}s_0\{H_1^{(-)}, H_{j-1}^{(-)}\}, \quad x \to -\infty \quad (j = 2,4,\ldots,p), \tag{6.30}$$

and, consequently, because the Hamiltonians $H^{(\pm)}$ are in involution with respect to the Poisson bracket $\{\ ,\ \}$, d_j and $a_j \to 0$ as $x \to -\infty$ for any $j \neq 0$.

The $\bar d_j$ and $\bar a_j$ in $\bar V$ have, of course, the same asymptotic value.

From the recursion relations (2.18), (2.20) and (2.19), (2.23) one also obtains that f_j, e_j and c_j, b_j, together with their counterparts in $\bar V$, asymptotically vanish for any j.

Then, $V^{(-)}$ and $\bar V^{(-)}$ can be easily computed and, by taking into account Eq. (6.9) and $s_0 = \bar s_0$, we finally derive that *δ_0 and k must be independent on t.*

This result can be extended to any different choice of the constants of integration in the recursion relations (2.18)÷(2.25) and, therefore, to the more general NEE's in (3.17).

In order to get the explicit form of the t-component of the BT, we equate the coefficients of λ^0 in (6.5). It results that

$$\bar\omega_t + \omega_t = \frac{4i}{k}\csc\frac{\bar\omega + \omega}{2}\,\mathrm{Tr}[\sigma_3(\bar W_p - W_p)] + i\,\mathrm{Tr}[\sigma_1(\bar V_n + V_n)], \tag{6.31}$$

$$\bar\omega_t + \omega_t = \frac{4i}{k}\sec\frac{\bar\omega + \omega}{2}\,\mathrm{Tr}[\sigma_2(\bar W_p - W_p)] + i\,\mathrm{Tr}[\sigma_1(\bar V_n + V_n)]. \tag{6.32}$$

Consistency requires that these two equations be dependent. Any of them can be considered as the first differential equation of the t-component of the BT.

The previously obtained equation

$$k_t = 0 \tag{6.33}$$

must be considered as the second differential equation of the t-component of the BT.

The composition of two Bäcklund gauges $B(q_3,q_1;\lambda;k_2)B(q_1,q_0;\lambda;k_1)$—or for short $B_{31}^{(2)}B_{10}^{(1)}$—is a Bäcklund gauge that transforms the field q_0 into q_3. This can be verified by computing directly its x- and t-derivatives by using Eqs. (6.4) and (6.5) written successively for $B_{31}^{(2)}$ and $B_{10}^{(1)}$.

Moreover, the Bäcklund gauges satisfy the so-called permutability theorem

$$B_{31}^{(2)}B_{10}^{(1)} = B_{32}^{(1)}B_{20}^{(2)}. \tag{6.34}$$

Because the Bäcklund gauges are uniquely determined by their asymptotic values at $x = -\infty$, the theorem (6.34) follows from the corresponding trivial identity at $x = -\infty$.

From Eq. (6.24) written successively for the BT's, $B_{31}^{(2)}$, $B_{10}^{(1)}$, $B_{32}^{(1)}$, and $B_{20}^{(2)}$, and the trivial identity, $(u_3 - u_1) + (u_1 - u_0) = (u_3 - u_2) + (u_2 - u_0)$, one gets the double BT for ω

$$\tan\frac{\omega_3 - \omega_0}{4} = \frac{k_1 + k_2}{k_1 - k_2}\tan\frac{\omega_1 - \omega_2}{4}, \tag{6.35}$$

which furnishes ω_3 in terms of ω_1, ω_2, and ω_0 by means of pure algebraic and elementary transcendental operations.

From the same equation (6.24) one also gets the double BT for u,

$$\left[1 + \left(\frac{k_1 + k_2}{k_1 - k_2}\right)^2 \tan^2\frac{\omega_1 - \omega_2}{4}\right](u_3 - u_0) = \left(1 + \tan^2\frac{\omega_1 - \omega_2}{4}\right)\frac{k_1 + k_2}{k_1 - k_2}(u_1 - u_2), \tag{6.36}$$

which furnishes u_3 in terms of u_0, u_1, u_2, ω_1, and ω_2.

The found BT and double BT in the SG case reduce, as expected, to the well-known BT discovered by Bäcklund[24] and the double BT discovered by Bianchi.[25]

The two differential equations in the x-variable (6.24) and (6.25), solved for $u = 0$ and $\omega = 0$, furnish the one-soliton solution for all the equations in the hierarchy

$$\bar{\omega} = 4\tan^{-1}\exp\left[\frac{k}{2}\left(1 + \frac{16s_0}{k^2}\right)x + \varphi(k,t)\right]. \tag{6.37}$$

The dispersion function $\varphi(k,t)$ must be determined by inserting the found solution $\bar{\omega}$ into the differential equation in the t-variable (6.31) or (6.32).

For instance for the generalized SG equation (5.15) one gets

$$\left(1 + \frac{16s_0}{k^2}\right)\bar{\omega}_x - \left(\alpha - \frac{16s_0}{k^2}\beta\right)\bar{\omega}_t = 0, \tag{6.38}$$

and, consequently, the one-solition solution (5.19).

Finally, let us note that, because the double BT for ω (6.35) is the same as for the SG equation, the form of the N-soliton solution is the same for all the equations in the hierarchy. Only the dispersion functions $\varphi(k,t)$ depend sensitively on the specific chosen equation.

VII. THE CASE $v = s$, LIOUVILLE EQUATION AND GENERALIZATIONS

Now we are to examine briefly the class of reductions corresponding to $s_0 = 0$ in (1.7) and we choose for simplicity $s = v$ (the case $s = -v$ can be studied through the same procedure). As noted before, this reduction has already been studied[6] in the case where $W_j = 0$ for all j in (1.6) and a degenerated Hamiltonian structure has been found.

Using the results obtained in the previous sections one first realizes that the operator L^+ given by (2.22) possesses the following properties in the case $v = s$ where it is written L_0^+.

(i) It is defined and can be factorized under the form

$$L_0 = J_0\Lambda_0, \tag{7.1}$$

with

$$J_0 = \begin{pmatrix} D & 2v \\ -2v & 0 \end{pmatrix}, \quad \Lambda_0 = \begin{pmatrix} u & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} Iu & I \\ Iu & I \end{pmatrix} - \frac{1}{4}\begin{pmatrix} D & 2 \\ -2 & 0 \end{pmatrix}. \tag{7.2}$$

(ii) It is singular since, for any asymptotically vanishing function $f(x)$ we have

$$L_0^+\begin{pmatrix} f \\ -\frac{1}{2}f_x - uf \end{pmatrix} = 0. \tag{7.3}$$

Thus, in opposition to the case when $s_0 \neq 0$, the operator L_0^+ no longer is invertible and the class of evolution equations that one would obtain for $p \geqslant 1$ in (1.6) would not be local. So we restrict our study to the case $p = 0$ and obtain the following class of solvable NEE's:

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = J_0\left(\sum_{j=0}^{m}\mu_j L_0^j + \nu N\right)\begin{pmatrix} u \\ 1 \end{pmatrix}, \tag{7.4}$$

where μ_j and ν are arbitrary functions of t, L_0 is the adjoint of L_0^+, and

$$N = \begin{pmatrix} 1 & (1/2v)Dv \\ -(1/2v)D & 1 - (1/4v)D(1/v)Dv \end{pmatrix}, \tag{7.5}$$

for the asymptotic behaviors

$$u \to 0, \quad v \to a \neq 0 \quad (a_x = a_t = 0), \quad \text{for } |x| \to \infty. \tag{7.6}$$

Note that for $a = 0$ and $\nu \neq 0$, although (7.4) cannot be written in the standard canonical form,

$$J_0 N\begin{pmatrix} u \\ 1 \end{pmatrix}$$

can still be defined since it reads

$$J_0 N\begin{pmatrix} u \\ 1 \end{pmatrix} = \begin{pmatrix} 2v \\ -v_x - 2uv \end{pmatrix}. \tag{7.7}$$

Moreover, the relation (3.16) for the $(+)$ sign still holds and thus all the equations in the hierarchy (7.4) are Hamiltonian systems (with commuting flows), that is,

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = -i\sum_{j=0}^{m}\mu_j J_0\frac{\delta H_{2j+1}^{(+)}}{\delta q} + i\nu J_0\frac{\delta H_1^{(-)}}{\delta q}, \tag{7.8}$$

where the Hamiltonians $H_n^{(+)}$ and $H_1^{(-)}$ are obtained from those given in Sec. IV by making $v = s$.

Another important result is that, in the case $v = s$, and $p = 0$, we do not have to choose the particular solution (2.15) for $a_0(x,t)$. Indeed, the relations (2.4) and (2.5) are automatically consistent for any a_0 and thus, the hierarchy (7.4) can be generalized to involve an arbitrary function of x and t as

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = J_0 \sum_{j=0}^{m} \mu_j L_0^j \begin{pmatrix} u \\ 1 \end{pmatrix} + i \begin{pmatrix} 2a_0 \\ -a_{0,x} - 2ua_0 \end{pmatrix} \tag{7.9}$$

(note that the right-hand side vector functional of a_0 belongs to the null space of L_0^+ and of Λ_0). As an instance of such new NEE we set

$$m = 0, \quad \mu_0 = 1, \tag{7.10}$$

and (7.9) writes

$$u_t = u_x + 2v + 2ia_0, \tag{7.11a}$$

$$v_t = -2uv - ia_{0,x} - 2iua_0. \tag{7.11b}$$

The above set of couplet NEE's can be cast into an equivalent system of two Riccati equations:

$$u^2 + u_t - 2ia_0 = f(x+t), \tag{7.12a}$$

$$u^2 + u_x + 2v = f(x+t), \tag{7.12b}$$

where f is an arbitrary function of $(x+t)$.

Any specific functions of u, v, and a_0 can be arbitrarily given. If a_0 (or v) is arbitrarily chosen, the problem of finding the solution of (7.11) reduces, via (7.12), to a linear problem. Otherwise, the system can be solved by using the IST method.

To illustrate the found system of solvable NEE, let us choose the following transformation:

$$v(x,t) = \tfrac{1}{16}\exp[\phi(x,t)],$$

$$a_0(x,t) = iv(x,t)[1 - 2\rho(x,t)]. \tag{7.13}$$

It decouples (7.11) as follows:

$$u = -(1/4\rho)(\phi_t - (1-2\rho)\phi_x + 2\rho_x), \tag{7.14}$$

$$\phi_{tt} + (1-2\rho)\phi_{xx} + 2(\rho-1)\phi_{xt} + \rho^2 \exp[\phi]$$
$$= (1/\rho)(\rho_t - \rho_x)(\phi_t - \phi_x + 2\rho_x) - 2(\rho_{xt} - \rho_{xx}). \tag{7.15}$$

In the last equation, any specific function of ϕ and ρ can be arbitrarily chosen. In particular, for $\rho = 1$ one gets the Liouville equation for $\rho = \frac{1}{2}$ the equation studied in Refs. 6 and 23.

ACKNOWLEDGMENTS

Two of us (M. B. and F. P.) are grateful to Professor P. C. Sabatier and the Laboratoire de Physique Mathématique for kind hospitality. We would also like to thank Professor J. M. Brami for useful comments. One of us (J. L.) acknowledges enlightening discussions with Professor E. Jen.

This work has been done as part of the program RCP 264 "Etude Interdisciplinaire des Problèmes Inverses" and has been supported in part by the Instituto Nazionale di Física Nucleare, Bari, Italy.

[1]See, for instance, Calogero and A. Degasperis, "Spectral transform and solitons 1," in *Studies in Applied Mathematics* (North Holland, Amsterdam, 1982), Vol 13; M. J. Ablowitz and J. Segur, "Solitons and the inverse scattering transform," in *Studies Applied Mathematics* (SIAM, Philadelphia, 1981).
[2]See, for instance, A. R. Bishop, J. A. Krumhansl, and S. E. Trullinger, Physica D **1**, 1 (1980) and the references quoted there.
[3]G. L. Lamb, Jr., *Elements of Soliton Theory* (Wiley, New York, 1979); M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Phys. Rev. Lett. **30**, 1262 (1973); L. A. Takhtadzhyan and L. D. Faddeev, Theor. Math. Phys. USSR **21**, 1046 (1974); D. J. Kaup, Stud. Appl. Math. **54**, 165 (1975).
[4]M. Boiti and G. Z. Tu, Nuovo Cimento B **75**, 145 (1983).
[5]A compatible reduction is a reduction that is conserved by any of the evolutions of the hierarchy, see, for instance, F. Calogero and A. Degasperis, J. Math. Phys. **22**, 23 (1981); A. Degasperis and J. JP. Leon, Nuovo Cimento B **78**, 129 (1983).
[6]G. Z. Tu, Phys. Lett. A **94**, 340 (1983).
[7]M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974).
[8]R. Sasaki and R. K. Bullough, Proc. R. Soc. London Ser. A **376**, 401 (1981); see also Ref. 11; R. K. Dodd and R. K. Bullough, Proc. R. Soc. London Ser. A **351**, 499 (1976).
[9]L. A. Takhtadzhyan and L. D. Faddeev, Proc. Steklov Inst. Math. **3**, 277 (1979).
[10]M. G. Forest and D. W. McLaughlin, Stud. Appl. Math. **68**, 11 (1983).
[11]F. Magri, "A geometrical approach to the nonlinear solvable equations," in *Lecture Notes in Physics*, edited by M. Boiti, F. Pempinelli, and G. Soliani (Springer-Verlag, Berlin, 1980), Vol. 120.
[12]B. Fuchssteiner, Nonlinear Anal. Theory Math. Appl. **3**, 849 (1979).
[13]A. S. Fokas and B. Fuchssteiner, Lett. Nuovo Cimento **28**, 299 (1980); and Physica D **4**, 47 (1981).
[14]R. L. Anderson and A. S. Fokas, J. Math. Phys. **23**, 1066 (1982).
[15]P. J. Olver, J. Math. Phys. **18**, 1212 (1977); and Math. Proc. Cambridge Philos. Soc. **88**, 71 (1980).
[16]T. M. Alberty, T. Koikawa, and R. Sasaki, Physica D **15**, 43 (1982).
[17]M. Boiti, F. Pempinelli, and G. Z. Tu, "Canonical structure of soliton equations via isospectral eigenvalue problem," to appear in Nuovo Cimento B (1984).
[18]A more precise characterization of the behavior is needed for studying direct and inverse spectral problems for (1.5). In general it is sufficient for instance that the components decay faster than exponentials.
[19]V. S. Gerdjikov, M. I. Ivanov, and P. P. Kulish, Teor. Math. Fiz. **44**, 342 (1980).
[20]M. Boiti, C. Laddomada, F. Pempinelli, and G. Z. Tu, Physica D **9**, 425 (1983).
[21]C. Tian and Y. S. Li, "Prolongation Methods applied to nonlinear evolution equations," preprint, Heifei, 1983.
[22]For example, M. B. Fogel, S. E. Trullinger, A. R. Bishop, and J. A. Krumhansl, Phys. Rev. B **15**, 1578 (1977); D. W. McLaughlin and A. C. Scott, Phys. Rev. A **18**, 1652 (1978); A. R. Bishop, K. Fesser, P. S. Lomdahl, and S. E. Trullinger, Physica D **7**, 259 (1983) and the references quoted there.
[23]J. JP. Leon, G. Reinisch, and J. C. Fernandez, Phys. Rev. B **27**, 5817 (1983).
[24]A. V. Bäcklund, Lund Univ. Arsskr. Avd. 2 **19** (1883).
[25]L. Bianchi, *Lezioni di Geometria Differenziale*, edited by Enrico Spoerri (Pisa, 1902), Vol. II, p. 418.

Quantum solitons of the nonlinear Schrödinger field as Galilean particles

L. Martínez Alonso
Departamento de Métodos Matemáticos de la Física, Facultad de Ciencias Físicas, Universidad Complutense, Madrid-3, Spain

(Received 1 November 1983; accepted for publication 27 January 1984)

The quantum nonlinear Schrödinger field with attractive coupling is considered through the quantum inverse scattering method. The algebra of scattering data operators is formulated in terms of an infinite family of independent boson fields. It is shown that the unitary representation induced by the Galilean invariance of the model is equivalent to a direct product of simple representations. As a consequence, the quantum solitons turn out to be associated with representations describing quantum elementary Galilean particles. The characterization of quantum solitons as stable asymptotic fragments arising in the scattering of the fundamental bosons of the model is also analyzed.

PACS numbers: 02.30.Jr, 11.30.Na

I. INTRODUCTION

We have recently analyzed some of the nonlinear wave equations solvable through the inverse scattering method (ISM) from a group-theoretical point of view.[1,2] Our main interest was to provide a precise mathematical characterization of the analogies between solitons and classical particles. Classical particles are mathematically described by canonical realizations of invariance Lie groups. Thus we looked for a description of the degrees of freedom associated with solitons in terms of canonical realizations of Lie groups on finite-dimensional phase spaces. This description was found by using the structure of independent degrees of freedom provided by the scattering data variables of the ISM.

In the last years great progress has been made in studying complete integrable quantum models by means of an appropriate formulation of the ISM, the so-called quantum inverse scattering method (QISM).[3–5] The QISM applies to the quantum versions of several of the nonlinear wave equations previously solved by the ISM. Moreover, at the quantum level some aspects of the ISM acquire a simpler and more intuitive meaning.[3] In the present paper we will be concerned with the quantum nonlinear Schrödinger equation with attractive coupling

$$i\psi_t = -\psi_{xx} + 2c\psi^\dagger\psi\psi, \quad c<0, \tag{1.1}$$

where $\psi(x)$ denotes the one-dimensional boson field verifying the canonical commutation relations

$$[\psi(x),\psi(y)] = 0, \quad [\psi(x),\psi^\dagger(y)] = \delta(x-y). \tag{1.2}$$

This model describes a field of nonrelativistic bosons of mass $\frac{1}{2}$ interacting via the attractive potential $c\delta(x-y)$, in one space dimension. The exact integrability of (1.1) was proved[6,7] by means of an explicit diagonalization of the corresponding Hamiltonian through the use of Bethe ansatz solutions.[8] Nevertheless the rich algebraic structure underlying this model was revealed by the application of the (QISM).[9–12]

The quantum nonlinear Schrödinger equation is a Galilean-invariant system. The corresponding projective unitary representation of the Galilei group comes from a unitary representation U of the extended Galilei group $G = \{g = (\theta,b,a,v)\in\mathbb{R}^4\}$. The group law of G is

$$g_1g_2 = (\theta_1+\theta_2+\tfrac{1}{2}v_1^2b_2 + v_1a_2, b_1+b_2, a_1+a_2+v_1b_2, v_1+v_2), \tag{1.3}$$

and the representation U is given by

$$U(g) = \exp(i\theta M)\exp(ibH)\exp(-iaP)\exp(-ivK), \tag{1.4}$$

where

$$M = \frac{1}{2}\int_{-\infty}^{\infty} dx\,\psi^\dagger\psi, \quad H = \int_{-\infty}^{\infty} dx(\psi_x^\dagger\psi_x + c\psi^\dagger\psi^\dagger\psi\psi),$$
$$P = \int_{-\infty}^{\infty} dx\,\psi^\dagger(-i\psi_x), \quad K = -\frac{1}{2}\int_{-\infty}^{\infty} dx\quad \psi^\dagger x\psi. \tag{1.5}$$

The generators M, H, and P represent, respectively, the mass, the energy, and the momentum observables. The generator K is equal to $-MQ$, where Q represents the position of the center of mass of the system. These operators satisfy the commutation relations

$$[M,H] = [M,P] = [M,K] = [H,P] = 0,$$
$$[H,K] = iP, \quad [P,K] = iM. \tag{1.6}$$

The solutions of (1.1) transform under the action of G according to

$$\psi'(t,x) \equiv U(g)\psi(t,x)U(g)^{-1} = \exp[-(i/2)(\theta+\tfrac{1}{2}v^2t+vx)]\,\psi(t+b,x+vt+a). \tag{1.7}$$

The classical version of Eq. (1.1) is one of the representative models of the ISM.[13] Its dynamical structure presents two basic components which become independent asymptotically in time, namely, solitons and radiation. A group-theoretical analysis[1] reveals that the solitons are Galilean classical particles and that the radiation may be identified with a classical free Schrödinger field. On the other hand, the quantum nonlinear Schrödinger field (1.1) may be described in terms of an infinite spectrum of excitation quanta of momentum p and energy[6,7]

$$p^2/n - (c^2/12)(n^3-n), \quad n = 1,2,\ldots. \tag{1.8}$$

For $n = 1$ these quanta coincide with the fundamental bosons of mass $\frac{1}{2}$. For $n > 1$ they correspond to bound states of n fundamental bosons.

 0022-2488/84/061735-07$02.50

In this paper we use the QISM to analyze the structure of the representation U associated with the Galilean invariance of (1.1). We have two main objectives. The first is to provide a characterization of the excitation quanta of (1.1) in terms of quantum Galilean particles. The second aim is to clarify the correspondence between the dynamical objects arising in the classical and the quantum versions of the nonlinear Schrödinger field. Our analysis is based on the set of scattering data operators introduced by Göckeler.[11] In this way, we find that the quantum nonlinear Schrödinger field may be described in terms of an infinite family of independent boson fields ϕ_n $(n = 1,2,...)$. Then we use these fields to decompose the representation U into a direct product of simple representations of the extended Galilei group. As a consequence it follows that the quanta associated with the fields ϕ_n are characterized as quantum elementary Galilean particles of mass $\frac{1}{2}n$ and internal energy $-(c^2/12)(n^3-n)$. These results are seen to be in complete agreement with the predictions of quasiclassical quantization methods. In particular, we analyze in detail the identification of the radiation quanta with the lowest soliton quanta.[3] Thus, we find that the states obtained by repeated application of the operators $\phi_n^\dagger$ to the vacuum vector are the quantum analogs of the classical pure soliton solutions. This property leads to the characterization of the quantum solitons as the quanta of the fields ϕ_n. In order to get a deeper understanding of the quantum solitons, we provide a method for obtaining the wave functions in configuration space for multisoliton states. It allows us to describe quantum solitons in terms of the asymptotic fragments arising in the scattering processes of the fundamental bosons of (1.1). Finally, we investigate the quantum model

$$i\psi_t = -\psi_{xx} + 2c\psi^\dagger\psi\psi - eEx\psi, \tag{1.9}$$

which describes the interaction of the nonlinear Schrödinger field with a uniform constant electric field. By using a simple transformation[14] it is proved that (1.9) is exactly integrable. We prove also that quantum solitons of mass $\frac{1}{2}n$ react to the electric field as charged particles with charge ne.

II. QISM AND SCATTERING DATA OPERATORS

A. The auxiliary spectral problem

The starting point of the QISM for solving (1.1) is an operator version of the Zakharov–Shabat spectral problem[11]

$$\frac{\partial}{\partial x}v_1 + ikv_1 - \sqrt{-c}\,v_2\psi(x) = 0,$$
$$\frac{\partial}{\partial x}v_2 - ikv_2 + \sqrt{-c}\,\psi^\dagger(x)v_1 = 0. \tag{2.1}$$

Here k is the spectral parameter and $v_i(k,x)$ $(i = 1,2)$ are normal ordered operator functionals of the fields ψ and $\psi^\dagger$. Let us denote by $\phi(k,x)$ the Jost solution of (2.1) verifying

$$\lim_{x\to-\infty} \exp(ikx\sigma_3)\varphi(k,x) = \begin{pmatrix}1\\0\end{pmatrix}. \tag{2.2}$$

Equations (2.1) and condition (2.2) are equivalent to the integral equations

$$\varphi_1(k,x) = e^{-ikx} + \sqrt{-c}\int_{-\infty}^{x} dy\, e^{-ik(x-y)}\varphi_2(k,y)\psi(y), \tag{2.3a}$$

$$\varphi_2(k,x) = -\sqrt{-c}\int_{-\infty}^{x} dy\, e^{ik(x-y)}\psi^\dagger(y)\varphi_1(k,y). \tag{2.3b}$$

By iterating Eqs. (2.3) we may generate series expansions for $\varphi_i(k,x)$ $(i = 1,2)$ whose terms are Wick monomials of the fields ψ and $\psi^\dagger$. For Im $k\geqslant 0$ it follows that the kernels of these Wick monomials are bounded and that the corresponding expressions for the components of φ are well defined as quadratic forms on a suitable domain of the Fock space. Following the same procedure as in the classical case, we define two scattering data operators

$$\begin{pmatrix}A(k)\\B(k)\end{pmatrix} = \lim_{x\to\infty} \exp(ikx\sigma_3)\varphi(k,x). \tag{2.4}$$

Through the operator expression for $\varphi(k,x)$ and (3.4) one finds that $A(k)$ and $B(k)$ are well-defined quadratic forms for Im $k\geqslant 0$ and Im $k = 0$, respectively. Moreover, one finds the following commutation relations[11]:

$$[A(k_1),A(k_2)] = [A(k_1),A^\dagger(k_2)] = [B(k_1),B(k_2)] = 0, \tag{2.5a}$$

$$A^\dagger(k_1)B(k_2) = \frac{2k_1-2k_2-ic}{2k_1-2k_2-i0}B(k_2)A^\dagger(k_1), \tag{2.5b}$$

$$B^\dagger(k_2)A^\dagger(k_1) = \frac{2k_1-2k_2-ic}{2k_1-2k_2-i0}A^\dagger(k_1)B^\dagger(k_2), \tag{2.5c}$$

$$\begin{aligned}B^\dagger(k_1)B(k_2) = {} & -\pi c\delta(k_1-k_2)A^\dagger(k_1)A(k_1)\\ & + \left(\frac{2k_1-2k_2-ic}{2k_1-2k_2-i0}\right)\\ & \times\left(\frac{2k_1-2k_2+ic}{2k_1-2k_2+i0}\right)B(k_2)B^\dagger(k_1).\end{aligned} \tag{2.5d}$$

The operators $\{A(k),B(k):k\in\mathbb{R}\}$ provide a complete description of the quantum nonlinear Schrödinger field in the repulsive case $(c>0)$ (Ref.9). However, in the attractive case additional degrees of freedom are present which require the introduction of a wider set of scattering data operators.

The relevant set of scattering data operators for (1.1) was found by Göckeler.[11] These operators are constructed through a suitable analytic continuation of products of the operators $A(k)$ and $B(k)$. An essential role in Göckeler's method is played by the generator K of pure Galilean transformations defined in (1.5). This operator satisfies the following important property:

$$\begin{aligned}&\exp(-\alpha K)\varphi(k,x)\exp(\alpha K)\\ &\quad = \varphi(k+i\alpha/4,x), \quad \mathrm{Im}k\geqslant 0, \quad \mathrm{Im}(k+i\alpha/4)\geqslant 0,\end{aligned} \tag{2.6}$$

where φ is the Jost solution of (2.1). An analogous property is also satisfied at the classical level[1] and its proof extends without any difficulty for proving (2.6). For each $n = 1,2,...,$ Göckeler defines two scattering data operators

$$A_n(k) = \lim_{x\to\infty}\Big[\exp[c(1-n)K] \times\prod_{j=1}^{n}\hat{\varphi}_1\Big(k-\frac{i}{2}c(n-j),x\Big)\exp[c(n-1)K]\Big], \tag{2.7a}$$

$$B_n^\dagger(-2k)=\lim_{x\to\infty}\Big[\exp[c(1-n)K]\times\prod_{j=1}^{n}\hat\varphi_2\Big(k-\frac{i}{2}c(n-j),x\Big)\times\exp[c(n-1)K]\Big], \tag{2.7b}$$

where $\hat\varphi(k,x)=\exp(ikx\sigma_3)\varphi(k,x)$ and $k\in\mathbb{R}$. The order in the products of $\hat\varphi_i$'s involved in Eqs. (2.6) is not important since it is verified that $[\hat\varphi_i(k,x),\hat\varphi_i(k',x)]=0$ $(i=1,2)$. Observe that by means of a formal application of (2.6), Eqs. (2.7) reduce to the formal expressions

$$A_n(k)=\prod_{j=1}^{n}A\left[k-i\frac{c}{2}\left(\frac{n+1}{2}-j\right)\right],$$

$$B_n^\dagger(-2k)=\prod_{j=1}^{n}B\left[k-i\frac{c}{2}\left(\frac{n+1}{2}-j\right)\right]. \tag{2.8}$$

B. Reflection coefficient operators

The algebraic properties of the scattering data operators adopt their simplest form when expressed in terms of the "reflection coefficient operators" defined by

$$R_n^\dagger(p)=[2\pi|c|^{1/2}n]^{-1}B_n^\dagger(p/n)\times[A_n(-p/2n)]^{-1},\quad n=1,2,\ldots,\quad p\in\mathbb{R}. \tag{2.9}$$

This definition differs slightly from the one used by Göckeler[11] who sets $R_n^\dagger(p)=B_n^\dagger(p)[A_n(-p/2)]^{-1}$. The operators (2.9) satisfy the following simple algebraic relations:

$$R_n^\dagger(p)R_{n'}^\dagger(p')=S_{nn'}\Big(\frac{p}{n}-\frac{p'}{n'}\Big)R_{n'}^\dagger(p')R_n^\dagger(p), \tag{2.10a}$$

$$R_n(p)R_{n'}^\dagger(p')=\delta_{nn'}\delta(p-p')\times S_{nn'}^*\Big(\frac{p}{n}-\frac{p'}{n'}\Big)R_{n'}^\dagger(p')R_n(p), \tag{2.10b}$$

where the coefficients $S_{nn'}$ are given by

$$S_{nn'}(q)=\prod_{m=-s}^{s}\prod_{m'=-s'}^{s'}\frac{q+ic(m+m'+1)}{q+ic(m+m'-1)},\quad s=(n-1)/2,\quad s'=(n'-1)/2. \tag{2.11}$$

Note that

$$S_{nn'}(q)=S_{n'n}(q)=S_{nn'}^*(-q),\quad |S_{nn'}(q)|=1. \tag{2.12}$$

Equations (2.10) exhibit in a convenient form the algebra of the reflection coefficient operators. They may be derived from Göckeler's results[11] by observing that according to the formula for the composition of two quantum angular momenta, the coefficient $S_{nn'}$ defined in (2.11), may be written as

$$S_{nn'}(q)=\prod_{j=|s-s'|}^{s+s'}\prod_{m=-j}^{j}\frac{q+ic(m+1)}{q+ic(m-1)}, \tag{2.13}$$

which leads to the following two other expressions:

$$S_{nn'}(q)=\prod_{\alpha=1}^{\bar n}\frac{q+ic(2\alpha+|n-n'|)/2}{q-ic(2\alpha+|n-n'|)/2}\times\frac{q-ic(2\alpha-n-n')/2}{q+ic(2\alpha-n-n')/2}, \tag{2.14a}$$

$$S_{nn'}(q)=\frac{q^2+c^2(n+n')^2/4}{q^2+c^2(n-n')^2/4}\times\prod_{\alpha=1}^{\bar n}\left(\frac{q-ic(2\alpha-n-n')/2}{q-ic(2\alpha+|n-n'|)/2}\right)^2, \tag{2.14b}$$

where $\bar n=\min(n,n')$. From (2.14a) and (2.14b) the equivalence between (2.9) and the relations deduced by Göckeler follows at once.

In addition to (2.10) the reflection coefficient operators satisfy the following commutation relations[11] with the operators M, P, and H defined in (1.5):

$$[M,R_n^\dagger(p)]=\tfrac12 nR_n^\dagger(p),\quad [P,R_n^\dagger(p)]=pR_n^\dagger(p), \tag{2.15a}$$

$$[H,R_n^\dagger(p)]=[p^2/n-(c^2/12)(n^3-n)]R_n^\dagger(p). \tag{2.15b}$$

III. ANALYSIS OF THE GALILEAN ACTION

A. Galilean generators and scattering data operators

In order to analyze the unitary representation U of the extended Galilei group, we look for the expressions of its generators in terms of scattering data variables. They are the quantum versions of the trace relations arising in the classical context of the Zakharov–Shabat spectral problem. First, let us prove that

$$M=\sum_{n>1}\frac12 n\int_{-\infty}^{\infty}dp\,R_n^\dagger(p)R_n(p), \tag{3.1a}$$

$$P=\sum_{n>1}\int_{-\infty}^{\infty}dp\,pR_n^\dagger(p)R_n(p), \tag{3.1b}$$

$$H=\sum_{n>1}\int_{-\infty}^{\infty}dp\Big(\frac{p^2}{n}-\frac{c^2}{12}(n^3-n)\Big)R_n^\dagger(p)R_n(p). \tag{3.1c}$$

From the algebraic relations (2.10) it is clear that the expressions (3.1) for M, P, and H satisfy the commutation relations (2.15). On the other hand, the quantum Gel'fand–Levitan method[11,12] for (1.1) shows that the quantum field $\psi(x)$ may be reconstructed from the operators $\{R_n(p):n=1,2,\ldots;p\in\mathbb{R}\}$. This property implies that Eqs. (2.15) determine M, P, and H up to an additive constant. Then Eqs. (3.1) follow from (2.15) and the fact that all the operators M, P, H, and $R_n(p)$ annihilate the vacuum state.

The interpretation of Eqs. (3.1) becomes simple when we use a realization of the operators R_n in terms of boson fields. This kind of realization was first proposed by Grosse[15] in the repulsive case. It is based on the following property of boson fields $\phi(p)$:

$$\exp\Big(\int_{-\infty}^{\infty}dp'\,G(q,p')\phi^\dagger(p')\phi(p')\Big)\phi^\dagger(p)=\phi^\dagger(p)\exp[G(q,p)]\times\exp\Big(\int_{-\infty}^{\infty}dp'\,G(q,p')\phi^\dagger(p')\phi(p')\Big). \tag{3.2}$$

From this property it follows at once that given a family of independent boson fields $\phi_n(p)$ $(n\geqslant1)$

$$[\phi_n(p),\phi_{n'}(p')]=0,$$
$$[\phi_n(p),\phi_{n'}^\dagger(p')]=\delta_{nn'}\delta(p-p'), \tag{3.3}$$

then the operators

$$R_n^\dagger(p) = \phi_n^\dagger(p) \prod_{n'>1} \exp\left[\int_{-\infty}^{\infty} dp'\, \theta\left(\frac{p}{n} - \frac{p'}{n'}\right)\right.$$
$$\left. \times \ln S_{nn'}\left(\frac{p}{n} - \frac{p'}{n'}\right) \phi_{n'}^\dagger(p')\phi_{n'}(p')\right] \qquad (3.4)$$

satisfy the algebraic relations (2.10). In addition, (3.4) implies that

$$R_n^\dagger(p) R_n(p) = \phi_n^\dagger(p)\phi_n(p), \qquad (3.5)$$

for all $n \geqslant 1$. Therefore, Eq. (3.4) may be inverted and we get that

$$\phi_n^\dagger(p) = R_n^\dagger(p) \prod_{n'>1} \exp\left[\int_{-\infty}^{\infty} dp'\, \theta\left(\frac{p}{n} - \frac{p'}{n'}\right)\right.$$
$$\left. \times \ln S_{nn'}\left(\frac{p'}{n'} - \frac{p}{n}\right) R_{n'}^\dagger(p') R_{n'}(p')\right]. \qquad (3.6)$$

Equations (3.4) and (3.6) show how to pass from the reflection coefficient operators to a set of independent boson fields satisfying the canonical commutation relations (3.3). In terms of these boson fields the Galilean generators adopt the following form:

$$M = \sum_{n>1} \frac{1}{2} n \int_{-\infty}^{\infty} dp\, \phi_n^\dagger(p)\phi_n(p), \qquad (3.7a)$$

$$P = \sum_{n>1} \int_{-\infty}^{\infty} dp\, p \phi_n^\dagger(p)\phi_n(p), \qquad (3.7b)$$

$$H = \sum_{n>1} \int_{-\infty}^{\infty} dp \left(\frac{p^2}{n} - \frac{c^2}{12}(n^3 - n)\right) \phi_n^\dagger(p)\phi_n(p), \qquad (3.7c)$$

$$K = -\sum_{n>1} \frac{1}{2} in \int_{-\infty}^{\infty} dp\, \phi_n^\dagger(p) \frac{\partial}{\partial p} \phi_n(p). \qquad (3.7d)$$

The first three formulas are a trivial consequence of Eqs. (3.1) and (3.5). To prove (3.7d) we note that from (2.6) and (2.7) we have that

$$\exp(-ivK) R_n(p) \exp(ivK) = R_n(p + \tfrac{1}{2}nv), \qquad (3.8)$$

and then from (3.6) it follows that

$$\exp(-ivK)\phi_n(p)\exp(ivK) = \phi_n(p + \tfrac{1}{2}nv), \qquad (3.9)$$

which leads at once to (3.7d).

The expressions (3.7) show that the unitary representation U associated with the Galilean invariance of the nonlinear Schrödinger field is equivalent to a direct product of the form

$$U_1 \otimes U_2 \otimes \cdots \otimes U_n \otimes \cdots, \qquad (3.10)$$

where $U_n (n \geqslant 1)$ is the unitary representation of the extended Galilei group associated with a field of free bosons of mass $\frac{1}{2}n$ and internal energy $-(c^2/12)(n^3 - n)$. In this way the fundamental interacting field $\psi(t,x)$ gives rise to an infinite spectrum of quantum elementary Galilean particles.

It is instructive to see how the structure (3.10) of U may be derived from the quasiclassical quantization[16] of the Galilean realization R associated with the classical version of (1.1) (Ref. 1). The realization R is a composition of two Galilean invariant systems. One represents the radiation component and is described by a free classical Schrödinger field of mass $\frac{1}{2}$, which upon quantization gives rise to the factor U_1 in (3.10). The other system corresponds to the solitons and is described by a variable number of classical Galilean particles. These classical particles have a phase space with two pairs of action-angle variables (p,q) and (m,τ). The variable m is the mass of the particle and its associated angle variable τ is defined mod 4π. Moreover these particles have an internal energy equal to $-\frac{4}{3}c^2m^3$. By applying the standard quantization rule to the pair (m,τ) we get a spectrum of quantum particles with masses $\frac{1}{2}n$ and internal energies $-(c^2/12)n^3$ $(n \geqslant 1)$. The presence of the additional term $(c^2/12)n$ in the quantum internal energy may also be justified by the introduction of a further quantum correction.[10,17] Hence, we see that the quantization of the solitons produces the whole spectrum $U_n (n \geqslant 1)$ of the unitary representation U. In particular, it must be observed that the factor U_1 in (3.10) associated with the quantization of the radiation component appears also as the lowest state of the quantized solitons.[3] This suggests that the quantum nonlinear Schrödinger field involves soliton degrees of freedom only. In order to understand the meaning of this fact it is necessary to recall the properties of classical pure soliton solutions. As it is well known[13] these solutions correspond to a scattering data coefficient $a(k)$ of the form

$$a(k) = \prod_{l=1}^{r} \frac{k - k_l}{k - k_l^*}, \quad \operatorname{Im} k_l > 0. \qquad (3.11)$$

The classical field associated with (3.11) contains r solitons whose masses and momenta are related with the zeros of $a(k)$ according to[1]

$$k_l = -p_l/4m_l - i(c/2)m_l. \qquad (3.12)$$

At the quantum level the role of the coefficient $a(k)$ is played by the operator $A(k)$ defined in (2.4). Moreover, the Hilbert space of states is generated by the vectors

$$|p_1n_1;p_2n_2;\cdots;p_rn_r\rangle = \phi_{n_1}^\dagger(p_1)\phi_{n_2}^\dagger(p_2)\cdots\phi_{n_r}^\dagger(p_r)|0\rangle, \quad n_l \geqslant 1, \qquad (3.13)$$

with $|0\rangle$ being the vacuum state. By using (3.4) it follows that

$$|p_1n_1;\cdots;p_rn_r\rangle = \prod_{i<j}\left[\theta\left(\frac{p_j}{n_j} - \frac{p_i}{n_i}\right) + \theta\left(\frac{p_i}{n_i} - \frac{p_j}{n_j}\right) S_{n_in_j}\left(\frac{p_j}{n_j} - \frac{p_i}{n_i}\right)\right] \times R_{n_1}^\dagger(p_1)\cdots R_{n_r}^\dagger(p_r)|0\rangle. \qquad (3.14)$$

On the other hand, $A(k)$ satisfies the following commutation relations with the reflection coefficient operators[11]:

$$A(k)R_n^\dagger(p) = \frac{k + p/2n + ic[(n+1)/4]}{k + p/2n - ic[(n-1)/4]} R_n^\dagger(p) A(k). \qquad (3.15)$$

To get a closer analogy between the classical and the quantum equations it is convenient to use the translated operator $\hat{A}(k) = A(k - ic/4)$. Then from (3.14) and (3.15) and taking into account that $A(k)|0\rangle = |0\rangle$, we have that

$$\hat{A}(k)|p_1n_1;\cdots;p_rn_r\rangle = \hat{a}(k)|p_1n_1;\cdots;p_rn_r\rangle, \qquad (3.16)$$

where

$$\hat{a}(k) = \prod_{l=1}^{r} \frac{k - k_l}{k - k_l^*}, \quad k_l = -\frac{p_l}{2n_l} - i\frac{c}{4}n_l. \qquad (3.17)$$

Observe the complete analogy between (3.17) and the classical expressions (3.11) and (3.12). Therefore, in terms of the ISM language, a state $|p_1n_1;\cdots;p_rn_r\rangle$ represents a pure soliton state with r quantum solitons with masses $\frac{1}{2}n_l$, momenta p_l, and internal energies $-(c^2/12)(n_l^3-n_1)$ $(l=1,\dots,r)$.

B. Quantum solitons and scattering states

Our next objective is to study the wave functions associated with the basic states (3.13). That is to say, we want to determine the scalar products

$$(1/\sqrt{n!})\langle x_1\cdots x_n|p_1n_1;\cdots;p_rn_r\rangle,\quad n=n_1+\cdots+n_r, \tag{3.18}$$

where $|x_1\cdots x_n\rangle=\psi^\dagger(x_1)\cdots\psi^\dagger(x_n)|0\rangle$. Firstly, we observe that according to (3.14)

$$|p_1n_1;\dots;p_rn_r\rangle=R^\dagger_{n_r}(p_r)\cdots R^\dagger_{n_1}(p_1)|0\rangle, \tag{3.19}$$

provided that

$$p_1/n_1>p_2/n_2>\cdots>p_r/n_r. \tag{3.20}$$

It may be shown[18] that the wave function associated with

$$R^\dagger_1(k_n)R^\dagger_1(k_{n-1})\cdots R^\dagger_1(k_1)|0\rangle \tag{3.21}$$

takes on the regions $x_1\leqslant x_2\leqslant\cdots\leqslant x_n$ the Bethe ansatz form

$$\frac{(-1)^n}{(2\pi)^{n/2}\sqrt{n!}}\sum_\sigma\xi(\sigma,k)\exp\Big(i\sum_j k_{\sigma(j)}x_j\Big), \tag{3.22}$$

where the sum extends to all the elements σ of the permutation group of n objects and

$$\xi(\sigma,k)=(-1)^{\epsilon(\sigma)}\prod_{i<j}\frac{k_{\sigma(i)}-k_{\sigma(j)}+ic}{k_i-k_j+ic},$$
$$k=(k_1,\dots,k_n), \tag{3.23}$$

with $\epsilon(\sigma)$ being the parity of the permutation σ. For other values of x the function (3.22) is extended by symmetry.

The wave function for a general vector (3.19) can be found by means of a suitable analytical continuation of Eq. (3.22) in the following way. By repeated use of the commutation relation (2.5c) one obtains that

$$\begin{aligned}R^\dagger_1(k_n)\cdots R^\dagger_1(k_1)&=(2\pi|c|)^{-n/2}\Big[B\Big(-\frac{k_n}{2}\Big)A\Big(-\frac{k_n}{2}\Big)^{-1}\Big]\cdots\Big[B\Big(-\frac{k_1}{2}\Big)A\Big(-\frac{k_1}{2}\Big)^{-1}\Big]\\&=(2\pi|c|)^{-n/2}\Big(\prod_{i<j}\frac{k_i-k_j}{k_i-k_j+ic}\Big)\\&\quad\times\Big[\prod_{i=1}^n B\Big(-\frac{k_i}{2}\Big)\Big]\Big[\prod_{i=1}^n A\Big(-\frac{k_i}{2}\Big)\Big]^{-1}.\end{aligned} \tag{3.24}$$

If we now use the formal expressions (2.8) in the definition (2.9) of $R^\dagger_n(p)$, then from (3.24) we get

$$R^\dagger_n(p)=(n-1)!(2\pi|c|)^{(n-1)/2}\times R^\dagger_1(k_n)R^\dagger_1(k_{n-1})\cdots R^\dagger_1(k_1), \tag{3.25}$$

where

$$k_j=p/n-ic[j-(n+1)/2],\quad j=1,\dots,n. \tag{3.26}$$

Equation (3.25) has only a formal meaning. However, it suggests that the wave function of a state of the form (3.19)–(3.20) obtains through the analytic continuation of Eq. (3.22) to a complex vector k given by

$$k=p^{(1)}\oplus\cdots\oplus p^{(r)}, \tag{3.27}$$

where $p^{(l)}$ $(l=1,\dots,r)$ denotes the n_l-dimensional vector of the form (3.26) with $p=p_l$ and $n=n_l$. The resulting function turns out to be bounded. To prove this it is convenient to introduce some notation conventions. Let C be a subset of $\{1,2,\dots,n\}$ with n' elements, then we define

$$Q(C)=\frac{1}{n'}\sum_{i\in C}x_i,$$
$$\varphi(C,x)=\exp\Big(-\frac{|c|}{4}\sum_{i,j\in C}|x_i-x_j|\Big). \tag{3.28}$$

Observe that given a vector k of the form (3.26), then on the region $x_1\leqslant x_2\leqslant\dots\leqslant x_n$ we have that

$$\exp\Big(i\sum_j k_jx_j\Big)=\exp[ipQ(C)]\varphi(C,x),$$
$$C=\{1,2,\dots,n\}. \tag{3.29}$$

Let us consider now a vector k of the form (3.27). We denote by $C_1,\dots,C_r$ the cluster decomposition of the set $\{1,2,\dots,n=n_1+\cdots+n_r\}$ given by

$$C_1=\{1,2,\dots,n_1\},$$
$$C_2=\{n_1+1,n_1+2,\dots,n_1+n_2\},\dots,$$
$$C_r=\{n-n_r+1,\dots,n\}. \tag{3.30}$$

It is not difficult to prove the following two properties.

(i) The coefficient $\xi(\sigma,k)$ is different from zero only for those σ such that

$$\forall i,j\in C_l(l=1,\dots,r)/i<j\Rightarrow\sigma^{-1}(i)<\sigma^{-1}(j). \tag{3.31}$$

(ii) Given σ verifying (3.31), then on the region $x_1\leqslant x_2\leqslant\cdots\leqslant x_n$,

$$\exp\Big(i\sum_j k_{\sigma(j)}x_j\Big)=\prod_{l=1}^r\exp\{ip_lQ[\sigma^{-1}(C_l)]\}\varphi[\sigma^{-1}(C_l),x], \tag{3.32}$$

where $\sigma^{-1}(C_l)=\{\sigma^{-1}(i):i\in C_l\}$.

Therefore the wave function of a state of the form (3.19)–(3.20) on the region $x_1\leqslant x_2\leqslant\cdots\leqslant x_n$ is

$$N\sum_\sigma{}'\xi(\sigma,k)\prod_{l=1}^r\exp\{ip_lQ[\sigma^{-1}(C_l)]\}\varphi[\sigma^{-1}(C_l),x], \tag{3.33}$$

where $k=p^{(1)}\oplus\cdots\oplus p^{(r)}$ and the sum extends to all the permutations σ satisfying (3.31). The normalization constant N is given by

$$N=(-1)^n(2\pi|c|)^{-r/2}|c|^{n/2}\times\prod_{l=1}^r\Big(\frac{(n_l-1)!}{n_l}\Big)^{1/2},\quad n=n_1+\cdots+n_r. \tag{3.34}$$

Each term in the sum (3.33) describes a motion of n particles of mass $\frac{1}{2}$ distributed into r fragments of n_l ($l = 1,...,r$) particles. The wave function for each fragment with $n_l > 1$ represents a bound state of its constituent particles whose center of mass moves freely. By considering a wave packet formed from (3.33), one sees that due to (3.20) the only term which remains in the region $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$ as $t \to -\infty$ is that corresponding to the identical permutation $\sigma = e$. Moreover, from (3.23) we have that $\xi(e,k) = 1$. Thus, the wave function (3.33) is a state of n bosons which as $t \to -\infty$ describes a motion of r freely moving fragments with masses $\frac{1}{2}n_l$, momenta p_l, and internal energy $-(c^2/12)(n_l^3 - n_l)$. This result provides a clear picture of the concept of quantum soliton. We may identify the quantum solitons with the asymptotic fragments arising in the evolution of the scattering state (3.19)–(3.20) as $t \to -\infty$. In this way, the fields $\phi_n^\dagger(k)$ introduced in (3.6) are creation fields for incoming solitons.

A similar analysis shows that the fields defined by

$$\hat{\phi}_n^\dagger(p) = R_n^\dagger(p) \prod_{n'>1} \exp\left[\int_{-\infty}^{\infty} dp'\, \theta\left(\frac{p'}{n'} - \frac{p}{n}\right) \times \ln S_{nn'}\left(\frac{p'}{n'} - \frac{p}{n}\right) R_{n'}^\dagger(p') R_{n'}(p')\right]. \quad (3.35)$$

are creation fields for outgoing solitons. In terms of these fields the Galilean generators take the same form as in (3.7). Moreover, we may construct outgoing states of r solitons

$$\begin{aligned} |p_1 n_1;...;p_r n_r\rangle' &= \hat{\phi}_{n_1}^\dagger(p_1) \cdots \hat{\phi}_{n_r}^\dagger(p_r)|0\rangle = \prod_{i<j}\left[\theta\left(\frac{p_i}{n_i} - \frac{p_j}{n_j}\right) \right. \\ &\quad \left. + \theta\left(\frac{p_j}{n_j} - \frac{p_i}{n_i}\right) S_{n_i n_j}\left(\frac{p_j}{n_j} - \frac{p_i}{n_i}\right)\right] \\ &\quad \times R_{n_1}^\dagger(p_1) \cdots R_{n_r}^\dagger(p_r)|0\rangle, \end{aligned} \quad (3.36)$$

and from (3.14) and (3.36) we have that the S matrix for the scattering of r solitons, with masses n_i and momenta p_i such that $p_1/n_1 > p_2/n_2 > \cdots > p_r/n_r$, adopts the factorized form[19]

$$\prod_{i<j} S_{n_i n_j}\left(\frac{p_j}{n_j} - \frac{p_i}{n_i}\right). \quad (3.37)$$

IV. INTERACTION WITH A UNIFORM CONSTANT ELECTRIC FIELD

Let us consider the action of a uniform constant electric field E on the quantum nonlinear Schrödinger field. The Hamiltonian of the system is now given by

$$H' = \int_{-\infty}^{\infty} dx(\psi_x^\dagger \psi_x + c\psi^\dagger\psi^\dagger\psi\psi - eEx\psi^\dagger\psi), \quad (4.1)$$

where it is assumed that the charge of the fundamental bosons of the model is equal to e. The Heisenberg equation of motion for the field ψ has the form

$$i\psi_t = -\psi_{xx} + 2c\psi^\dagger\psi\psi - eEx\psi. \quad (4.2)$$

Like its classical version,[20] the quantum equation (4.2) is exactly integrable. Indeed, by defining the field[14]

$$\hat{\psi}(t,x) = \psi(t,x + eEt^2)\exp\{-i[eEtx + \tfrac{2}{3}(eE)^2t^3]\}, \quad (4.3)$$

Eq. (4.2) reduces to the quantum nonlinear Schrödinger equation (1.1) for $\hat{\psi}$. This property may be formulated in terms of wave functions as follows. Given a function $f(t,x_1,...,x_n)$ verifying

$$i\frac{\partial f}{\partial t} = -\sum_i \frac{\partial^2 f}{\partial x_i^2} + 2c\sum_{i<j}\delta(x_i - x_j)f, \quad (4.4)$$

then the function

$$F(t,x_1,...,x_n) = f(t,x_1 - eEt^2,...,x_n - eEt^2) \times \exp\left[i\left(eEt\sum_i x_i - \frac{n}{3}(eE)^2t^3\right)\right], \quad (4.5)$$

satisfies the equation

$$i\frac{\partial F}{\partial t} = -\sum_i \frac{\partial^2 F}{\partial x_i^2} + 2c\sum_{i<j}\delta(x_i - x_j)F - eE\sum_i x_i F. \quad (4.6)$$

By using (4.5) we may determine the effect of the electric field on the quantum solitons of the nonlinear Schrödinger field. If we take as f the wave function (3.33) associated with an incoming state of solitons, then we have that under the action of the electric field solitons move asymptotically with an acceleration equal to $2eE$. That is to say, solitons of mass $\frac{1}{2}n$ behave as charged particles with charge ne.

The exact integrability of (4.2) is a consequence of the properties of the nonlinear Schrödinger field under Galilean transformations. Indeed, the Hamiltonian (4.1) is equal to $H + 2eEK$, where H and K are the Galilean generators defined in (1.4). Then, due to the expressions (3.7c) and (3.7d) the creation fields for incoming solitons verify that

$$[H', \phi_n^\dagger(p)] = \left(\frac{p^2}{n} - \frac{c^2}{12}(n^3 - n)\right)\phi_n^\dagger(p) + ineE\frac{\partial\phi_n^\dagger(p)}{\partial p}, \quad (4.7)$$

and therefore the incoming states of solitons evolve according to

$$\begin{aligned} &\exp(-itH')|p_1n_1;...;p_rn_r\rangle \\ &= \exp\left[-it\sum_l\left(\frac{p_l^2}{n_l} - \frac{c^2}{12}(n_l^3 - n_l) + eEp_l t + \frac{1}{3}n_l(eEt)^2\right)\right]|p_1(t)n_1;...;p_r(t)n_r\rangle, \end{aligned} \quad (4.8)$$

where $p_l(t) = p_l + eEtn_l$.

ACKNOWLEDGMENTS

I am indebted to H. de Vega for guiding me through the literature on the quantum inverse scattering method.

This work was financially supported by the Comisión Asesora de Investigación Científica y Técnica.

[1]L. Martínez Alonso, J. Math. Phys. **23**, 1518 (1982); **24**, 982 (1983).
[2]L. Martínez Alonso, J. Math. Phys. **24**, 2652 (1983).
[3]L. D. Faddeev, Sov. Sci. Rev. C **1**, 107 (1980).
[4]H. B. Thacker, Rev. Mod. Phys. **53**, 253 (1981).
[5]P. P. Kulish and E. K. Sklyanin, in *Integrable Quantum Field Theories, Proceedings of the Tvarmine Symposium, Lecture Notes in Physics* 151, edited by J. Hietarinta and C. Montonen (Springer-Verlag, Berlin, 1982).

[6]J. B. McGuire, J. Math. Phys. **5**, 622 (1964).
[7]C. N. Yang, Phys. Rev. **168**, 1920 (1968).
[8]H. A. Bethe, Z. Phys. **71**, 205 (1931).
[9]E. K. Sklyanin, Zap. Nauch. Semin. LOMI **95**, 55 (1980).
[10]The relevance of the ISM in this quantum model was first studied by D. J. Kaup, J. Math. Phys. **16**, 2036 (1975).
[11]M. Göckeler, Z. Phys. C-Part. Fields **7**, 263 (1981); **11**, 125 (1981).
[12]F. A. Smirnov, Sov. Phys. Dokl. **27**, 34 (1982).
[13]V. E. Zakharov and A. B. Shabat, Sov. Phys. JETP **34**, 62 (1972).
[14]D. J. Kaup and A. C. Newell, Proc. R. Soc. London Ser. A **361**, 413 (1978).
[15]H. Grosse, Phys. Lett. B **86**, 267 (1979).
[16]A. Kirillov, *Elements de la Theorie des Representations* (Editions Mir, Moscow, 1974).
[17]L. D. Faddeev, in *Solitons*, edited by R. K. Bullough and P. J. Caudrey (Springer-Verlag, Berlin, 1980).
[18]This result is a straightforward consequence of Proposition 2.4.1 of Ref. 9. We notice that the operators $A_1(k)$ and $B_1(k)$ are related with Sklyanin's operators $A(k)$ and $B(k)$ in the form $A_1(k) = A^\dagger(-2k)$, $B_1(k) = i|c|^{1/2}B(k)$.
[19]A. B. Zamolodchikov and Al. B. Zamolodchikov, Ann. Phys. **120**, 253 (1979).
[20]H. H. Chen and C. S. Liu, Phys. Rev. Lett. **37**, 693 (1976).

Fermi pseudopotential in higher dimensions

Alexander Grossmann
Centre de Physique Théorique II, CNRS, Marseille, France
Tai Tsun Wu
Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany, and Gordon McKay Laboratory, Harvard University, Cambridge, Massachusetts 02138

(Received 16 June 1983; accepted for publication 9 December 1983)

The Fermi pseudopotential is generalized from three to five dimensions, and the case of an infinite, uniform, equidistant, linear chain of such pseudopotentials is studied in detail. Similar to the three-dimensional case, zero-width resonances are also present in five dimensions. While this generalization is natural and can be carried through formally when the strength is negative, there are basic changes in the underlying structure. These results in five dimensions also apply in four dimensions.

PACS numbers: 02.30.Tb, 12.90. + b

I. INTRODUCTION

The Fermi pseudopotential was first used nearly half a century ago.[1] The first applications were to problems of nuclear physics,[2,3] and the later ones to many-body problems.[4–7] In three dimensions, it is

$$4\pi\delta^3(\mathbf{r})\frac{\partial}{\partial r}r, \tag{1.1}$$

and is called a pseudopotential because it is an operator instead of a function. In the simplest case, a is a parameter with dimension length; more generally, a can be a function of the energy.[4] Here we are not going to deal with such energy dependence.

Recently, potentials of the form

$$\alpha\delta^3(\mathbf{r}) \tag{1.2}$$

were studied,[8–10] where α is infinitesimal. It is now known that $-\nabla^2+(1.1)$ and $-\nabla^2+(1.2)$ are the same with α and a suitable functions of each other.

There is a general belief that the Fermi pseudopotential exists only in two and three dimensions, but not in higher dimensions.[10] The reason for this belief, in a nutshell, is as follows. Consider for $\kappa>0$ the free-space Green's function in d dimensions

$$(-\nabla^2+\kappa^2)G_0(r)=\delta^d(\mathbf{r}). \tag{1.3}$$

It is square integrable in two and three dimensions, but not for $d\geqslant 4$.

Intuitively, it seems that this condition of being square integrable is rather far removed from physics. One of the original ways of motivating the Fermi pseudopotential is to take a repulsive potential such as a hard sphere, and to continue analytically the s-wave part of the wave function outside the potential. Thus the wave function for the Fermi pseudopotential (1.1) is physically meaningful only for r away from the origin. With this motivation, square integrability in the vicinity of the origin has no physical interpretation.

It is the purpose of the present paper to initiate the study of Fermi pseudopotentials in higher dimensions by considering the case $d=5$. This value of d has been chosen judiciously. On the one hand, because of the form of the free-space Green's function, the Fermi pseudopotential is simpler in form when d is odd than when d is even, as already evident by comparing the cases $d=2$ and $d=3$. On the other hand, the case $d=7$ is much more complicated than the case $d=5$, for reasons that are not yet understood.

In three dimensions, since the free-space Green's function has the singularity r^{-1}, the operator $(\partial/\partial r)r$ in (1.1) is chosen to annihilate this singularity. In an analogous way, since the free-space Green's function in five dimensions has the singularity r^{-3}, the Fermi pseudopotential is chosen to be

$$4\pi^2a^3\delta^5(\mathbf{r})\frac{\partial^3}{\partial r^3}r^3, \tag{1.4}$$

where the coefficient $4\pi^2$ is only for convenience. Here a also has the dimension of length. An alternative, entirely equivalent choice is

$$4\pi^2a^3\delta^5(\mathbf{r})\frac{\partial}{\partial r}\frac{2}{r}\frac{\partial}{\partial r}r^3. \tag{1.5}$$

In Sec. II, the Green's function in the presence of this Fermi pseudopotential (1.4) or (1.5) is obtained explicitly in a completely straightforward way. However, the result is interesting in that it makes sense only for

$$a\leqslant 0. \tag{1.6}$$

As in the three-dimensional case, $a=\infty$ is allowed. In Sec. III, this Green's function is applied to the case of an infinite equispaced linear chain of Fermi pseudopotentials. It is found that the phenomenon of infinitely narrow resonances, previously known in three dimensions, also occurs in five dimensions. Section IV is devoted to the simplest property of the Green's function of Sec. II as the resolvent operator, and the more mathematical questions of this Fermi pseudopotential in five dimensions are raised in Sec. V, but by no means solved.

The failure of the present procedure in seven dimensions stems from the fact that there the condition (1.6) is replaced by $a\leqslant 0$ and $a\geqslant 0$, or more precisely $a=0$ or ∞. Assuming that the mathematical problems discussed in Sec. V can be solved, then the present situation with Fermi pseudopotentials in higher dimensions is as follows. A Fermi pseudopotential can be defined in all dimensions corresponding to $a=\infty$, meaning that the Green's function is defined and has the necessary analytic properties. In four

 0022-2488/84/061742-04$02.50

and five dimensions, a one-parameter family of Fermi pseudopotentials can be defined, analogous to the known cases of two and three dimensions. The situation is as yet unclear in six dimensions.

II. GREEN'S FUNCTION

A. Three dimensions

Before obtaining the Green's function in five dimensions in the presence of the Fermi pseudopotential (1.4), we review the corresponding problem in three dimensions with (1.1). The Green's function is defined through the partial differential equation

$$\left(\nabla^2 + k^2 - 4\pi a \delta^3(\mathbf{r})\frac{\partial}{\partial r} r\right) G(\mathbf{r}, \mathbf{r}_0; k) = -\delta^3(\mathbf{r} - \mathbf{r}_0) \tag{2.1}$$

together with the usual boundary conditions at infinity. Let

$$A = 4\pi a \frac{\partial}{\partial r} rG(\mathbf{r}, \mathbf{r}_0; k)|_{r=0}, \tag{2.2}$$

then

$$(\nabla^2 + k^2) G(\mathbf{r}, \mathbf{r}_0; k) = -\delta^3(\mathbf{r} - \mathbf{r}_0) + A\delta^3(\mathbf{r}), \tag{2.3}$$

and hence

$$G(\mathbf{r}, \mathbf{r}_0; k) = G_0(\mathbf{r}, \mathbf{r}_0; k) - AG_0(\mathbf{r}, 0; k), \tag{2.4}$$

where $G_0(\mathbf{r}, \mathbf{r}_0; k)$ is the free-space Green's function given explicitly by

$$G_0(\mathbf{r}, \mathbf{r}_0; k) = e^{ik|\mathbf{r} - \mathbf{r}_0|}/4\pi|\mathbf{r} - \mathbf{r}_0|. \tag{2.5}$$

The coefficient A is determined by substituting (2.4) into (2.2) and the result is

$$G(\mathbf{r}, \mathbf{r}_0; k) = G_0(\mathbf{r}, \mathbf{r}_0; k) - [4\pi a/(1 + ika)] G_0(\mathbf{r}, 0; k) G_0(\mathbf{r}_0, 0; k). \tag{2.6}$$

This same Green's function is also obtained from the "point interaction" (1.2). This shows that the Fermi pseudopotential and the point interaction are one and the same in three dimensions.

B. Five dimensions

The derivation in five dimensions is step-by-step the same. We begin with the partial differential equation

$$\left(\nabla^2 + k^2 - 4\pi^2 a^3 \delta^5(\mathbf{r})\frac{\partial^3}{\partial r^3} r^3\right) G(\mathbf{r}, \mathbf{r}_0; k) = -\delta^5(\mathbf{r} - \mathbf{r}_0) \tag{2.7}$$

in five dimensions. Let

$$A = 4\pi^2 a^3 \frac{\partial^3}{\partial r^3} r^3 G(\mathbf{r}, \mathbf{r}_0; k)\Big|_{r=0}, \tag{2.8}$$

then

$$(\nabla^2 + k^2) G(\mathbf{r}, \mathbf{r}_0; k) = -\delta^5(\mathbf{r} - \mathbf{r}_0) + A\delta^5(\mathbf{r}), \tag{2.9}$$

and hence

$$G(\mathbf{r}, \mathbf{r}_0; k) = G_0(\mathbf{r}, \mathbf{r}_0; k) - AG_0(\mathbf{r}, 0; k), \tag{2.10}$$

where $G_0(\mathbf{r}, \mathbf{r}_0; k)$ is the free-space Green's function in five dimensions

$$G_0(\mathbf{r}, \mathbf{r}_0; k) = (8\pi^2)^{-1}|\mathbf{r} - \mathbf{r}_0|^{-3} e^{ik|\mathbf{r} - \mathbf{r}_0|}(1 - ik|\mathbf{r} - \mathbf{r}_0|). \tag{2.11}$$

Again, the coefficient A is determined by substituting (2.10) into (2.8), and the result is

$$G(\mathbf{r}, \mathbf{r}_0; k) = G(\mathbf{r}, \mathbf{r}_0; k) - B(k) G_0(\mathbf{r}, 0; k) G_0(\mathbf{r}_0, 0; k), \tag{2.12}$$

where

$$B(k) = 24\pi^2 a^3/(1 + ik^3 a^3). \tag{2.13}$$

In spite of the close similarity of the results, $B(k)$ has new and interesting features. If $a < 0$, there is a bound state with binding energy a^{-2} and bound-state wave function

$$r^{-3} e^{-r/|a|}(1 + r/|a|). \tag{2.14}$$

When $a > 0$, there are *complex* poles located at

$$k = a^{-1} e^{i\pi/6}$$

and

$$k = a^{-1} e^{5i\pi/6}. \tag{2.15}$$

If such complex poles are not acceptable, in five dimensions a cannot be positive. More precisely, in five dimensions, a can be zero (which is trivial), negative, or infinity. In the last case, $B(k)$ is simply

$$-i24\pi^2 k^{-3}. \tag{2.16}$$

III. ONE-DIMENSIONAL ARRAY

As an example of utilizing the Fermi pseudopotential (1.4) in five dimensions, we study in this section the case of the one-dimensional equispaced linear chain of infinite length. Before specializing to this case, let a finite or infinite number of Fermi pseudopotentials be located in general at $\mathbf{r}_j$. With $R_j = |\mathbf{r} - \mathbf{r}_j|$, the Hamiltonian is

$$H = -\nabla^2 + \sum_j \delta^5(\mathbf{r} - \mathbf{r}_j) 4\pi^2 a_j^3 \frac{\partial^3}{\partial R_j^3} R_j^3. \tag{3.1}$$

The solution of the Schrödinger equation

$$H\psi = k^2\psi \tag{3.2}$$

follows closely the procedure of Sec. II B. Equation (3.2) is explicitly

$$(\nabla^2 + k^2)\psi = \sum_j A_j \delta^5(\mathbf{r} - \mathbf{r}_j), \tag{3.3}$$

where

$$A_j = 4\pi^2 a_j^3 \frac{\partial^3}{\partial R_j^3} R_j^3 \psi\Big|_{R_j = 0}. \tag{3.4}$$

In the absence of an incident field, the solution of (3.3) is

$$\psi = -\sum_j A_j (8\pi^2 R_j^3)^{-1} e^{ikR_j}(1 - ikR_j). \tag{3.5}$$

Substitution into (3.4) then yields the following linear equations for the determination of A_j and hence ψ from (3.5):

$$(1 + ik^3 a_j^3) A_j + 3a_j^3 \sum_{l \neq j} r_{jl}^{-3} e^{ikr_{jl}}(1 - ikr_{jl}) A_l = 0, \tag{3.6}$$

where $r_{jl} = |\mathbf{r}_j - \mathbf{r}_l|$ is the five-dimensional distance between the points $\mathbf{r}_j$ and $\mathbf{r}_l$. Given a_j and $\mathbf{r}_j$, (3.6) admits a nontrivial solution only for certain values of k: these are the resonances.

We now specialize to the case of an infinite uniform linear array of equal spacing b. Thus

$$r_{jl} = |j - l| b,$$

$$a_j = a,$$

independent of j, and

$$A_j = e^{ij\beta}, \tag{3.7}$$

where β specifies the reduced Hamiltonian. For this case, the fundamental equations (3.6) simplify to one transcendental equation

$$1 + ik^3a^3 + 3\left(\frac{a}{b}\right)^3 \sum_{j\neq 0} |j|^{-3} e^{ik|j|b}(1 - ik|j|b)e^{ij\beta} = 0. \tag{3.8}$$

In three dimensions, the salient feature of such an infinite uniform linear chain is the presence of infinitely narrow resonances. The zero width of the resonance is intimately related to the infinite length of the array. If the array is suitably terminated or bent into a closed curve such as a circle, the width becomes nonzero but small.[11] Such narrow resonances are the quantum-mechanical analog of the corresponding electromagnetic phenomenon associated with the Yagi–Uda antenna array[12] often used with television sets.

This phenomenon also occurs in the present case of a linear chain in five dimensions. This means the existence of real k's as solutions to (3.8) provided that a, b, and β are in suitable ranges. First of all, the zero width of the resonance implies the absence of a radiation field, and hence

$$\beta > kb. \tag{3.9}$$

Let the condition (3.8) be rewritten in the form

$$\frac{1}{3}\left(\frac{b}{a}\right)^3 + \frac{1}{3}ik^3b^3 = -\sum_{j\neq 0} |j|^{-3} e^{ik|j|b}(1 - ik|j|b)e^{ij\beta}, \tag{3.10}$$

then

$$\text{rhs of (3.10)} = \int_0^{e^{i(\beta + kb)}} z^{-1}\,dz \times [\ln(1-z)](i\beta - \ln z) + (\beta \to -\beta). \tag{3.11}$$

Deform the path of integration in the z plane so that it goes from 0 to 1 along the positive real axis and then from 1 to $e^{i(\pm\beta - \ln z)}$ along the unit circle. As the first part of the contour gives a Riemann ζ-function $\zeta(3)$, (3.11) reduces to, by (3.9),

$$\begin{aligned}\text{rhs of (3.10)} = {} & -2\zeta(3) + \int_0^{\beta + kb} d\theta\,(\theta - \beta) \\ & \times\left[\ln\left(2\sin\frac{\theta}{2}\right) - \frac{1}{2}i(\pi - \theta)\right] \\ & + \int_0^{-\beta + kb} d\theta\,(\theta + \beta) \\ & \times\left[\ln\left(-2\sin\frac{\theta}{2}\right) + \frac{1}{2}i(\pi + \theta)\right].\end{aligned} \tag{3.12}$$

Thus the imaginary part can be explicitly calculated as

$$\text{Imaginary part of rhs of (3.10)} = \tfrac{1}{3}k^3b^3, \tag{3.13}$$

and consequently (3.10) has the real form

$$\frac{1}{3}\left(\frac{b}{a}\right)^3 + 2\zeta(3) - \left(\int_0^{\beta + kb} + \int_0^{\beta - kb}\right)d\theta\,(\theta - \beta) \times \ln\left(2\sin\frac{\theta}{2}\right) = 0. \tag{3.14}$$

Remember that $a < 0$ while $b > 0$. The solutions of (3.14) give the locations of infinitely narrow resonances.

IV. RESOLVENT EQUATION

The example of the last section shows that the Fermi pseudopotential in five dimensions is of use. We now return to the case of a single Fermi pseudopotential already treated in Sec. III B to study some simple properties of the Green's function (2.12). Let the Hamiltonian be

$$H = -\nabla^2 + 4\pi^2 a^3 \delta^5(\mathbf{r}) \frac{\partial^3}{\partial r^3} r^3, \tag{4.1}$$

then $G(\mathbf{r}, \mathbf{r}_0; k)$ is the coordinate representation of $(H - k^2)^{-1}$. Since H commutes with itself, this operator satisfies the resolvent equation

$$\begin{aligned}&(k^2 - k'^2)(H - k^2)^{-1}(H - k'^2)^{-1} \\ &\quad = (H - k^2)^{-1} - (H - k'^2)^{-1}.\end{aligned} \tag{4.2}$$

Expressed in terms of $G(\mathbf{r}, \mathbf{r}_0; k)$, (4.2) is

$$\begin{aligned}&(k^2 - k'^2)\int d^5r''\, G(\mathbf{r}, \mathbf{r}''; k)G(\mathbf{r}', \mathbf{r}''; k') \\ &\quad = G(\mathbf{r}, \mathbf{r}'; k) - G(\mathbf{r}, \mathbf{r}'; k').\end{aligned} \tag{4.3}$$

In what sense is this formula correct?

The difficulty stems from the fact that in five dimensions the free-space Green's function is not square integrable, as discussed in the Introduction. Thus, in the usual sense of integration, the integral on the left-hand side of (4.3) does not exist.

Fortunately, since the divergent integral is of the form $\int r^{-6}\, d^5\mathbf{r}$, formulas familiar in dimensional regularization[13,14] can be invoked. The relevant one is

$$\int r^{-6}\, d^5\mathbf{r} = 0. \tag{4.4}$$

Equation (4.4) gives a meaning to the left-hand side of (4.3), because the G_0 of (2.11) is of the form

$$G_0(\mathbf{r}, 0; k) = (8\pi^2)^{-1} r^{-3}[1 + O(r^2)], \tag{4.5}$$

for small r. The rest of the verification for (4.3) is straightforward.

V. DISCUSSIONS

We have seen that the Fermi pseudopotential (1.1) in three dimensions can be naturally generalized to (1.4) or equivalently (1.5) in five dimensions. With this natural generalization, the formal manipulations are only slightly changed. The underlying mathematical structure, however, is altered in a profound manner. While $-\nabla^2 + (1.1)$ gives rise to a self-adjoint operator in the space of square integrable functions, it is not clear whether (4.1) is self-adjoint in any sense. A first question that needs to be answered is: Does there exist a Hilbert space of functions such that the resolvent as explicitly given by (2.12) is bounded and satisfies the resolvent equation (4.2)? We do not know the answer to this and numerous related questions. By explicit examples, we do know, however, that the Fermi pseudopotential (1.4) in five dimensions is useful. It is a challenge to mathematical physicists to construct a consistent theory of this new Fermi pseudopotential.

There is one other fundamental difference between (1.1) and (1.4). In three dimensions, (1.1) can be approximated by a short-range square-well attractive potential of suitable strength. This is the basis for the successful application of nonstandard analysis.[8] On the contrary, in five dimensions, such an approximation does not seem to exist for the Fermi pseudopotential (1.4).

We conclude with a list of Fermi pseudopotentials in two to six dimensions, where the overall constant is omitted: two dimensions,

$$\delta^2(\mathbf{r}) r (\ln r)^2 \frac{\partial}{\partial r} (\ln r)^{-1};$$

three dimensions,

$$\delta^3(\mathbf{r}) \frac{\partial}{\partial r} r;$$

four dimensions,

$$\delta^4(\mathbf{r}) r (\ln r)^2 \frac{\partial}{\partial r} (\ln r)^{-1} r^{-1} \frac{\partial}{\partial r} r^2;$$

five dimensions,

$$\delta^5(\mathbf{r}) \frac{\partial^3}{\partial r^3} r^3;$$

and six dimensions

$$\delta^2(\mathbf{r}) r (\ln r)^2 \frac{\partial}{\partial r} (\ln r)^{-1} r^{-1} \frac{\partial}{\partial r} r^{-1} \frac{\partial}{\partial r} r^4.$$

There may be some problems with the last one.

ACKNOWLEDGMENTS

One of us (TTW) thanks Professor Hans Joos, Professor Erich Lohrmann, Professor Volker Soergel, and Professor Kurt Symanzik for their kind hospitality at DESY.

Work supported in part by the United States Department of Energy under Contract No. DE-AS02-76ER03227.

[1]E. Fermi, Ric. Sci. **7**, 13 (1936).
[2]G. Breit, Phys. Rev. **71**, 215 (1947); G. Breit and P. R. Zilsel, Phys. Rev. **71**, 232 (1947).
[3]J. M. Blatt and V. F. Weisskopf, *Theoretical Nuclear Physics* (Wiley, New York, 1952), pp. 74–75.
[4]K. Huang and C. N. Yang, Phys. Rev. **105**, 767 (1957).
[5]K. Huang, C. N. Yang, and J. M. Luttinger, Phys. Rev. **105**, 776 (1957).
[6]T. D. Lee, K. Huang, and C. N. Yang, Phys. Rev. **106**, 1134 (1957).
[7]T. T. Wu, Phys. Rev. **115**, 1390 (1959).
[8]S. Albeverio, J. E. Fenstad, and R. Hoegh-Krọhn, Trans. Am. Math. Soc. **252**, 275 (1979).
[9]L. E. Thomas, J. Math. Phys. **20**, 1848 (1979).
[10]A. Grossmann, R. Hoegh-Krøhn, and M. Mebkhout, J. Math. Phys. **21**, 2376 (1980); Commun. Math. Phys. **77**, 87 (1980).
[11]A. Grossmann and T. T. Wu, Preprint Centre Physique Theorique, Marseilles, 1981.
[12]R. W. P. King, R. B. Mack, and S. S. Sandler, *Arrays of Cylindrical Dipoles* (Cambridge U.P., Cambridge, 1968), Chap. 6.
[13]C. G. Bollini and J. J. Giambiagi, Nuovo Cimento B **12**, 20 (1972).
[14]G. 't Hooft and M. Veltman, Nuclear Phys. B **44**, 189 (1972).

General nonlinear realization of chiral $SU_2 \times SU_2$ symmetry in curved isospin space

Abdul Ebrahim and Mustafa Savci
Physics Department, Middle East Technical University, Ankara, Turkey

(Received 30 August 1983; accepted for publication 7 October 1983)

The geometric method for constructing the nonlinear chiral $SU_2 \times SU_2$ invariant Yang–Mills Lagrangian in curved isospin space given by Meetz for a particular choice of coordinates is generalized to an arbitrary system of coordinates in curved isospin space. The resulting Lagrangian coincides with the general nonlinear chiral $SU_2 \times SU_2$ invariant Yang–Mills Lagrangian constructed previously using the matrix method.

PACS numbers: 02.40.Ky, 11.15.Kc, 11.30.Na, 11.30.Rd

I. INTRODUCTION

The purpose of this paper is to generalize the geometric method suggested by Meetz[1] for constructing the nonlinear chiral $SU_2 \times SU_2$ invariant Yang–Mills Lagrangian for the pion–nucleon–ρ-meson (π–N–ρ) system using curved isospin space. Meetz' construction was based on a particularly simple choice of coordinates in curved isospin space such that the chiral $SU_2 \times SU_2 \approx SO(4)$ four-vector ξ^α ($\alpha = 1,2,3,4$) was identified with the isovector pion field φ^a ($a = 1,2,3$) according to $\xi^\alpha = (\varphi^a, \xi^4)$. The four-vector ξ^α transforms linearly as the $(\frac{1}{2},\frac{1}{2})$ representation of the chiral $SU_2 \times SU_2$ group, and hence its length $\delta_{\alpha\beta}\ \xi^\alpha \xi^\beta$ is an $SU_2 \times SU_2$ invariant. A suitable normalization of this length then determines ξ^4 in terms of φ^a and hence defines a three-dimensional curved isospin space embedded in the four-dimensional Euclidean space spanned by ξ^α. The three components of the pion field φ^a can be used as general curvilinear coordinates in this three-dimensional curved isospin space. The methods of Riemannian geometry as well as the techniques of Yang–Mills theory can then be employed for constructing a nonlinear chiral $SU_2 \times SU_2$ invariant Yang–Mills Lagrangian for the π–N–ρ system. The resulting Lagrangian turns out to be a special case of the general nonlinear chiral $SU_2 \times SU_2$ invariant Yang–Mills Lagrangian for the π–N–ρ system constructed previously using the matrix method.[2] The reason for this lack of generality in the geometric method of construction used by Meetz is that the identification of the four-vector ξ^α in terms of the pion field φ^a assumed by Meetz is not the most general possible. In the following we will identify the components of the four-vector ξ^α in terms of the pion field φ^a according to $\xi^\alpha = (2f\rho\varphi^a, \sigma)$, where σ and ρ are functions of the dimensionless variable $f^2\varphi^2$, f is a coupling parameter with the dimension of length, and $\varphi^2 = \delta_{ij}\,\varphi^i\varphi^j$ ($i,j = 1,2,3$). Even though a normalization of the length of the four-vector ξ^α will put some restriction on the form of the functions σ and ρ, the identification of the four-vector ξ^α in terms of the pion field adopted here is the most general possible. This identification coupled with the geometric method of Meetz leads to the most general nonlinear chiral $SU_2 \times SU_2$ invariant Yang–Mills Lagrangian for the π–N–ρ system that was obtained previously using the matrix method.

II. CURVED ISOSPIN SPACE

We begin by considering a four-dimensional internal Euclidean space spanned by a real, dimensionless four-vector field $\xi^\alpha(x)$. Under the chiral $SU_2 \times SU_2 \approx SO(4)$ group, ξ^α is assumed to transform linearly as the $(\frac{1}{2},\frac{1}{2})$ representation according to

$$\xi^{\alpha'} = R^\alpha{}_\beta\, \xi^\beta, \quad \alpha, \beta = 1,2,3,4, \tag{1a}$$

where the 4×4 matrix R satisfies the conditions

$$RR^T = R^TR = I, \tag{1b}$$

$$\det R = 1\,. \tag{1c}$$

The length $\delta_{\alpha\beta}\ \xi^\alpha\xi^\beta$ of the four-vector ξ^α is an $SU_2 \times SU_2 \approx SO(4)$ invariant, and without loss of generality we will normalize this length to unity so that

$$\delta_{\alpha\beta}\ \xi^\alpha\xi^\beta \equiv \boldsymbol{\xi}\cdot\boldsymbol{\xi} + (\xi^4)^2 = 1, \tag{1d}$$

where $\boldsymbol{\xi} = (\xi^1, \xi^2, \xi^3)$ and $\boldsymbol{\xi}\cdot\boldsymbol{\xi} = \delta_{ij}\,\xi^i\xi^j$. This normalization condition defines a three-dimensional curved isospin space with constant curvature embedded in the four-dimensional Euclidean space. We will restrict our attention to the upper hemisphere of the three-dimensional curved isospin space defined by

$$\xi^4 = +\sqrt{1 - \boldsymbol{\xi}\cdot\boldsymbol{\xi}}, \tag{1e}$$

and we will choose the three components of the pion field φ^a as general curvilinear coordinates in this three-dimensional curved isospin space. Further, we will identify the components of the four-vector ξ^α in terms of the components of the pion field according to

$$\xi^\alpha = (2f\rho\varphi^a, \sigma). \tag{2a}$$

The normalization condition (1d) gives a relation between σ and ρ of the form

$$\sigma^2 + 4f^2\rho^2\varphi^2 = 1, \tag{2b}$$

so that, in general, σ and ρ are given by[3]

$$\sigma(f^2\varphi^2) = 1 - 2f^2\varphi^2 + O(f^4), \tag{2c}$$

$$\rho(f^2\varphi^2) = 1 + O(f^2), \tag{2d}$$

but are otherwise arbitrary. Our identification of the four-vector ξ^α in terms of the pion field given in Eq. (2a) is the most general possible. The special case considered by Meetz

 0022-2488/84/061746-06$02.50

corresponds to the choice $\rho = 1$.

The chiral $SU_2\times SU_2 \approx SO(4)$ group is represented linearly in the four-dimensional Euclidean space as

$$\xi^{i'} = R^i{}_j\, \xi^j + R^i{}_4\, \xi^4, \tag{3a}$$

$$\xi^{4'} = R^4{}_j\, \xi^j + R^4{}_4\, \xi^4. \tag{3b}$$

However, in the three-dimensional curved isospin space with the pion field φ^a as coordinates, the chiral $SU_2\times SU_2$ group induces a nonlinear transformation on the pion field according to

$$(2f\rho\varphi^a)' = R^a{}_b(2f\rho\varphi^b) + R^a{}_4\sigma, \tag{3c}$$

$$(\sigma)' = R^4{}_4\sigma + R^4{}_b(2f\rho\, \varphi^b). \tag{3d}$$

In order to obtain the explicit form of the infinitesimal chiral $SU_2\times SU_2$ transformations of the pion field, we write the chiral $SU_2\times SU_2$ transformation matrix $R^\alpha{}_\beta$ as

$$R^\alpha{}_\beta = \delta^\alpha{}_\beta + \epsilon^\alpha{}_\beta, \tag{4a}$$

where $\epsilon^\alpha{}_\beta$ are the infinitesimal chiral $SU_2\times SU_2$ parameters satisfying

$$\epsilon^\alpha{}_\beta + \epsilon_\beta{}^\alpha = 0. \tag{4b}$$

Then from Eqs. (3c) and (3d) we find that under the infinitesimal chiral $SU_2\times SU_2$ transformations, the quantities $(2f\rho\varphi^a)$ and σ transform according to

$$\delta(2f\rho\varphi^a) = \epsilon^a{}_b(2f\rho\varphi^b) + \epsilon^a{}_4\sigma, \tag{4c}$$

$$\delta\sigma = \epsilon^4{}_b(2f\rho\varphi^b), \tag{4d}$$

and if we define

$$\epsilon^a{}_b\, \varphi^b = \epsilon^{abc}\, \varphi_b\omega_c, \tag{4e}$$

$$\epsilon^a{}_4 = -\nu^a, \tag{4f}$$

where ω^a and ν^a are, respectively, the real infinitesimal isospin and chiral parameters and ϵ^{abc} is the Levi-Civita tensor with $\epsilon^{123} = 1$, then the infinitesimal chiral $SU_2\times SU_2$ transformation for the pion field can be written as[4]

$$\delta\varphi^a = -f^{ab}(\varphi)\omega_b - F^{ab}(\varphi)\nu_b, \tag{4g}$$

where

$$f^{ab}(\varphi) = \epsilon^{abc}\varphi_c, \tag{4h}$$

$$F^{ab}(\varphi) = (1/2f)[(\sigma/\rho)\delta^{ab} + 4f^2\varphi^a\varphi^b(\rho'/\sigma')], \tag{4i}$$

and a prime on σ and ρ denotes a derivative with respect to the dimensionless variable $f^2\varphi^2$. We note that the isospin subgroup of the chiral $SU_2\times SU_2$ group is represented linearly on the pion field, whereas under the chiral transformations the pion field transforms nonlinearly.

The infinitesimal line element in the four-dimensional Euclidean space given by

$$ds^2 = (1/4f^2)\delta_{\alpha\beta}\, d\xi^\alpha\, d\xi^\beta \tag{5a}$$

is an $SU_2\times SU_2 \approx SO(4)$ invariant. In the three-dimensional curved isospin space, the four-vector ξ^α is specified in terms of the pion field by Eq. (2a) so that

$$\xi^\alpha = \xi^\alpha(\varphi^a), \tag{5b}$$

and hence

$$d\xi^\alpha = \frac{\partial\xi^\alpha}{\partial\varphi^a}\, d\varphi^a. \tag{5c}$$

Thus the $SU_2\times SU_2$ invariant line element (5a) can be written in terms of the pion field as

$$\begin{aligned} ds^2 &= (1/4f^2)\delta_{\alpha\beta}\, \frac{\partial\xi^\alpha}{\partial\varphi^a}\frac{\partial\xi^\beta}{\partial\varphi^b}\, d\varphi^a\, d\varphi^b \\ &\equiv g_{ab}(\varphi) d\varphi^a\, d\varphi^b, \end{aligned} \tag{5d}$$

where $g_{ab}(\varphi)$ is the covariant metric in the three-dimensional curved isospin space defined by

$$g_{ab}(\varphi) = (1/4f^2)\frac{\partial\xi^\alpha}{\partial\varphi^a}\frac{\partial\xi^\beta}{\partial\varphi^b}\, \delta_{\alpha\beta}. \tag{5e}$$

From Eqs. (2a) and (5e) we find that[3],

$$g_{ab}(\varphi) = \rho^2\left[\delta_{ab} + f^2\varphi_a\,\varphi_b\left\{\frac{\sigma'^2 - 4\rho^4}{\rho^2(1-\sigma^2)}\right\}\right], \tag{5f}$$

and the corresponding contravariant metric defined according to

$$g^{ab}g_{bc} = \delta^a{}_c, \tag{5g}$$

is given by

$$g^{ab}(\varphi) = \frac{1}{\rho^2}\left[\delta^{ab} - f^2\varphi^a\varphi^b\left\{\frac{4\rho^2(\sigma'^2 - 4\rho^4)}{\sigma'^2(1-\sigma^2)}\right\}\right]. \tag{5h}$$

If we define the square roots of the covariant and contravariant metrices as

$$(g^{1/2})_{ab}(g^{1/2})_{cd}\, g^{bc} = g_{ad}, \tag{5i}$$

$$(g^{1/2})^{ab}(g^{1/2})_{bc} = \delta^a{}_c, \tag{5j}$$

then we find that

$$(g^{1/2})_{ab} = \rho\left[\delta_{ab} + 2f^2\varphi_a\,\varphi_b\left\{\frac{\rho'}{\rho} - \frac{\sigma'}{1+\sigma}\right\}\right], \tag{5k}$$

$$(g^{1/2})^{ab} = \frac{1}{\rho}\left[\delta^{ab} - 4f^2\varphi^a\varphi^b\left\{\frac{\rho^2}{1+\sigma} - \frac{\rho\rho'}{\sigma'}\right\}\right]. \tag{5l}$$

Under an $SU_2\times SU_2$ coordinate transformation in curved isospin space defined by

$$\varphi^{a'} = \varphi^{a'}(\varphi), \tag{6a}$$

the infinitesimal $d\varphi^a$ transforms as a contravariant vector. Moreover, the space-time derivative $\partial_\mu\,\varphi^a$ also transforms as a contravariant vector since we have

$$\partial\mu\,\varphi^{a'} = \frac{\partial\varphi^{a'}}{\partial\varphi^b}\,\partial_\mu\,\varphi^b. \tag{6b}$$

The transformation law for the covariant metric $g_{ab}(\varphi)$ follows from Eq. (5e) as

$$\begin{aligned} g_{ab}{}'(\varphi') &\equiv \frac{\partial\xi^{\alpha'}}{\partial\varphi^{a'}}\frac{\partial\xi^{\beta'}}{\partial\varphi^{b'}}\,\delta_{\alpha\beta} \\ &= \frac{\partial\varphi^c}{\partial\varphi^{a'}}\frac{\partial\varphi^d}{\partial\varphi^{b'}}\, g_{cd}(\varphi), \end{aligned} \tag{6c}$$

so that it transforms as a second-rank covariant tensor, and hence by Eq. (5g) the contravariant metric $g^{ab}(\varphi)$ transforms as a second-rank contravariant tensor. Thus from Eqs. (6b) and (6c), we can define the covariant space-time derivative of φ^a as

$$\partial_\mu\varphi_a = g_{ab}(\varphi)\partial_\mu\,\varphi^b. \tag{6d}$$

Also since the chiral $SU_2\times SU_2$ group is an invariance group of the infinitesimal line element (5a), we have the isometry relation[5]

$$g_{ab}'(\varphi') = g_{ab}(\varphi'). \tag{6e}$$

The chiral $SU_2\times SU_2$ invariant Lagrangian for the pion field that is at most quadratic in space-time derivatives of the pion field and that reduces to the free pion Lagrangian in lowest order in the coupling parameter f is now uniquely given by

$$\mathscr{L}_\pi = \tfrac{1}{2} g_{ab}(\varphi)\partial_\mu \varphi^a \partial_\nu \varphi^b \eta^{\mu\nu}, \tag{7a}$$

where $\eta^{\mu\nu}$ is the Lorentz metric tensor defined as $\eta_{\mu\nu} = \eta^{\mu\nu} = \mathrm{diag}(1, -1, -1, -1)$. The corresponding chiral $SU_2\times SU_2$ invariant action for the pion field given by

$$I = \tfrac{1}{2}\int d^4x\, \eta^{\mu\nu}\partial_\mu \varphi^a \partial_\nu \varphi^b g_{ab}(\varphi), \tag{7b}$$

has the form of field theories that are harmonic mappings of Riemannian manifolds.[6] The chiral covariant equation of motion for the pion field obtained by varying the action given in Eq. (7b) is

$$\eta_{\mu\nu}[\partial^\mu\partial^\nu\varphi^a + \Gamma^a_{bc}(\varphi)\partial^\mu\varphi^b\partial^\nu\varphi^c] = 0, \tag{7c}$$

where $\Gamma^a_{bc}(\varphi)$ is the Christoffel symbol for the curved isospin metric $g_{ab}(\varphi)$ defined as

$$\Gamma^a_{bc} = \tfrac{1}{2} g^{ad}[\partial_b g_{cd} + \partial_c g_{bd} - \partial_d g_{bc}], \tag{7d}$$

where

$$\partial_c g_{ab} = \frac{\partial g_{ab}}{\partial \varphi^c}.$$

The chiral covariant derivative of a contravariant vector $v^a(\varphi)$ in curved isospin space can be defined as usual in terms of the Christoffel symbol according to[5]

$$\nabla_b v^a = \partial_b v^a + \Gamma^a_{bc} v^c, \tag{8a}$$

where

$$\partial_a v^b = \frac{\partial v^b}{\partial \varphi^a},$$

so that $\nabla_b v^a$ transforms as a second-rank mixed tensor in curved isospin space. If we define a chiral covariant space-time derivative of a contravariant vector in curved isospin space according to

$$D_\mu v^a = (\nabla_b v^a)\partial_\mu \varphi^b = \partial_\mu v^a + \Gamma^a_{bc} v^c \partial_\mu \varphi^b, \tag{8b}$$

where

$$\partial_\mu v^a = (\partial_b v^a)\partial_\mu \varphi^b, \tag{8c}$$

then $D_\mu v^a$ transforms as a covariant vector in Lorentz space and a contravariant vector in curved isospin space. In case the vector v^a has an explicit space-time dependence not arising from its dependence on the pion field, Eq. (8c) must be amended as

$$\partial_\mu v^a = (\partial_b v^a)\partial_\mu \varphi^b + \hat{\partial}_\mu v^a, \tag{8d}$$

where the caret denotes a space-time derivative due to the explicit space-time dependence of v^a.

We close this section by defining the triad or dreibein fields $e^a{}_i(\varphi)$ according to

$$e^a{}_i(\varphi) e^b{}_j(\varphi)\delta^{ij} = g^{ab}(\varphi), \tag{9a}$$

$$e^a{}_i(\varphi) e^b{}_j(\varphi) g_{ab}(\varphi) = \delta_{ij}. \tag{9b}$$

Then from Eqs. (5i) and (5j) we find that

$$e^a{}_i(\varphi) = (g^{1/2})^a{}_i. \tag{9c}$$

We note that the first index on $e^a{}_i(\varphi)$ transforms contravariantly in curved isospin space while the second index is a flat isospin index.

The chiral $SU_2\times SU_2$ invariant pion Lagrangian given in Eq. (7a) can now be written as

$$\mathscr{L}_\pi = \tfrac{1}{2} g_{ab}(\varphi)\partial_\mu \varphi^a \partial_\nu \varphi^b \eta^{\mu\nu} = 2f^2 X^i_{5\mu} X^j_{5\nu}\delta_{ij}\eta^{\mu\nu}, \tag{9d}$$

where

$$X^i_{5\mu} = (1/2f)\delta^{ij} e^a{}_j(\varphi)\partial_\mu \varphi_a, \tag{9e}$$

so that $X^i_{5\mu}$ is the flat isospin-space component of the curved isospin-space vector $\partial_\mu \varphi_a$.

III. CHIRAL $SU_2\times SU_2$ PION–NUCLEON LAGRANGIAN

In this section we proceed to determine the chiral $SU_2\times SU_2$ transformation law for the nucleon field in curved isospin space in order to be able to define a chiral covariant space-time derivative of the nucleon field. This and the result of the preceding section will then permit the construction of a general nonlinear chiral $SU_2\times SU_2$ invariant pion–nucleon Lagrangian.

In flat isospin space, the nucleon field transforms according to the two-dimensional fundamental representation of the isospin SU(2) group generated by the 2×2 Pauli matrices τ^i satisfying the anticommutation relation

$$\{\tau^i, \tau^j\} = 2\delta^{ij}. \tag{10a}$$

For the curved isospin space with metric $g^{ab}(\varphi)$, we generalize the flat isospin Clifford algebra (10a) by defining the 2×2 matrices $L^a(\varphi)$ that depend on the pion field and that satisfy[7]

$$\{L^a(\varphi), L^b(\varphi)\} = 2g^{ab}(\varphi). \tag{10b}$$

Under the chiral $SU_2\times SU_2$ coordinate transformations in curved isospin space, the matrices $L^a(\varphi)$ are assumed to transform as components of a contravariant vector so that

$$L^{a'}(\varphi') = \frac{\partial \varphi^{a'}}{\partial \varphi^b} L^b(\varphi). \tag{10c}$$

An explicit form for the matrices $L^a(\varphi)$ is obtained by observing that the 2×2 matrices defined by

$$L^a(\varphi) = (g^{1/2})^a{}_i \tau^i = e^a{}_i \tau^i, \tag{10d}$$

satisfy the Clifford algebra (10b) by virtue of the relation (9a). However, this solution for the matrices $L^a(\varphi)$ is not unique since a transformation on the flat isospin index of $e^a{}_i(\varphi)$ of the form

$$e^a{}_i(\varphi)' = A_i{}^j e^a{}_j(\varphi), \tag{10e}$$

will leave the Clifford algebra (10b) unchanged provided the 2×2 matrices $A_i{}^j$ satisfy the condition

$$\delta^{ij} A_i{}^n A_j{}^m = \delta^{nm}. \tag{10f}$$

Thus the solution (10d) for the matrices $L^a(\varphi)$ is unique up to an orthogonal transformation on the flat isospin space index of $e^a{}_i(\varphi)$. We will adopt the particular solution given by Eqs. (10d) and (51).

From the Clifford algebra (10b) and the solution (10d) for the matrices $L^a(\varphi)$, it follows that the ordinary derivative of $L^a(\varphi)$ with respect to the pion field can be put in the form

$$\partial_b L^a(\varphi) = \tfrac{1}{2}(\partial_b g^{ac})L_c(\varphi) + [\Sigma_b, L^a(\varphi)], \qquad (11a)$$

where

$$\Sigma_b = -\tfrac{1}{8}[L_a(\varphi), \partial_b L^a(\varphi)], \qquad (11b)$$

Since g^{ab} and $L^a(\varphi)$ transform, respectively, as second-rank contravariant tensor and contravariant vector in curved isospin space, the isometry relation (6e) implies that the matrices $L^{a'}(\varphi')$ satisfy the relation

$$\{L^{a'}(\varphi'), L^{b'}(\varphi')\} = 2g^{ab'}(\varphi') = 2g^{ab}(\varphi'), \qquad (12a)$$

and from Eq. (10b) we have

$$\{L^a(\varphi'), L^b(\varphi')\} = 2g^{ab}(\varphi'), \qquad (12b)$$

and hence there exists a unitary matrix $S(\varphi)$ depending on the pion field such that

$$L^{a'}(\varphi') = S^{-1}L^a(\varphi')S. \qquad (12c)$$

To determine the chiral transformation law for the nucleon field $N(\varphi)$ we require that the bilinear $\overline{N}(\varphi)L^a(\varphi)N(\varphi)$ transform as a contravariant vector in curved isospin space so that

$$\overline{N}'(\varphi')L^a(\varphi')N'(\varphi') = \frac{\partial\varphi^{a'}}{\partial\varphi^b}\overline{N}(\varphi)L^b(\varphi)N(\varphi), \qquad (12d)$$

where we assume that the matrices $L^a(\varphi)$ have the same functional form in all coordinate systems related by chiral $SU_2 \times SU_2$ transformations, and from Eqs. (12c) and (10c) we have

$$\overline{N}'(\varphi')L^a(\varphi')N'(\varphi') = \frac{\partial\varphi^{a'}}{\partial\varphi^b}\overline{N}'(\varphi')SL^b(\varphi)S^{-1}N'(\varphi'), \qquad (12e)$$

so that a comparison of Eqs. (12d) and (12e) gives

$$N'(\varphi') = SN(\varphi). \qquad (12f)$$

In order to determine the unitary matrix S, we write its infinitesimal form as

$$S = I + T. \qquad (13a)$$

Then from Eqs. (12b) and (12e) it follows that

$$[T, L^a(\varphi)] = (\partial_b L^a)\delta\varphi^b - L^b(\varphi)\partial_b(\delta\varphi^a), \qquad (13b)$$

where from Eqs. (4g) and (4i)

$$\delta\varphi^a = -F^{ab}(\varphi)\nu_b. \qquad (13c)$$

Using the form of $L^a(\varphi)$ given in Eq. (10d), we find that Eq. (13b) gives

$$T = \Sigma_a\delta\varphi^a + \tfrac{1}{8}[L_a, L^b]\partial_b(\delta\varphi^a), \qquad (13d)$$

which can be written as

$$T = (i4)\epsilon^{ijk}\tau_k e_{ai}(\varphi)[e^b{}_j(\varphi)\partial_b(\delta\varphi^a) - \partial_b e^a{}_j(\varphi)\delta\varphi^b]. \qquad (13e)$$

Finally using the explicit form for $e^a{}_i(\varphi)$ and $\delta\varphi^a$ given in Eqs. (9c) and (13c), respectively, we obtain

$$T = [if\rho/(1+\sigma)]\epsilon^{ijk}\tau_i\nu_j\varphi_k \equiv (i/2)\boldsymbol{\tau}\cdot\boldsymbol{\beta}, \qquad (13f)$$

where

$$\beta^i = [2f\rho/(1+\sigma)]\epsilon^{ijk}\nu_j\varphi_k \equiv [2f\rho/(1+\sigma)](\boldsymbol{\nu}\times\boldsymbol{\varphi})^i, \qquad (13g)$$

and hence the matrix S is given by

$$S = \exp[(i/2)]\boldsymbol{\tau}\cdot\boldsymbol{\beta}. \qquad (13h)$$

We observe that the chiral transformation law for the nucleon field $N(\varphi)$ is similar to a local isospin transformation with the transformation parameters $\boldsymbol{\beta}$ depending on the pion field. This similarity will be exploited below to construct a chiral $SU_2 \times SU_2$ invariant Yang–Mills Lagrangian for the π-N-ρ system.

The matrix S given in Eq. (13h) determines the chiral transformation law for the nucleon field N according to Eq. (12f) as well as the chiral transformation law for the 2×2 matrices $L^a(\varphi)$ according to Eq. (12c). Further, since the dreibein fields $e^a{}_i(\varphi)$ are related to $L^a(\varphi)$ according to Eq. (10d), the same matrix S also determines the chiral transformation law for the dreibeins. If we write this transformation law in the form

$$e^{a'}{}_i(\varphi') = B_{ij}e^{aj}(\varphi'), \qquad (13i)$$

then we find from Eqs. (10d), (12c), and (13h) that the 3×3 matrix B is given by

$$B_{ij}(\varphi) = \cos\beta\delta_{ij} - \left(\frac{\sin\beta}{\beta}\right)\epsilon_{ijk}\beta^k + \left(\frac{1-\cos\beta}{\beta^2}\right)\beta_i\beta_j, \qquad (13j)$$

where $\beta = |\boldsymbol{\beta}|$. We note that the isometry relation (6e) for the metric and the defining relation (9a) for the dreibeins give

$$g'_{ab}(\varphi') = e'_{ai}(\varphi')e'_{bj}(\varphi')\delta^{ij} = g_{ab}(\varphi') = e_{ai}(\varphi')e_{bj}(\varphi')\delta^{ij}, \qquad (13k)$$

which implies that the transformation (13i) must be orthogonal. It is readily verified that the matrix B given above is indeed orthogonal.

The ordinary derivative of $L^a(\varphi)$ with respect to the pion field is given in Eq. (11a) and we define the covariant derivative of $L^a(\varphi)$ using Eq. (8a) as

$$\nabla_b L^a = \partial_b L^a + \Gamma_b{}^a{}_c L^c. \qquad (14a)$$

Further for the metric tensor $g^{ab}(\varphi)$ we have in general

$$\nabla_c g^{ab} = 0. \qquad (14b)$$

Then if we write the covariant derivative of $L^a(\varphi)$ as

$$\nabla_b L^a = [\Omega_b, L^a], \qquad (14c)$$

then it follows that the 2×2 matrix Ω_b is given by

$$\Omega_b = \epsilon_b + \tfrac{1}{8}[L^a, L^c]\partial_a g_{bc} = -\tfrac{1}{8}[L_a, \nabla_b L^a], \qquad (14d)$$

and using the solution for $L^a(\varphi)$ given in Eq. (10d), we find that

$$\Omega_b = (i/4)\epsilon^{ijk}\tau_k e^c{}_j[e^a{}_i\partial_a g_{bc} + (\partial_b e^a{}_i)g_{ac}]. \qquad (14e)$$

Finally, using the explicit expression for $e^a{}_i(\varphi)$ given in Eq. (9c) we obtain

$$\Omega_b = [2if^2\rho^2/(1+\sigma)]\, \epsilon_{bij}\, \varphi^i \tau^j. \tag{14f}$$

The covariant derivative of the nucleon field $N(\varphi)$ in curved isospin space is defined to satisfy

$$\Delta_b L^a = \nabla_b L^a - [\Omega_b, L^a] = 0, \tag{15a}$$

$$\Delta_b(\overline{N}L^a N) \equiv \nabla_b(\overline{N}L^a N) = (\overline{\Delta_b N})L^a N + \overline{N}L^a(\Delta_b N). \tag{15b}$$

Then it follows that

$$\Delta_a N = \partial_a N - \Omega_a N, \tag{15c}$$

$$(\overline{\Delta_a N}) = (\overline{\partial_a N}) + \overline{N}\Omega_a, \tag{15d}$$

where a bar over a Dirac spinor ψ is defined as

$$\bar{\psi} = \psi^\dagger \gamma^0. \tag{15e}$$

In analogy with Eq. (8b), we now define a chiral covariant space-time derivative of the nucleon field $N(\varphi)$ as

$$D_\mu N = (\Delta_a N)\partial_\mu \varphi^a = \partial_\mu N - \Omega_a N \partial_\mu \varphi^a, \tag{16a}$$

$$\overline{D_\mu N} = (\overline{\Delta_a N})\,\partial_\mu \varphi^a = \overline{\partial_\mu N} + \overline{N}\Omega_a\, \partial_\mu \varphi^a, \tag{16b}$$

where

$$\partial_\mu N = (\partial_a N)\partial_\mu \varphi^a + \hat{\partial}_\mu N, \tag{16c}$$

and the last term in Eq. (16c) is present if the nucleon field has an explicit space-time dependence not arising from its dependence on the pion field. Using the explicit expression for Ω_a given in Eq. (14f), we can write the chiral covariant space-time derivative of the nucleon field as

$$D_\mu N = (\partial_\mu + if^2\, \boldsymbol{\tau}\cdot\mathbf{X}_\mu)N, \tag{17a}$$

where

$$\boldsymbol{\tau}\cdot\mathbf{X}_\mu = \delta_{ij}\, \tau^i X_\mu{}^j, \tag{17b}$$

$$X_\mu{}^i = [2\rho^2/(1+\sigma)]\, \epsilon^{ijk} \varphi_j \partial_\mu \varphi_k. \tag{17c}$$

The chiral $SU_2\times SU_2$ invariant Lagrangian for the nucleon field $N(\varphi)$ can now be written as

$$\mathscr{L}_N = \overline{N}i\gamma^\mu(\partial_\mu + if^2\boldsymbol{\tau}\cdot\mathbf{X}_\mu)N - m\overline{N}N \tag{18a}$$

and if we add to this the chiral $SU_2\times SU_2$ invariant pion Lagrangian given in Eq. (7a), we obtain the chiral $SU_2\times SU_2$ invariant pion–nucleon Lagrangian in the form

$$\mathscr{L}^{(0)}_{\pi N} = \overline{N}i\gamma^\mu(\partial_\mu + if^2\boldsymbol{\tau}\cdot\mathbf{X}_\mu)N - m\overline{N}N + \tfrac{1}{2} g_{ab}(\varphi)\partial_\mu \varphi^a \partial^\mu \varphi^b. \tag{18b}$$

This Lagrangian is invariant under the infinitesimal isospin transformations

$$\varphi^a \to \varphi^a - f^{ab}(\varphi)\omega_b, \tag{19a}$$

$$N \to (1 + (i/2)\boldsymbol{\tau}\cdot\boldsymbol{\omega})N, \tag{19b}$$

as well as under the infinitesimal chiral transformations

$$\varphi^a \to \varphi^a - F^{ab}(\varphi)\nu_b, \tag{19c}$$

$$N \to (1 + (i/2)\, \boldsymbol{\tau}\cdot\boldsymbol{\beta})N. \tag{19d}$$

The chiral covariant equation of motion for the nucleon and pion fields obtained from Eq. (18b) have the form

$$i\gamma^\mu D_\mu N - mN = 0, \tag{20a}$$

$$\eta^{\mu\nu}[\partial_\mu\partial_\nu \varphi^a + \Gamma_b{}^a{}_c \partial_\mu \varphi^b \partial_\nu \varphi^c] = -i\overline{N}\gamma^\mu \Omega^a{}_b \partial_\mu \varphi^b N, \tag{20b}$$

where

$$\Omega^a{}_b = g^{ac}\Omega_{cb}, \tag{20c}$$

and

$$\Omega_{ab} = \partial_a \Omega_b - \partial_b \Omega_a - [\Omega_a, \Omega_b]. \tag{20d}$$

Using the expression for Ω_a given in Eq. (14d), the equation of motion for the pion field can be put in a manifestly chiral covariant geometric form

$$\eta^{\mu\nu}[\partial_\mu\, \partial_\nu\, \varphi^a + \Gamma_b{}^a{}_c \partial_\mu\, \varphi^b \partial_\nu\, \varphi^c] = \tfrac{1}{8}\overline{N}i\gamma^\mu R^a{}_{bcd} \times [L^c, L^d]\partial_\mu\, \varphi^b N, \tag{20e}$$

where $R^a{}_{bcd}$ is the Reimann curvature tensor in curved isospin space defined according to

$$R^a{}_{bcd} = \partial_c \Gamma_b{}^a{}_d - \partial_d\, \Gamma_b{}^a{}_c + \Gamma_c{}^a{}_e\, \Gamma_b{}^e{}_d - \Gamma_d{}^a{}_e\, \Gamma_b{}^e{}_c. \tag{20f}$$

The pion–nucleon Lagrangian (18b) contains chiral $SU_2\times SU_2$ invariant pion–pion as well as pion–nucleon interactions such that in the limit of vanishing curvature in curved isospin space ($f\to 0$) these interaction terms vanish and the Lagrangian (18b) reduces to a sum of free pion and nucleon Lagrangians that are invariant under the isospin transformations only. Thus the chiral $SU_2\times SU_2$ invariant pion–pion and pion–nucleon interactions contained in the Lagrangian (18b) have a geometric origin in that their form is completely determined by the curvature in isospin space. However, the chiral $SU_2\times SU_2$ symmetry does not restrict the pion–nucleon interaction to those terms that arise from the curvature of the isospin space only, and it is possible to construct chiral $SU_2\times SU_2$ invariant pion–nucleon interaction that does not vanish in the flat isospin space limit. Thus, since $\overline{N}L^a N$ and $\partial_\mu\, \varphi^a$ transform as contravariant vectors in curved isospin space, an interaction term of the form

$$\mathscr{L}_{\text{int}} = G\overline{N}\gamma^\mu\gamma^5 L^a(\varphi)\partial_\mu\, \varphi^b N g_{ab}(\varphi), \tag{21a}$$

where G is a coupling parameter with the dimension of length, will be invariant under the chiral $SU_2\times SU_2$ transformations. Using Eqs. (9e) and (10d) this can be written as

$$\mathscr{L}_{\text{int}} = 2fG\overline{N}\gamma^\mu\, \gamma^5\, \boldsymbol{\tau}\cdot\mathbf{X}_{5\mu}\, N, \tag{21b}$$

and in the limit of vanishing curvature in curved isospin space, it reduces to

$$\mathscr{L}_{\text{int}} \to G\overline{N}\gamma^\mu\gamma_5\, \boldsymbol{\tau}\cdot\partial_\mu\, \boldsymbol{\varphi}N \tag{21c}$$

which is invariant under isospin transformations only.

The complete chiral $SU_2\times SU_2$ invariant pion–nucleon Lagrangian can now be written as

$$\mathscr{L}_{\pi N} = \overline{N}i\gamma^\mu D_\mu\, N - m\overline{N}N + \tfrac{1}{2} g_{ab}(\varphi)\partial_\mu\, \varphi^a \partial^\mu\, \varphi^b + G\overline{N}\gamma^\mu\gamma^5 L^a(\varphi)\partial_\mu\, \varphi^b N g_{ab}(\varphi), \tag{22}$$

and its invariance under the chiral $SU_2\times SU_2$ transformations is manifest in its geometric structure.

IV. CHIRAL $SU_2\times SU_2$ π–N–ρ LAGRANGIAN

In this section we extend the general chiral $SU_2\times SU_2$ invariant pion–nucleon Lagrangian constructed in the last section by incorporating a massive isovector ρ-meson field without breaking the chiral $SU_2\times SU_2$ symmetry of the Lagrangian.

The chiral transformation law for the nucleon field given in Eqs. (12f) and (13h) is similar in structure to a local isospin transformation of an isospinor in Yang–Mills theory. Further, the chiral covariant derivative of the nucleon field given in Eqs. (16a) and (17a) is also similar in structure to the covariant derivative of an isospinor in Yang–Mills theory. Hence, if we replace the chiral covariant derivative of the nucleon field in Eqs. (16a) and (17a) by a Yang–Mills covariant derivative containing an isovector ρ-meson field according to

$$(\partial_\mu - \Omega_a \partial_\mu \varphi^a)N = (\partial_\mu + if^2 \boldsymbol{\tau}\cdot\mathbf{X}_\mu)N \rightarrow (\partial_\mu + (i/2)\, g\boldsymbol{\tau}\cdot \boldsymbol{\rho}_\mu)N, \tag{23a}$$

then the Lagrangian (18b) will still be chiral $SU_2\times SU_2$ invariant provided that the ρ-meson field matrix $\rho_\mu = \boldsymbol{\tau}\cdot \boldsymbol{\rho}_\mu$ has a chiral $SU_2\times SU_2$ transformation law identical to that of

$$(2i/g)\Omega_a \partial_\mu \varphi^a = (2f^2/g)\boldsymbol{\tau}\cdot\mathbf{X}_\mu .$$

In order to determine the chiral transformation law for $\Omega_a\, \partial_\mu\, \varphi^a$ we recall that Ω_a is defined by Eq. (14c) as

$$\nabla_b L^a(\varphi) = [\Omega_b(\varphi), L^a(\varphi)], \tag{23b}$$

or equivalently, by Eq. (14d) as

$$\Omega_a(\varphi) = -\tfrac{1}{8}[L_b(\varphi), \nabla_a L^b(\varphi)]. \tag{23c}$$

Moreover, the matrices $\Gamma^a(\varphi)$ are assumed to have the same functional form in all coordinate systems in curved isospin space that are related by chiral $SU_2\times SU_2$ transformations. Consequently, under a chiral $SU_2\times SU_2$ transformation $\Omega_a(\varphi)$ transforms as

$$\Omega_a(\varphi) \rightarrow \Omega_a(\varphi') = -\tfrac{1}{8}[L_b(\varphi'), \nabla_a L^b(\varphi')], \tag{23d}$$

and hence the chiral covariant derivative of the nucleon field given in Eq. (23a) transforms as

$$(\partial_\mu - \Omega_a(\varphi)\partial_\mu \varphi^a)N \rightarrow (\partial_\mu - \Omega_a(\varphi')\partial_\mu \varphi^{a'})N' . \tag{23e}$$

From Eq. (23c), we find that

$$\Omega_a(\varphi)\partial_\mu \varphi^a = -\tfrac{1}{8}[L_b(\varphi), \nabla_a L^b(\varphi)]\, \partial_\mu \varphi^a. \equiv -\tfrac{1}{8}[L_b(\varphi), D_\mu L^b(\varphi)], \tag{23f}$$

and under chiral $SU_2\times SU_2$ transformation this transforms as

$$\Omega_a(\varphi)\partial_\mu \varphi^a \rightarrow \Omega_a(\varphi')\partial_\mu \varphi^{a'} = -\tfrac{1}{8}[L_b(\varphi'), D_\mu L^b(\varphi')]. \tag{23g}$$

Finally, from Eqs. (10c), (12c), and (23g), we find that

$$\Omega_a(\varphi')\partial_\mu \varphi^{a'} = S(\Omega_a(\varphi)\partial_\mu \varphi^a)S^{-1} - S\partial_\mu S^{-1}, \tag{23h}$$

so that according to Eq. (23a) the ρ-meson field matrix has a chiral $SU_2\times SU_2$ transformation law of the form

$$\rho_\mu \rightarrow \rho_\mu{}' = S\rho_\mu S^{-1} - (2i/g)S\partial_\mu S^{-1}. \tag{24a}$$

This chiral transformation for the ρ-meson field is similar in structure to the transformation of an isovector gauge field in Yang–Mills theory, and consequently, a covariant curl of the ρ-meson field defined according to

$$\boldsymbol{\tau}\cdot\mathbf{F}_{\mu\nu} = F_{\mu\nu} = \partial_\mu \rho_\nu - \partial_\nu \rho_\mu + (ig/2)[\rho_\mu, \rho_\nu], \tag{24b}$$

will have a chiral transformation law of the form

$$F'_{\mu\nu} = SF_{\mu\nu} S^{-1}, \tag{24c}$$

and hence a kinetic energy term for the ρ-meson of the form $-\frac{1}{4}\mathbf{F}_{\mu\nu}\cdot\mathbf{F}^{\mu\nu}$, where $\mathbf{F}_{\mu\nu} = \partial_\mu\, \boldsymbol{\rho}_\nu - \partial_\nu\, \boldsymbol{\rho}_\mu - g\, \boldsymbol{\rho}_\mu \times \boldsymbol{\rho}_\nu$ will be chiral invariant.

In contrast to the usual Yang–Mills theory, where the gauge field must remain massless in the limit of exact symmetry, we have in the present chiral Yang–Mills theory the possibility of adding a mass term for the ρ-meson field without breaking the chiral $SU_2\times SU_2$ symmetry. This is because under the chiral transformations ζ_μ is assumed to transform in the same way as $(2f^2/g)\, \mathbf{X}_\mu$, and hence the combination

$$\left(\boldsymbol{\rho}_\mu - \frac{2f^2}{g}\mathbf{X}_\mu\right)^2$$

will be invariant under the chiral $SU_2\times SU_2$ transformations and will provide a mass term for the ρ-meson without breaking the chiral $SU_2\times SU_2$ symmetry.

Thus with the incorporation of the ρ-meson field, the general nonlinear chiral $SU_2\times SU_2$ invariant π-N-ρ Lagrangian takes the form

$$\begin{aligned}\mathscr{L}_{\pi N\rho} = {}& \overline{N}i\gamma^\mu(\partial_\mu + (i/2)\, g\boldsymbol{\tau}\cdot \boldsymbol{\rho}_\mu)N - m\overline{N}N \\ & - \tfrac{1}{4}\mathbf{F}_{\mu\nu}\cdot\mathbf{F}^{\mu\nu} + \tfrac{1}{2}m_{\rho^2}(\boldsymbol{\rho}_\mu - (2f^2/g)\mathbf{X}_\mu)^2 \\ & + \tfrac{1}{2}g_{ab}(\varphi)\partial_\mu \varphi^a\partial^\mu\varphi^b \\ & + G\overline{N}\gamma^\mu\gamma^5 L^a(\varphi)\partial_\mu \varphi^b N g_{ab}(\varphi).\end{aligned} \tag{25}$$

This Lagrangian coincides with the most general nonlinear chiral $SU_2\times SU_2$ invariant π–N–ρ Lagrangian constructed previously using the matrix method.[2] Here we have constructed this Lagrangian by generalizing the geometric method of Meetz,[1] and the chiral $SU_2\times SU_2$ invariance of the Lagrangian (25) is manifest in its geometric structure.

[1]K. Meetz, J. Math. Phys. **10**, 589 (1969); see also L. Brown and D. Boulware, Ann. Phys. **138**, 392 (1982).
[2]A. Ebrahim and F. Gürsey, Nuovo Cimento Lett. **9** (9), 716E (1974).
[3]A. Ebrahim, J. Math. Phys. **16**, 1253 (1975).
[4]A. Ebrahim and M. Serdaroğlu, Nuovo Cimento A **21**, 249 (1974).
[5]See, e.g., S. Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1972).
[6]See, e.g., C. Misner, Phys. Rev. D **18**, 4510 (1978).
[7]W. Pauli, Ann. Phys. **18**, 337 (1933); D. Brill and J. Wheeler, Rev. Mod. Phys. **29**, 465 (1957); J. Anderson, *Principles of Relativity Physics* (Academic, New York, 1967), p. 358.

Explicit evaluation of a path integral with memory kernel

F. Brosens[a] and J. T. Devreese[b]
Physics Department, Universitaire Instelling Antwerpen, Universiteitsplein 1, B-2610 Antwerpen, Belgium

(Received 8 August 1983; accepted for publication 16 December 1983)

An explicit expression is presented for the path integral of an harmonic oscillator interacting with itself in the past. The time dependence of the memory kernel is assumed to be exponential. The relation with the Feynman trial action for the variational calculation of the polaron ground state energy is discussed.

PACS numbers: 02.50. + s, 71.38. + i

1. INTRODUCTION

In the present paper, the explicit evaluation is presented of the path integral

$$K_f(x,t\,|x',0) = \int \mathscr{D}\, x_\tau \exp \frac{i}{\hbar} S_f[\dot{x},x,t\,] \tag{1a}$$

with the boundary conditions $x_t = x$ and $x_0 = x'$ over the action

$$S_f[\dot{x},x,t\,] = \int_0^t d\tau\Big(\frac{m}{2}\dot{x}_\tau^2 - \frac{1}{2}m\Omega^2 x_\tau^2 + f(\tau)x_\tau\Big) + i\hbar 2C\int_0^t d\tau\int_0^\tau d\sigma\, x_\tau x_\sigma e^{-iw|\tau-\sigma|}. \tag{1b}$$

For imaginary times $t = -iT$, this propagator becomes

$$\mathscr{K}_f(x,T\,|x',0) = \int \mathscr{D} x_\tau \exp \frac{1}{\hbar}\mathscr{S}_f[\dot{x},x,T\,] \tag{2a}$$

with

$$\mathscr{S}_f[\dot{x},x,T\,] = -\int_0^T d\tau\Big(\frac{m}{2}\dot{x}_\tau^2 + \frac{1}{2}m\Omega^2 x_\tau^2 - f(\tau)x_\tau\Big) + \hbar C\int_0^T d\tau\int_0^T d\sigma\, x_\tau x_\sigma e^{-w|\tau-\sigma|}. \tag{2b}$$

Although this action at first glance resembles the Feynman trial action[1] for the study of the Fröhlich polaron, an essential difference is the fact that in (1b) and (2b) the coefficient of x_τ^2 is not time dependent. The memory effect only appears in the coupling with the trajectory in the past. Furthermore, the strength of the harmonic potential is independent of the strength of the memory terms, whereas in the Feynman trial action both are intimately connected.

Despite these differences (the first of which is essential for the actual evaluation), the calculation of the ground state energy of the large polaron with this trial action (1b) reproduces the Feynman ground state energy for the polaron. Indeed, the equation of motion for the classical path, as derived from the extremalization of the action (1), becomes exactly the asymptotic limit (i.e., $0 \ll \tau \ll T$ for $T \to \infty$) of the equation of motion corresponding to the Feynman action, provided one imposes the extra condition $\Omega^2 = 4\hbar C/mw$. This condition is not required in the action (2b), suggesting that it allows in principle to find a lower variational energy. However, as will be shown below, this extra degree of freedom in the variational determination of the energy, turns out to introduce an harmonic potential for the center of mass, which does not improve the upper bound for the energy.

Although the integration of (1) and (2) thus does not provide new results for the ground state energy of the polaron, and although the path integral itself is not even required within Feynman's treatment, the explicit evaluation of this type of path integrals remains a challenging problem in itself, and seems of particular interest from a mathematical point of view. Only a limited class of Gaussian path integrals with retardation has effectively been calculated.

The most general class of path integrals with a quadratic action and with a memory kernel treated so far has been worked out in Ref. 2, including several previously known functional integrals as limiting cases, and with explicit results for "β-periodic kernels." However, the kernel in the action (1) does not satisfy this periodicity condition, and to the best of our knowledge, the path integral (1) has not been obtained explicitly so far.

It should be noted that the exponential dependence of the memory term in the action (1), introduces a substantial difference with the Gaussian path integral for an electron gas in a random potential, as studied in Ref. 3.

2. EVALUATION OF THE PATH INTEGRAL

In order to evaluate the path integral (1), we proceed as follows. In the first place, we examine how this path integral derives from a Hamiltonian, describing an oscillator in an external field, and coupled to a boson. Given the averaging procedure for obtaining the propagator (1) from the Hamiltonian, the path integral can be explicitly evaluated if one subsequently diagonalizes the Hamiltonian, writes its propagator, and applies the averaging procedure. The relevant Hamiltonian turns out to be

$$H = \frac{p^2}{2m} + \frac{m}{2}\Omega^2 x^2 - f(t)x - i\hbar\sqrt{2C}\,(a - a^+)x + \hbar w a^+ a, \tag{3}$$

where a and a^+ are the standard annihilation and creation operators for bosons. Consider then the propagator

$$K(x,\alpha,t\,|x',\alpha',0) = \langle x,\alpha|\exp -\frac{i}{\hbar}\int_0^t d\tau\, H\,|\alpha',x'\rangle, \tag{4}$$

where the coherent states $|\alpha\rangle$ (with complex numbers α) are

[a] Research Associate of the National Fund for Scientific Research (Belgium).
[b] Institute of Applied Mathematics, Rijksuniversitair Centrum Antwerpen, Groenenborgerlaan 171, B-2020 Antwerpen, and Eindhoven University of Technology, Eindhoven, The Netherlands.

defined as[4–6]

$$|\alpha\rangle = \frac{1}{\pi^{1/2}} \sum_{n=0}^{\infty} \frac{\alpha^n e^{-1/2|\alpha|^2}}{(n!)^{1/2}} \Big| n \Big\rangle. \tag{5}$$

The propagator (4) can be obtained from the well-known properties of these coherent states (see, e.g., Ref. 6). The result is

$$K(x,\alpha,t\,|x',\alpha',0) = \int \mathscr{D}x_\tau \exp \frac{i}{\hbar}\int_0^t d\tau\Big(\frac{m}{2}\dot{x}_\tau^2 - \frac{m}{2}\Omega^2 x_\tau^2 + f(\tau)x_\tau\Big) K_{[x]}(\alpha,t\,|\alpha',0), \tag{6a}$$

where

$$K_{[x]}(\alpha,t\,|\alpha',0) = (1/\pi)\exp(-(|\alpha|^2/2 + |\alpha'|^2/2 - \alpha^*\alpha' e^{-iwt})) \times \exp(\alpha^* A_{[x]}(t) - \alpha' A^*_{[x]}(t) e^{-iwt} - (i/\hbar)\Phi_{[x]}(t)) \tag{6b}$$

with

$$A_{[x]}(t) = \sqrt{2C}\, e^{-iwt}\int_0^t d\tau\, x_\tau e^{iw\tau}, \tag{6c}$$

$$\Phi_{[x]}(t) = -i\hbar 2C \int_0^t d\tau \int_0^\tau d\sigma\, x_\tau x_\sigma e^{-iw(\tau-\sigma)}. \tag{6d}$$

This propagator can be transformed into the boson number representation

$$K(x,n,t\,|x',n',0) = \int d^2\alpha \int d^2\alpha' \langle n|\alpha\rangle \times K(x,\alpha,t\,|x',\alpha',0)\langle\alpha'|n'\rangle. \tag{7}$$

Performing the integrations explicitly, one finds that this propagator for $n = n' = 0$ is exactly the propagator (1) which we study.

$$K(x,n=0,t\,|x',n'=0,0) = K_f(x,t\,|x',0). \tag{8}$$

The path integral (1) is thus obtained by eliminating the boson in the propagator of the Hamiltonian (3), where the elimination means that the average of the propagator in the boson ground state is considered.

Once this relation is established, one can proceed to the evaluation of the path integral. A convenient way to do this, is first to diagonalize the Hamiltonian, which of course gives two independent oscillators, and then to eliminate the boson contribution from the resulting propagator. Introducing then the momentum and position operators p_2 and x_2,

$$\alpha = \frac{p_2}{\sqrt{2\hbar m_2 w}} - i\sqrt{\frac{m_2 w}{2\hbar}}\,x_2, \tag{9a}$$

$$\alpha^+ = \frac{p_2}{\sqrt{2\hbar m_2 w}} + i\sqrt{\frac{m_2 w}{2\hbar}}\,x_2, \tag{9b}$$

one readily obtains the Lagrangian, associated with the Hamiltonian (3)

$$L(\dot{x},x;\dot{x}_2,x_2;t) = \tfrac{1}{2}\hbar w + \tfrac{1}{2}m\dot{x}^2 - \tfrac{1}{2}m\Omega^2x^2 + f(t)x + \tfrac{1}{2}m_2\dot{x}_2^2 - \tfrac{1}{2}m_2w^2x_2^2 + 2xx_2\sqrt{\hbar C m_2 w}, \tag{10}$$

which is the Lagrangian of two coupled harmonic oscillators, one of them acted upon by an external force $f(t)$. Obviously, this system can be written as a sum of two independent oscillators, via the transformation

$$q_1 = x_2\sqrt{m_2}\sin\theta + x\sqrt{m}\cos\theta;$$
$$q_2 = x_2\sqrt{m_2}\cos\theta - x\sqrt{m}\sin\theta. \tag{11}$$

Introducing the notation

$$\gamma = 2\sqrt{\frac{\hbar w C}{m}};\quad W = \sqrt{(\Omega^2 - w^2)^2 + 4\gamma^2}, \tag{12}$$

the angle θ of rotation of the coordinate frame is defined by

$$\sin\theta = -\sqrt{\frac{w^2 - \Omega^2 + W}{2W}};\quad \cos\theta = \sqrt{\frac{\Omega^2 - w^2 + W}{2W}}. \tag{13}$$

The eigenfrequencies of the independent oscillators are

$$\Omega_1^2 = \frac{w^2 + \Omega^2 + W}{2};\quad \Omega_2^2 = \frac{w^2 + \Omega^2 - W}{2}, \tag{14}$$

and the forces which act upon them are

$$g_1(t) = \frac{\cos\theta}{m^{1/2}}f(t);\quad g_2(t) = -\frac{\sin\theta}{m^{1/2}}f(t). \tag{15}$$

The Lagrangian, obtained after this transformation, is

$$L = \frac{1}{2}\hbar w + \sum_{j=1}^{2}\Big(\frac{\dot{q}_j^2}{2} - \frac{1}{2}\Omega_j^2 q_j^2 + g_j(t)q_j\Big). \tag{16}$$

The propagator of this system is (see, e.g., Ref. 7, or Ref. 5; p. 38)

$$K(x,x_2,t\,|x',x_2',0) = \sqrt{mm_2}\,e^{iwt/2}k_{g_1}(q_1,t\,|q_1',0) \times k_{g_2}(q_2,t\,|q_2',0), \tag{17}$$

where the mass factor in front is due to the Jacobian from the transformation (11), and where the propagators $k_{g_j}(q,t\,|q',0)$ are given by

$$k_{g_j}(q_j,t\,|q_j',0) = \sqrt{\frac{\Omega_j}{2\pi i\hbar\sin\Omega_j t}}\exp\frac{i}{\hbar}\frac{\Omega_j}{2\sin\Omega_j t}\Big[(q_j^2 + q_j'^2)\cos\Omega_j t - 2q_jq_j' + \frac{2}{\Omega_j}\int_0^t d\tau\, g_j(\tau)(q_j\sin\Omega_j\tau + q_j'\sin\Omega_j(t-\tau)) - \frac{2}{\Omega_j^2}\int_0^t d\tau\int_0^\tau d\sigma\, g_j(\tau)g_j(\sigma)\sin\Omega_j\sigma\sin\Omega_j(t-\tau)\Big]. \tag{18}$$

As mentioned above, the propagator $K_f(x,t\,|x',0)$ is the propagator of the Hamiltonian, averaged over the phonon ground state, and thus it can be obtained by performing the integration

$$K_f(x,t\,|x',0)=\int_{-\infty}^{\infty}dx_2\int_{-\infty}^{\infty}dx_2'\;\Psi_0(x_2)K(x,x_2,t\,|x',x_2',0)\Psi_0(x_2') \tag{19}$$

with

$$\Psi_0(x_2)=\left(\frac{m_2w}{\pi\hbar}\right)^{1/4}\exp-\frac{m_2w}{2\hbar}x_2^2. \tag{20}$$

Using the transformation (11), one realizes that only Gaussian integrals are involved, which can be done. Introducing the shorthand notations

$$X=\sqrt{\frac{mw}{2\hbar}}x;\quad X'=\sqrt{\frac{mw}{2\hbar}}x', \tag{21}$$

$$b_1(t)=\frac{i\Omega_1}{w\sin\Omega_1 t};\quad b_2(t)=\frac{i\Omega_2}{w\sin\Omega_2 t}, \tag{22}$$

$$A(t)=1-b_1(t)\sin^2\theta\cos\Omega_1 t-b_2(t)\cos^2\theta\cos\Omega_2 t, \tag{23}$$

$$B(t)=b_1(t)\sin^2\theta+b_2(t)\cos^2\theta, \tag{24}$$

$$D(t)=2X\tan\theta-2b_1(t)(X\cos\Omega_1 t-X')\tan\theta+\sqrt{\frac{2w}{m\hbar}}\sin\theta\cos\theta\int_0^t d\tau f(\tau)\left(b_2(t)\frac{\sin\Omega_2\tau}{\Omega_2}-b_1(t)\frac{\sin\Omega_1\tau}{\Omega_1}\right), \tag{25}$$

$$D'(t)=2X'\tan\theta-2b_1(t)(X'\cos\Omega_1 t-X)\tan\theta+\sqrt{\frac{2w}{m\hbar}}\sin\theta\cos\theta\int_0^t d\tau f(\tau)\left(b_2(t)\frac{\sin\Omega_2(t-\tau)}{\Omega_2}-b_1(t)\frac{\sin\Omega_1(t-\tau)}{\Omega_1}\right), \tag{26}$$

one finally obtains

$$\begin{aligned}K_f(x,t\,|x',0)={}&2\sqrt{\frac{m\hbar}{w}}\frac{1}{\sqrt{A^2(t)-B^2(t)}}\sqrt{\frac{\Omega_1}{2\pi i\hbar\sin\Omega_1 t}}\sqrt{\frac{\Omega_2}{2\pi i\hbar\sin\Omega_2 t}}e^{iwt/2}\exp-(X^2+X'^2)\tan^2\theta\\&\times\exp\frac{b_1(t)}{\cos^2\theta}((X^2+X'^2)\cos\Omega_1 t-2XX')\exp\frac{2b_1(t)}{\Omega_1}\sqrt{\frac{w}{2m\hbar}}\int_0^t d\tau f(\tau)(X\sin\Omega_1\tau+X'\sin\Omega_1(t-\tau))\\&\times\exp-\frac{wb_1(t)\cos^2\theta}{m\hbar\Omega_1^2}\int_0^t d\tau\int_0^\tau d\sigma f(\tau)f(\sigma)\sin\Omega_1\sigma\sin\Omega_1(t-\tau)\exp-\frac{wb_2(t)\sin^2\theta}{m\hbar\Omega_2^2}\\&\times\int_0^t d\tau\int_0^\tau d\sigma f(\tau)f(\sigma)\sin\Omega_2\sigma\sin\Omega_2(t-\tau)\exp\frac{A(t)[D^2(t)+D'^2(t)]-2B(t)D(t)D'(t)}{4[A^2(t)-B^2(t)]}.\end{aligned} \tag{27}$$

Note that $D(t)$ and $D'(t)$ depend on the coordinates X and X'. The propagator (2) for imaginary times $t=-iT$ is readily obtained from (27).

3. REMARK ON THE GROUND STATE ENERGY OF THE POLARON

For the study of the ground state energy of the polaron, along the lines introduced by Feynman, the propagator for imaginary times is only required for $T\rightarrow\infty$. Therefore, we restrict ourselves to this asymptotic limit. In this case, $b_1(-iT)$ and $b_2(-iT)$ and thus also $B(-iT)$ decay exponentially with T. Combining the terms in X and X', and neglecting all contributions in the exponent which decrease exponentially with T, one ends up with

$$\mathscr{K}_f(x,T\rightarrow\infty\,|x',0)\approx\mathscr{K}_{f=0}(x,T\rightarrow\infty\,|x',0)\exp\frac{1}{m\hbar}\int_0^T\int_0^T d\tau\,d\sigma f(\tau)f(\sigma)\left[\frac{\cos^2\theta}{4\Omega_1}e^{-\Omega_1|\tau-\sigma|}+\frac{\sin^2\theta}{4\Omega_2}e^{-\Omega_2|\tau-\sigma|}\right], \tag{28}$$

where

$$\begin{aligned}\mathscr{K}_{f=0}(x,T\rightarrow\infty\,|x',0)\approx{}&2\frac{\sqrt{m\hbar w}}{w+\Omega_1\sin^2\theta+\Omega_2\cos^2\theta}\sqrt{\frac{\Omega_1}{\pi\hbar}}\sqrt{\frac{\Omega_2}{\pi\hbar}}e^{-(\Omega_1+\Omega_2-w)T/2}\\&\times\exp-(X^2+X'^2)\frac{\Omega_1\Omega_2/w+\Omega_1\cos^2\theta+\Omega_2\sin^2\theta}{w+\Omega_1\sin^2\theta+\Omega_2\cos^2\theta}.\end{aligned} \tag{29}$$

Following Feynman's variational treatment, but using the action (2b) without external force as the trial action in the Jensen inequality, one obtains an upper bound for the polaron ground state energy E_0. Of course, the ground state energy corresponding to the trial action equals $\Omega_1+\Omega_2-w$, as follows from the exponential decay with T in (29). Also for the expectation value of the interaction part in the polaron action, one can directly use the propagator (28), without relying on the equation of motion for the classical path. Without going into details, we quote the final result (expressed in units such that the electron mass m, the frequency of the longitudinal optical phonons and $\hbar$ equal 1)

$$E_0 \leqslant \frac{3}{4}\frac{(\Omega_1+\Omega_2-w)^2+\Omega_1\Omega_2}{\Omega_1+\Omega_2} - \frac{\alpha}{\sqrt{\pi}}\int_0^\infty d\sigma \times \frac{e^{-\sigma}}{\sqrt{[(\Omega_1^2-w^2)/(\Omega_1^2-\Omega_2^2)](1-e^{-\Omega_1\sigma})/\Omega_1 - [(\Omega_2^2-w^2)/(\Omega_1^2-\Omega_2^2)](1-e^{-\Omega_2\sigma})/\Omega_2}}, \tag{30}$$

where Ω_1, Ω_2, and w are variation parameters, and α is the polaron coupling constant.

Since in the Hamiltonian formulation of the polaron problem the total momentum associated with the electron and the phonons is a conserved quantity, one would intuitively expect that the trial action reflects this property. One thus expects that the minimum of the energy (30) is found for values of the parameters where the diagonalized Hamiltonian (3) consists of a free particle (i.e., $\Omega_2 = 0$) and an harmonic oscillator. This is indeed confirmed numerically. The minimum variational energy is obtained for $\Omega_2 = 0$, which means from (14) that $\Omega^2 = 4\hbar C/mw$. As mentioned in the introduction, this is precisely the condition for which the equation of motion for the classical path from the action (1), in the asymptotic limit for imaginary times becomes exactly the equation of motion corresponding to the Feynman trial action. Under this condition, one also readily checks that the minimum of the energy (30) becomes exactly the Feynman upper bound for the ground state energy of the polaron.

ACKNOWLEDGMENT

Financial support by the F.K.F.O. ("Fonds voor Kollektief Fundamenteel Onderzoek"), Belgium, Project No. 2.0072.80, is gratefully acknowledged.

[1] R. P. Feynman, Phys. Rev. **97**, 660 (1955).

[2] J. Adamowski, B. Gerlach, and H. Leschke, J. Math. Phys. **23**, 243 (1982).

[3] V. Bezak, Proc. R. Soc. London Ser. A **315**, 339 (1970); G. Papadopoulos, J. Phys. A **7**, 183 (1974); D. C. Khandekar, S. V. Lawande, and K. V. Bhagwat, Phys. Lett. A **93**, 167 (1983); K. L. Sebastian, Phys. Lett. A **95**, 131 (1983).

[4] J. R. Klauder, Ann. Phys. **11**, 123 (1960).

[5] L. S. Schulman, *Techniques and Applications of Path Integration* (Wiley, New York, 1981), p. 242.

[6] See, e.g., G. J. Papadopoulos, in *Path Integrals and their Applications in Quantum, Statistical and Solid State Physics*, edited by G. J. Papadopoulos and J. T. Devreese (Plenum, New York, 1978), p. 139.

[7] R. P. Feynman and A. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965), p. 64.

The occupation statistics for indistinguishable dumbbells on a rectangular lattice space. I

Alain J. Phares
Department of Physics, Villanova University, Villanova, Pennsylvania 19085

(Received 12 October 1983; accepted for publication 10 February 1984)

The method of Hock and McQuistan used recently to solve the occupation statistics for indistinguishable dumbbells (or dimers) on a $2\times 2\times N$ lattice is extended further to obtain, for the $L\times M\times N$ lattice, general expressions for the normalization, expectation, and dispersion of the statistics, and their limit as N becomes very large. In particular, an explicit expression of the partition function in the thermodynamic limit $\Xi(x)$ is obtained for any value of the absolute activity x of dimers. The developed mathematical formalism is then applied to planar lattices, $1\times M\times N$, with $M = 1, 2, 3$, and 4. The known results for $M = 1$ and 2 are recovered, and some new ones are obtained. The recurrence relation for the number $A(q,N)$ of arrangements of q dumbbells on a $1\times M\times N$ lattice which has 3 and 5 terms when $M = 1$ and 2, respectively, is found to have 15 and 65 terms for $M = 3$ and 4. Analysis and extrapolation of the results enable one to predict the expectation $\langle\theta\rangle_{1MN}$ on a planar $1\times M\times N$ lattice to be 63.4%, in the limit as both M and N become infinite. We also find an upper bound on the quantity $MN[\langle\theta^2\rangle - (\langle\theta\rangle)^2]$ in the limit as both M and N become infinite. In the thermodynamic limit (M & $N\to\infty$) the partition function $\Xi(1)$, for the absolute activity $x = 1$, is found to be equal to 1.95. By limiting the number M of rows of infinite extent ($N\to\infty$) to just 4, we find that the error in determining $\Xi(1)$ for the infinite two-dimensional lattice is just 4.5%. In this paper $\Xi(x)$ is obtained for any value of the absolute activity x for $M = 1$ and 2. A more thorough study of $\Xi(x)$, and its fast convergence with increasing values of M, and applications will be presented in a forthcoming article.

PACS numbers: 02.50. + s, 05.50. + q, 05.70.Ce, 68.90. + g

I. INTRODUCTION

The statistical treatment of the absorption and crystallization of diatomic molecules, or the treatment of the properties of binary alloys, usually leads to the problem of determining the occupational degeneracy for indistinguishable dumbbells (or dimers) distributed on a lattice. In such problems, if a site is occupied, then, at least one of its nearest neighbor sites must also be occupied. Consequently, there is a distribution of pairs (dumbbells or dimers) of occupied nearest neighbor sites. Exact solutions have been found by McQuistan and co-workers (referred to hereafter as MQC) for the following special cases $1\times 1\times N$, $1\times 2\times N$, and $2\times 2\times N$.[1] Exact solutions have also been obtained for a *completely* filled lattice $1\times M\times N$ using Pfaffians[2] and the transfer matrix method.[3]

Basic improvements on MQC's method[1] enable one to give a general formulation of the problem of dumbbells distributed on a $L\times M\times N$ lattice, where L and M are fixed, and N is allowed to become very large. We give a brief summary of MQC's method.

(1) MQC introduced sublattice spaces obtained from the original one by removing from the Nth array containing LM compartments one, two, or more cells. Since they only considered the $2\times 2\times N$ lattice space, the number of sublattices thus generated is rather limited.

(2) MQC then considered the number of ways q indistinguishable dumbbells could be arranged on the original lattice and also on every other sublattice they generated. By graphical analysis they were able to develop coupled recurrence relations among the various possible arrangements on different lattices.

(3) The major difficulty of extending MQC's method to larger lattices is that one must decouple the recurrence relations to obtain a recurrence relation involving $A(q,N)$, the number of ways of arranging q indistinguishable dumbbells on the nontruncated lattice, in the form

$$A(q,N) = \sum_{n=0}^{n_{\max}} \sum_{m=1}^{m_{\max}} c_{nm} A(q-n, N-m). \tag{1.1}$$

For a given set of values of L and M, the coefficients c_{nm}, and the quantities $n_{\max}$ and $m_{\max}$ do not depend on q or N. Equation (1.1) generally holds for $q\geqslant n_{\max}$ and $N\geqslant m_{\max}$. Consequently, the initial values of $A(q,N)$ for $0\leqslant q<q_{\max}$ and $0\leqslant N<m_{\max}$ have to be calculated numerically.

(4) MQC also introduced the bivariant generating function

$$G(x,y) = \sum_{N=0}^{\infty} \sum_{q=0}^{q_{\max}} A(q,N) x^q y^N. \tag{1.2}$$

Combining Eq. (1.2) with the decoupled recurrence relation Eq. (1.1), $G(x,y)$ is obtained explicitly as the ratio of the two polynomials in x and y,

$$G(x,y) = H(x,y)/D(x,y). \tag{1.3}$$

Essential to this procedure is the evaluation of the initial values of $A(q,N)$ for $0\leqslant q<q_{\max}$ and $0\leqslant N<m_{\max}$.

(5) The meaningful statistical quantities are the configurational grand-canonical partition function[4]

$$\Delta_N(x) = \sum_{q=0}^{q_{\max}} A(q,N) x^q, \tag{1.4}$$

the expectation value of the occupation of $L\times M\times N$ lattice

 0022-2488/84/061756-15$02.50

$$\langle \theta \rangle_N = \frac{1}{q_{\max} \Delta_N} \sum_{q=0}^{q_{\max}} qA(q,N), \quad (1.5)$$

the dispersion in θ

$$\langle \theta^2 \rangle_N = \frac{1}{q_{\max}^2 \Delta_N} \sum_{q=0}^{q_{\max}} q^2 A(q,N), \quad (1.6)$$

and the mean square deviation

$$\sigma_N^2 = [\langle \theta^2 \rangle_N - (\langle \theta \rangle_N)^2]. \quad (1.7)$$

The connection between the bivariant generating function and the above statistical quantities is made by recalling that

$$G(x,y) = \sum_{N=0}^{\infty} y^N \sum_{q=0}^{q_{\max}} x^q A(N,q) = \sum_{N=0}^{\infty} y^N \Delta_N(x). \quad (1.8)$$

The quantity x represents the absolute activity of the dumbbells or dimers, and MQC have only considered the case $x = 1$.

(6) MQC then calculate the z roots R_j of $D(1,1/z)$, and perform an expansion of $G(1,y)$ in the form

$$G(1,y) = \sum_j \frac{k_j}{(1 - yR_j)} = \sum_j k_j \sum_{N=0}^{\infty} (yR_j)^N$$
$$= \sum_{N=0}^{\infty} y^N \sum_j k_j R_j^N. \quad (1.9)$$

Comparing Eqs. (1.8) and (1.9), the normalization is obtained exactly as a finite sum over the roots of $D(1,1/z)$, i.e.,

$$\Delta_N(1) = \sum_j k_j R_j^N(1). \quad (1.10)$$

As N becomes large, the leading term in Eq. (1.10) gives a good approximation of the normalization Δ_N, namely, if R_1 is the largest root, then

$$\Delta_N \simeq k_1 R_1^N. \quad (1.11)$$

The method of calculating k_1 as suggested by MQC for the $2 \times 2 \times N$ lattice is rather complicated and cannot be easily extended to the general case. As a matter of fact, this is also true for the other quantities (1.5), (1.6), and (1.7) which they have calculated for the same lattice mentioned above.

This paper is the first in a series of two.[5] Here we intend only to present exact mathematical expressions for the grand-canonical partition function $\Delta_N(x)$, for any value of the activity x ($x \geqslant 0$). This will enable one to derive the closed form expression for the thermodynamic limit

$$\Xi(x) = \lim_{N \to \infty} [\Delta_N(x)]^{1/LMN}. \quad (1.12)$$

We will also obtain the large N behavior of $\langle \theta \rangle_{LMN}$ and σ_{LMN}^2 for the rectangular $L \times M \times N$ lattice. This is done in Sec. II. The general formalism is then applied to a planar lattice space, namely $1 \times M \times N$ lattices for $M = 1, 2, 3$, and 4. The closing section is devoted to an analysis and discussion of the numerical results obtained. Throughout this article, all the numerical results were obtained using a Hewlett Packard 34C pocket calculator.

II. GENERAL FORMALISM

Following step (1) of MQC's procedure, we first consider all the topologically distinct sublattices obtained from the original rectangular $L \times M \times N$ lattice by removing from the Nth array, containing LM compartments, one, two, etc., up to $(LM - 1)$ compartments. The capital letter "T" will be used to denote a whole class of topologically distinct sublattices obtained from the original lattice by removing a *given* number of compartments. In the applications presented in this paper we use the following notations:

"$\mathscr{A}$" represents the nontruncated original lattice;

"$\mathscr{B}$" represents the whole class of topologically distinct sublattices with one compartment removed from the Nth array;

"$\mathscr{C}$" represents the whole class of topologically distinct sublattices with two compartments removed from the Nth array; and so on. Thus, "$\mathscr{T}$" stands for "$\mathscr{A}$," "$\mathscr{B}$," "$\mathscr{C}$," etc.

In order to enumerate the various elements belonging to the same class $\mathscr{T}$, we choose a running lower index i; so that when referring to a particular sublattice belonging to the class $\mathscr{T}$, we would call it $\mathscr{T}_i$.

The number of ways q dumbbells can be arranged on a particular T_i-lattice is called $T_i(q,N)$, where it is understood that L and M are fixed and N is the only dimension allowed to vary. The initial values for these numbers are

$$T_i(0,N) = 1 \quad \text{for } N \neq 0, \quad (2.1a)$$
$$A(0,0) = 1 \quad (\text{i.e.,} \mathscr{T} = \mathscr{A}). \quad (2.1b)$$

To these initial values, we add the initial conditions

$$T_i(0,0) = 0, \quad \text{for } \mathscr{T} \neq \mathscr{A}, \quad (2.1c)$$

since for $N = 0$, no compartments exist on an $\mathscr{A}$-lattice and, therefore, other $\mathscr{T}$-lattices, obtained by removing one or more compartments for the $\mathscr{A}$-lattice, do not exist.

The next step is to generate the coupled recurrence relations among the various numbers of arrangements $T_i(q,N)$ in the form

$$T_i(q,N) = \sum_{\{T_j\}} \sum_n \sum_m c_{nm} [T_i, T_j] T_j(q - n, N - m). \quad (2.2)$$

There will be as many equations as there are T_i's. The number of equations involved will be studied in Paper II.[5] For every given set of values of L and M (specifying the $L \times M \times N$ lattice space) the coefficients $c_{nm}[T_i, T_j]$ have to be evaluated. These coefficients are positive integers. This is step (2) of MQC's method described in the introduction.

For the purpose of obtaining meaningful statistical quantities, the bivariant generating function expressions Eq. (1.3) is essential. It can be determined without explicitly having to perform the decoupling of the recurrence relations Eq. (2.2) and obtain a recurrence relation for A only, as in Eq. (1.1). This is done by introducing as many bivariant generating functions as there are $\mathscr{T}_i$- lattices, namely

$$G_{T_i}(x,y) = \sum_{N=0}^{\infty} \sum_{q=0}^{q_{\max}} T_i(q,N) x^q y^N. \quad (2.3)$$

Combining Eqs. (2.2), (2.3), and the initial value conditions, Eqs. (2.1a), (2.1b), and (2.1c), one finds

$$G_A(x,y) = 1 + \sum_{\{T_j\}} \sum_{n,m} c_{nm} [A, T_j] G_{T_j}(x,y) x^n y^m, \quad (2.4a)$$

$$G_{T_i}(x,y) = \sum_{\{T_j\}} \sum_{n,m} c_{nm} [T_i T_j] G_{T_j}(x,y) x^n y^m, \quad \text{for } T_i \neq A. \quad (2.4b)$$

Thus, the decoupling of recurrence relations is reduced to the much easier problem of solving a system of linear equations where the bivariant generating functions G_{T_i} play the role of unknowns.[6] Clearly, in this system of equations, the coefficients are $c_{nm}[T_i,T_j]x^n y^m$. It then follows that each one of the bivariant generating functions can be written as the ratio of two polynomials in x and y. Formally, the system of equations (2.4a) and (2.4b) may be represented by a matrix equation, namely

$$\begin{bmatrix} 1-\sum_{n,m} x^n y^m c_{nm}[A,A], & -\sum_{n,m} x^n y^m c_{nm}[A,B_1], & \cdots \\ -\sum_{n,m} x^n y^m c_{nm}[B_1,A], & 1-\sum_{n,m} x^n y^m c_{nm}[B_1,B_1], & \cdots \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} G_A \\ G_{B_1} \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \end{bmatrix}. \tag{2.5}$$

Solving for G_A, one finds

$$G_A(x,y) = H(x,y)/D(x,y), \tag{2.6}$$

where $D(x,y)$ is the determinant of the square-matrix appearing in Eq. (2.5) and $H(x,y)$ is the cofactor associated with its first diagonal element. The polynomial structure of $H(x,y)$ and $D(x,y)$ follows immediately from the structure of this matrix. Although irrelevant to the discussion of the problem of interest to us, it is worthwhile remarking that decoupling of the recurrence relation Eq. (2.2) can be achieved by straightforward manipulations on Eq. (2.6).[6]

From now on, we will assume that Eq. (2.6) has been worked out explicitly for a given set of values of L and M. Unlike MQC's method, we will not restrict the activity x to take on the value $x = 1$. The major advantage of this approach is the ability to obtain closed form expressions for the configurational grand-canonical partition function $\Delta_N(x)$, the expectation $\langle\theta\rangle_N$, and the variance σ_N^2. Thus, the z-roots of the determinant $D(x,1/z)$, $R_j(x)$, are x-dependent. These roots are the z-solutions of the equation

$$D(x,1/z) = 0. \tag{2.7}$$

It then follows that the bivariant generating function may be expanded in terms of the $R_j(x)$, namely

$$G_A(x,y) = \sum_j \frac{k_j(x)}{[1-yR_j(x)]} = \sum_{N=0}^{\infty} y^N \sum_j k_j(x)[R_j(x)]^N, \tag{2.8}$$

where the coefficients $k_j(x)$ will be determined shortly.

A. The calculations of Δ_N

Equation (2.8), when compared with the equivalent expression of the bivariant generating function $G_A(x,y)$,

$$G_A(x,y) = \sum_{N=0}^{\infty} y^N \sum_{q=0}^{q_{\max}} x^q A(q,N), \tag{2.9}$$

allows one to obtain $\Delta_N(x)$ in the form

$$\Delta_N(x) \equiv \sum_{q=0}^{q_{\max}} x^q A(q,N) = \sum_j k_j(x)[R_j(x)]^N. \tag{2.10}$$

Clearly, the normalization Δ_N calculated in particular cases by MQC, is the function $\Delta_N(x)$ evaluated at $x = 1$,

$$\Delta_N \equiv \Delta_N(1). \tag{2.11}$$

We now explicitly evaluate $k_i(x)$ in terms of the polynomials $H(x,y)$ and $D(x,y)$. This is achieved by expanding $G_A(x,y)$ in the neighborhood of one of the roots, $y = 1/R_i(x)$. The denominator $D(x,y)$ in Eq. (2.6) is given by

$$D(x,y) = D\left(x,\frac{1}{R_i}\right) + \left(y-\frac{1}{R_i}\right) \times\left(\frac{\partial D}{\partial y}\right)_{y=1/R_i(x)} + O\left(y-\frac{1}{R_i}\right)^2. \tag{2.12}$$

Since R_i is one of the solutions of Eq. (2.7), $D(x,1/R_i) = 0$. Using the relation,

$$\frac{H(x,y)}{D(x,y)} = \sum_j \frac{k_j(x)}{[1-yR_i(x)]} \tag{2.13}$$

and isolating the term in the j-summation corresponding to $j = 1$, then replacing $D(x,y)$ by its equivalent from (2.12) and multiplying both sides of Eq. (2.13) by $(y - 1/R_i)$, one obtains an expression valid in the neighborhood of $y = 1/R_i$, namely

$$\frac{H(x,y)}{(\partial D/\partial y)_{yR_i=1} + O(y-1/R_i)} = -\frac{k_i(x)}{R_i(x)} + \left(\frac{y-1}{R_i}\right) \times \sum_{j\neq i} \frac{k_j(x)}{[1-yR_j(x)]}. \tag{2.14}$$

We then set $y = 1/R_i$ in Eq. (2.14) and obtain

$$k_i(x) = \frac{-R_i(x)H(x,1/R_i)}{[\partial D/\partial y]_{yR_i=1}}. \tag{2.15}$$

This gives the closed form expression of $\Delta_N(x)$,

$$\Delta_N(x) = -\sum_j \frac{[R_j(x)]^{N+1}H(x,1/R_j)}{[\partial D/\partial y]_{y=1/R_j(x)}}. \tag{2.16}$$

Let $R_1(x)$ be the largest root for any given activity x ($x \geqslant 0$), then in the thermodynamic limit the partition function is given by

$$\Xi(x) = \lim_{N\to\infty} [\Delta_N(x)]^{1/LMN} = [R_1(x)]^{1/LM}. \tag{2.17}$$

By setting $x = 1$ in Eq. (2.16), we obtain the exact closed form expression of the normalization, $\Delta_N \equiv \Delta_N(1)$. $R_j(1)$ are the z-roots of $D(1,1/z)$. Should there exist nonsimple roots of $D(1,1/z)$, the analysis above, which assumes simple roots, can be easily adjusted. Nevertheless, it is possible to show that the forthcoming discussion will remain unaffected as long as the root of largest modulus is real, positive and simple. Under these circumstances, let R_1 be the largest real root of $D(1,1/z)$. Setting $x=1$ in Eq. (2.16) and factoring out R_1^{N+1}, one finds

$$\Delta_N(1) = \Delta_N = -R_1^{N+1} \sum_j \left(\frac{R_j}{R_1}\right)^{N+1} \frac{H(1,1/R_j)}{[\partial D/\partial y]_{yR_j=1,\ x=1}}. \tag{2.18}$$

This clearly shows that the leading term of Eq. (2.18), when N becomes large, is

$$\Delta_N = -R_1^{N+1} \frac{H(1,1/R_i)}{[\partial D/\partial y]_{yR_1=1}} \left\{1 + O\left(\frac{R_j}{R_1}\right)^{N+1}\right\}. \tag{2.19}$$

B. The calculation of $\langle\theta\rangle_N$

Consider the first derivative of $\Delta_N(x)$, Eq. (2.10)

$$\Delta_N'(x) = \sum_{q=0}^{q_{\max}} x^{q-1} qA(q,N). \tag{2.20}$$

Since there are LMN compartments on a $L \times M \times N$-lattice space, and since each dumbbell occupies two adjacent compartments, the maximum number $q_{\max}$ of dumbbells that may occupy such a lattice is the integer part of $LMN/2$. We use the symbolic bracket notation $[LMN/2]$ to refer to integer division, i.e.,

$$q_{\max} = [LMN/2]. \tag{2.21}$$

Generalizing definition (1.5), the expectation on the lattice may then be given as a function of x as

$$\langle\theta\rangle_N = \frac{1}{[LMN/2]} \left(\frac{\Delta_N'(x)}{\Delta_N(x)}\right), \tag{2.22}$$

where $\Delta_N'(x)$ is defined as $\partial\Delta_N(x)/\partial x$,

$$\Delta_N'(x) = \sum_j \{k_j' R_j^N + N k_j R_j^{N-1} R_j'\}. \tag{2.23}$$

This enables one to obtain the closed form expression

$$\langle\theta\rangle_N = \frac{1}{[LMN/2]} \left(\frac{\sum_j [k_j' R_j^N + N k_j R_j^{N-1} R_j']}{\sum_j k_j R_j^N}\right). \tag{2.24}$$

So $R_j(x)$ are numerically calculated from $D(x,1/z)=0$ and the associated values $k_j(x)$ are calculated using Eq. (2.15). On the other hand, the quantities $R_j(x)$ have still to be evaluated. The values, $k_j'(x)$ may then be calculated after taking the derivative of Eq. (2.15).

Since $R_i(x)$ are completely independent of the polynomial $H(x,y)$, if we call $t_i(x)$ the expression of $k_i(x)$ when $H(x,y)\equiv 1$, it then follows that

$$\frac{1}{D(x,y)} = \sum_j \frac{t_j(x)}{[1 - yR_j(x)]}, \tag{2.25}$$

and

$$t_i(x) = \frac{-R_i(x)}{[\partial D/\partial y]_{yR_i=1}}. \tag{2.26}$$

We now take the partial derivative of Eq. (2.24) with respect to x:

$$-\frac{1}{D^2}\frac{\partial D}{\partial x} = \sum_j \frac{[t_j'(1 - yR_j) + yR_j' t_j]}{[1 - yR_j(x)]^2}. \tag{2.27}$$

According to Eq. (2.12), an expansion of D^2 in the neighborhood of $y = 1/R_i(x)$ is

$$D^2(x,y) = \left(y - \frac{1}{R_i}\right)^2 \left(\frac{\partial D}{\partial y}\right)^2_{yR_i=1} + O\left(y - \frac{1}{R_i}\right)^3. \tag{2.28}$$

We multiply both sides of Eq. (2.27) by $(y - 1/R_i)^2$, replace D^2 by its approximate expansion (2.28), and then take the limit as y approaches $1/R_i(x)$. We obtain

$$\frac{\partial D(x,1/R_i)}{\partial x}\frac{\partial D(x,1/R_i)}{\partial y} = \frac{t_i(x) R_i'(x)}{[R_i(x)]^3}. \tag{2.29}$$

Finally, the expresson of $t_i(x)$ given by Eq. (2.26) may be used in Eq. (2.28) and yields

$$R_i'(x) = R_i^2(x) \left(\frac{\partial D/\partial x}{\partial D/\partial y}\right)_{yR_i=1}. \tag{2.30}$$

The forthcoming derivation remains unaffected as long as the root of largest modulus $R_1(x)$ is a real, positive, and simple root. Then, as N becomes large, the leading terms in Eq. (2.23) are

$$\langle\theta\rangle_N = \frac{1}{[LMN/2]} \left\{\frac{k_1'(x)}{k_1(x)} + N\frac{R_1'(x)}{R_1(x)} + N \times O\left(\frac{R_j}{R_1}\right)^N\right\}. \tag{2.31}$$

Taking the limit as N becomes infinitely large, and using Eq. (2.30), one finds

$$\langle\theta\rangle_\infty = \frac{2}{LM}\left(\frac{1}{y}\frac{\partial D/\partial x}{\partial D/\partial y}\right)_{yR_i=1}. \tag{2.32}$$

First-order corrections are readily available from Eq. (2.31).

C. The calculation of σ_N^2

The method of calculating σ_N^2 is straightforward. The starting point is the expression of the second derivative of $\Delta_N(x)$ as computed from Eq. (2.20):

$$\Delta_N''(x) = \sum_{q=0}^{q_{\max}} x^{q-2}[q^2 A(q,N) - qA(q,N)]. \tag{2.33}$$

Generalizing Eq. (1.6) and combining Eqs. (2.20), and (2.33), it then follows that

$$\langle\theta^2\rangle_N = \frac{1}{[LMN/2]^2}\left(\frac{\Delta_N''(x) + \Delta_N'(x)}{\Delta_N(x)}\right). \tag{2.34}$$

The ratio $\Delta_N'(x)/\Delta_N(x)$ was calculated previously [see Eqs. (2.22) and (2.31)]. Similar calculations give

$$\frac{\Delta_N''}{\Delta_N} = \frac{\Sigma_j \{k_j'' + 2Nk_j'(R_j'/R_1) + Nk_j(R_j''/R_1) + N(N-1)k_j(R_j'/R_1)^2\}(R_j/R_1)^N}{\Sigma_j k_j(R_j/R_1)^N}. \tag{2.35}$$

The leading terms of Eq. (2.35) as N becomes large are

$$\frac{\Delta''_N}{\Delta_N} = \frac{k''_1}{k_1} + 2N\frac{k'_1}{k_1}\frac{R'_1}{R_1} + N\frac{R''_1}{R_1} + N(N-1)\Big(\frac{R'_1}{R_1}\Big)^2 + O\Big(\frac{R_j}{R_1}\Big)^N N^2. \tag{2.36}$$

Thus, the large N behavior of $\langle\theta\rangle_N$ is given by

$$\langle\theta^2\rangle_N = \frac{4}{L^2M^2}\Big\{\frac{R'^2_1}{R_1^2} + \frac{1}{N}\Big(2\frac{k'_1}{k_1}\frac{R'_1}{R_1} + \frac{R'_1}{R_1} + \frac{R''_1}{R_1} - \frac{R'^2_1}{R_1^2}\Big) + \frac{1}{N^2}\Big(\frac{k''_1}{k_1} + \frac{k'_1}{k_1}\Big) + O\Big(\frac{R_j}{R_1}\Big)^N\Big\}. \tag{2.37}$$

Clearly, in the limit as N becomes infinite, one has

$$\langle\theta^2\rangle_\infty = \frac{4}{L^2M^2}\Big(\frac{R'_1}{R_1}\Big)^2 = (\langle\theta\rangle_\infty)^2. \tag{2.38}$$

This shows that

$$\sigma^2_\infty = 0. \tag{2.39}$$

The first-order correction to σ^2_N may be calculated by taking the square of Eq. (2.31), keeping the leading terms, i.e.,

$$(\langle\theta\rangle_N)^2 = \frac{4}{L^2M^2}\Big\{\frac{R'^2_1}{R_1^2} + \frac{2}{N}\frac{k'_1}{k_1}\frac{R'_1}{R_1} + \frac{1}{N^2}\frac{k'^2_1}{k_1^2} + O\Big(\frac{R_j}{R_1}\Big)^N\Big\}. \tag{2.40}$$

We take the difference between Eqs. (2.36) and (2.39) and find

$$\sigma^2_N = \frac{4}{L^2M^2}\Big\{\frac{1}{N}\frac{R'_1}{R_1}\Big(1 + \frac{R''_1}{R'_1} - \frac{R'_1}{R_1}\Big) + \frac{1}{N^2}\frac{k'_1}{k_1}\Big(1 + \frac{k''_1}{k'_1} - \frac{k'_1}{k_1}\Big) + O\Big(\frac{R_j}{R_1}\Big)^N\Big\}. \tag{2.41}$$

Again, the above result holds when the root R_1 of largest modulus is a simple, real and positive root. We now calculate the first-order correction by looking at an explicit expression for R''_1/R'_1. This is done by taking the logarithmic derivative of R'_1 using Eq. (2.30) and the fact that in this equation one has

$$y = \frac{1}{R_i(x)} \quad\text{and}\quad \frac{dy}{dx} = -\frac{R'_i(x)}{R^2_i(x)}. \tag{2.42}$$

It follows that

$$\frac{R''_i}{R'_i} = 2\frac{R'_i}{R_i} + \Big(\frac{\partial^2 D/\partial x^2 - (\partial^2 D/\partial x\,\partial y)(R'_i/R^2_i)}{\partial D/\partial x}\Big)_{yR_i=1} - \Big(\frac{\partial^2 D/\partial x\,\partial y - (\partial^2 D/\partial y^2)(R'_i/R^2_i)}{\partial D/\partial y}\Big)_{yR_i=1}. \tag{2.43}$$

TABLE I. The entries A, B, C, etc., correspond to sublattices generated from the original lattice space $1\times M\times N$ by removing from the Nth array, no compartment, one, two, etc., up to $(M-1)$ compartments, respectively. This table gives the number of topologically distinct sublattices belonging to the same class, and for various values of the lattice size M. The last column in this table gives the total number of bivariant generating functions $G_{T_i}(x,y)$ for various values of the lattice size M. Since this number is equal to the total number of T_i- lattices, it is obtained by adding the number in the table falling on the same horizontal line.

SIZE M \ CLASS T	A	B	C	D	E	TOTAL NUMBER OF GENERATING FUNCTIONS
1	1					1
2	1	1				2
3	1	2	2			5
4	1	2	4	2		9
5	1	3	6	6	3	19

This result may be combined with the properties

$$\frac{R'_i(x)}{R^2_i(x)} = \Big(\frac{\partial D/\partial x}{\partial D/\partial y}\Big)_{yR_i=1}, \quad \langle\theta\rangle_\infty = \frac{2}{LM}\frac{R'_1(1)}{R_1(1)} \tag{2.44}$$

to give the leading term of σ^2_N for large N:

$$\sigma^2_N = \frac{2}{LMN}\langle\theta\rangle_\infty\Big\{1 + \frac{LM}{2}\langle\theta\rangle_\infty + \frac{\partial^2 D/\partial x^2}{\partial D/\partial x} + \frac{(\partial D/\partial x)(\partial^2 D/\partial y^2)}{(\partial D/\partial y)^2} - 2\frac{\partial^2 D/\partial x\,\partial y}{\partial D/\partial y}\Big\}_{yR_i=1} + O(1/N)^2. \tag{2.45}$$

To complete this theoretical study of the occupation statistics for indistinguishable dumbbells on a rectangular lattice space, $L\times M\times N$, for finite N and for the limit as N becomes very large, we conclude with the Gaussian representation of $A(q,N)$. Introducing the normalized quantity

$$\theta = q/q_{\max} \tag{2.46}$$

we obtain the Gaussian distribution valid for large values of N, namely

$$A(q,N) = A_{\max}\exp(-[\theta - \langle\theta\rangle_N]^2/\sigma^2_N), \tag{2.47}$$

where $\langle\theta\rangle_N$ and σ_N are calculated from Eqs. (2.32) and (2.45) setting $x = 1$. $A_{\max}$ is the maximum number of possible arrangements on the $L\times M\times N$ lattice given approximately by

$$A_{\max} = \Delta_N/\sqrt{8\pi N\sigma^2_N}. \tag{2.48}$$

We now proceed to applying this general formalism to $1\times M\times N$ lattices for $M = 1, 2, 3$, and 4. In Table I, we give the number of distinct sublattices of given class $\mathscr{T}$ for every case to be investigated.

III. THE $1\times1\times N$ LATTICE

This one-dimensional problem has already been solved exactly.[1] We reproduce the known results for the sake of completeness and as a mere application of the general formalism developed in Sec. II. Following MQH, the recurrence relation for the number of arrangements $A(q,N)$ of q dumbbells on this lattice is straightforward. This number is equal to the sum of all possible arrangements of q dumbbells leaving the Nth compartment unoccupied $A(q,N-1)$, and all possible arrangements, $A(q-1,N-2)$, of $(q-1)$ dumbbells when the last two compartments are occupied by one dumbbell. This is illustrated in Fig. 1. So, very simple, one has

$$A(q,N) = A(q,N-1) + A(q-1,N-2). \tag{3.1}$$

This is a special case of Eq. (2.2) which stands for a set of equations, equal in number to the number of distinct T_i-lattices. In the one-dimensional problem, we have two $c_{nm}[T_i,T_j]$ coefficients, namely

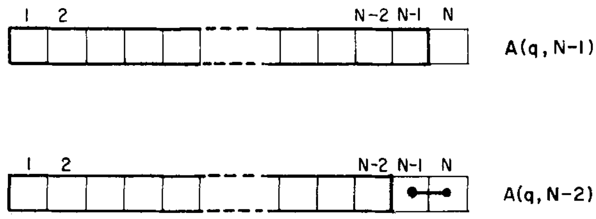


FIG. 1. Generating a recurrence relation for the number of arrangements $A(q,N)$ of q dumbbells on a $1\times1\times N$-lattice.

$$c_{01}[A,A]=1, \quad c_{12}[A,A]=1. \tag{3.2}$$

For this problem, Eq. (2.4a) becomes

$$G_A(x,y)=1+yG_A(x,y)+xy^2G_A(x,y) \tag{3.3}$$

and we end up with one equation with one unknown. Then matrix equation (2.5) has a trivial form

$$(1-y-xy^2)G_A(x,y)=1 \tag{3.4}$$

leading to the simplest expressions for $H(x,y)$ and $D(x,y)$,

$$H(x,y)=1, \quad D(x,y)=1-y-xy^2. \tag{3.5}$$

The z-roots of $D(1,1/z)$ are easily obtained from setting

$$D(1/z)=(1/z^2)(z^2-z-x)=0. \tag{3.6}$$

These are

$$R_j(x)=\tfrac{1}{2}[1-(-1)^j\sqrt{1+4x}\,], \quad j=1,2. \tag{3.7}$$

The root of largest modulus is real and positive and we proceed in the application of our formalism, by calculating the first- and second-order partial derivatives of $D(x,y)$ with respect to x and y, namely

$$\frac{\partial D}{\partial D}=-y^2, \quad \frac{\partial^2 D}{\partial x^2}=0, \quad \frac{\partial^2 D}{\partial x\partial y}=-2y;$$
$$\frac{\partial D}{\partial y}=-(1+2xy), \quad \frac{\partial^2 D}{\partial y^2}=-2x. \tag{3.8}$$

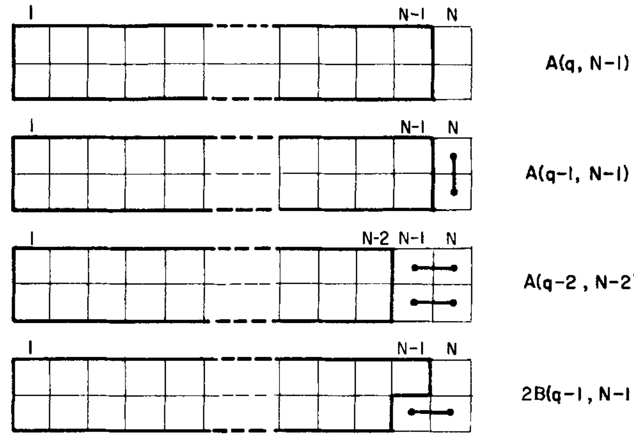


FIG. 2A. Generating a recurrence relation for the number of arrangements $A(q,N)$ of q dumbbells on a $1\times2\times N$-lattice.

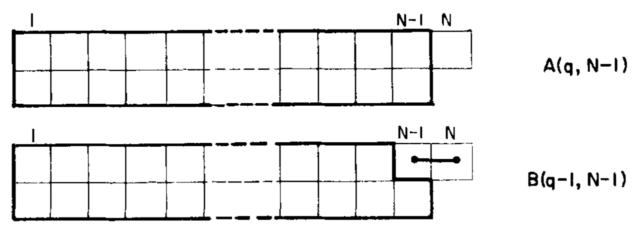


FIG. 2B. Generating a recurrence relation for the number of arrangements $B(q,N)$ of q dumbbells on a truncated $1\times2\times N$-lattice of class B.

Using Eq. (2.16), we obtain the exact value of the grand-canonical partition function,

$$\Delta_N(x)=\left(\frac{1+\sqrt{1+4x}}{2}\right)^{N+1}\sum_{j=1}^{2}\times\left(\frac{1-(-1)^j\sqrt{1+4x}}{1+\sqrt{1+4x}}\right)^{N+1}\frac{1-(-1)^j\sqrt{1+4x}}{1+4x-(1)^j\sqrt{1+4x}}. \tag{3.9}$$

TABLE II. Arrangements of q dumbbells on a $1\times1\times N$ lattice and the normalization factor Δ_N.

N \ q	0	1	2	3	4	5	EXACT Δ_N	APPROXIMATE Δ_N	ERROR %
0	1						1	0.7236	27.64
1	1						1	1.1708	17.08
2	1	1					2	1.8944	5.28
3	1	2					3	3.0652	2.17
4	1	3	1				5	4.9597	0.807
5	1	4	3				8	8.0249	0.312
6	1	5	6	1			13	12.9846	0.118
7	1	6	10	4			21	21.00952	0.045
8	1	7	15	10	1		34	33.99412	0.017
9	1	8	21	20	5		55	55.00364	0.0066
10	1	9	28	35	15	1	89	88.99775	0.0025

In the thermodynamic limit, Eq. (2.17) yields

$$\Xi(x) = \lim_{N\to\infty} [\Delta_N(x)]^{1/N} = [1 + \sqrt{1+4x}]/2. \quad (3.10)$$

For large values of N, and for $x = 1$, the normalization Δ_N,

$$\Delta_N = (1 + \sqrt{5})^{N+2}/2^{N+1}(5 + \sqrt{5}). \quad (3.11)$$

The accuracy of this relation improves with increasing values of N. Table II shows how good this approximation is for small values of N. This approximation gives a better than one percent agreement for $N = 4$. The expectation $\langle\theta\rangle_\infty$, on the infinite one-dimensional lattice is calculated using Eqs. (2.32) and (3.8):

$$\begin{aligned}\langle\theta\rangle_\infty &= 2\left(\frac{1}{y}\frac{-y^2}{-(1+2xy)}\right)_{\substack{yR_1=1\\ x=1}}\\ &= \frac{1}{x}\left[1 - \frac{1}{\sqrt{1+4x}}\right]_{x=1}\\ &= 1 - \sqrt{5}/5 = 0.552\,786\,405. \quad (3.12)\end{aligned}$$

Similarly, combining Eqs. (2.44) and (3.8) one obtains for $x = 1$

$$\sigma_N^2 = \frac{4\sqrt{5}}{25}\frac{1}{N} = \frac{0.357\,770\,876}{N}. \quad (3.13)$$

IV. THE $1\times2\times N$ LATTICE

This lattice problem has also been solved.[1] Our general formalism may be applied again to give, in just a few steps, the interesting statistical quantities. Table I shows that there is only one sublattice belonging to the B-class, and consequently, two bivariant generating functions G_A and G_B. Again, following MQC, Figs. 2A and 2B show how to find the recurrence relations Eq. (2.2). One finds

$$\begin{aligned}A(q,N) &= A(q,N-1) + A(q-1,N-1)\\ &\quad + A(q-2,N-2) + 2B(q-1,N-1), \quad (4.1)\\ B(q,N) &= A(q,N-1) + B(q-1,N-1).\end{aligned}$$

By comparison with Eq. (2.2) we obtain the $c_{nm}[T_i,T_j]$ coefficients

$$\begin{aligned}&C_{01}[A,A] = C_{11}[A,A] = C_{22}[A,A] = 1,\\ &C_{11}[A,B] = 2,\ \ C_{01}[B,A] = C_{11}[B,B] = 1. \quad (4.2)\end{aligned}$$

Consequently, for this lattice, Eqs. (2.4a) and (2.4b) reduce to

$$\begin{aligned}G_A(x,y) &= 1 + (y + xy + x^2y^2)G_A(x,y) + 2xyG_B(x,y),\\ G_B(x,y) &= yG_A(x,y) + xyG_B. \quad (4.3)\end{aligned}$$

The matrix equation becomes

$$\begin{pmatrix}(1 - y(x+1) - x^2y^2) & -2xy\\ -y & 1-xy\end{pmatrix}\begin{pmatrix}G_A\\ G_B\end{pmatrix} = \begin{pmatrix}1\\ 0\end{pmatrix}. \quad (4.4)$$

The determinant of the square matrix obtained is

$$\begin{aligned}D(x,y) &= (1-xy)[1 - y(x+1) - x^2y^2] - 2xy^2\\ &= 1 - y(1+2x) - xy^2 + x^3y^3 \quad (4.5)\end{aligned}$$

and the cofactor of its first diagonal element is

$$H(x,y) = 1 - xy. \quad (4.6)$$

It then follows that

$$G_A(x,y) = \frac{H(x,y)}{D(x,y)} = \frac{1-xy}{1 - y(1+2x) - xy^2 + x^3y^3}. \quad (4.7)$$

As an illustration of the flexibility of the general formalism developed, we would like to recover the decoupled recurrence relation from the expression of $D(x,y)$. It is sufficient to mention here, without any additional derivations, that with each term of the form $\lambda x^n y^m$ one associates $\lambda A(q-n,N-m)$. The one-to-one correspondence that follows produces the uncoupled recurrence relation

$$\begin{aligned}&A(q,N) - A(q,N-1) - 2A(q-1,N-1)\\ &\quad - A(q-1,N-2) + A(q-3,N-3) = 0. \quad (4.8)\end{aligned}$$

This is the recurrence relation derived by McQuistan and Lichtman (MQL).[1] As found by MQL, the roots of $D(1,1,z)$ are

$$\begin{aligned}&R_1 = 3.214\,319\,743,\quad R_2 = 0.460\,811\,121,\\ &R_3 = -0.675\,130\,871. \quad (4.9)\end{aligned}$$

Again, the root of largest modulus is a simple root, real and positive. For this value of the z-root, we calculate the first and second partial derivatives of $D(x,y)$

$$\frac{\partial D(1,1/R_1)}{\partial x} = -2y - y^2 + 3x^2y^3 = -0.628\,699\,130, \quad (4.10a)$$

$$\frac{\partial^2 D(1,1/R_1)}{\partial x^2} = 6xy^3 = 0.180\,699\,159, \quad (410b)$$

$$\frac{\partial^2 D(1,1/R_1)}{\partial x\,\partial y} = -2 - 2y + 9x^2y^2 = -1.751\,122\,968, \quad (4.10c)$$

$$\begin{aligned}\frac{\partial D(1,1/R_1)}{\partial y} &= -(1+2x) - 2xy + 3x^3y^2\\ &= -3.331\,851\,413, \quad (4.10d)\end{aligned}$$

$$\frac{\partial^2 D(1,1/R_1)}{\partial y^2} = -2x + 6x^3y = 0.133\,353\,095, \quad (4.10e)$$

$$H(1,1/R_i) = 1 - xy = 0.688\,921\,83. \quad (4.10f)$$

With these numerical results used in Eqs. (2.18), (2.31), and (2.44), we recover MQL's results, for $x = 1$, namely,

$$\Delta_N = 0.664\,591\,384(3.214\,319\,743)^N, \quad (4.11)$$

$$\langle\theta\rangle_\infty = 0.606\,492\,711, \quad (4.12)$$

and we also obtain the value of σ_N^2 that MQL do not report in their paper,

$$\sigma_N^2 = 0.167\,100\,836(1/N) + O(1/N)^2. \quad (4.13)$$

For the purpose of comparing the rate of convergence for the various lattice problems discussed in this paper, we reproduce in Table III, MQL's table for $A(q,N)$ and add to it the exact Δ_N and the approximate Δ_N calculated from Eq. (4.11). We notice from this table that the agreement of the approximate formula with the exact value is better than (0.1) percent for $N = 4$.

Finally, for completeness we give the explicit x-dependence of the z-roots of $D(x,1/z)$, since

$$D(x,1/z) = (1/z^3)[z^3 - (1+2x)z^2 - xz + x^3] = 0 \quad (4.14)$$

is a cubic equation. The three roots of this equation are real and given by

TABLE III. Arrangements of q dumbbells on a $1\times2\times N$ lattice and the normalization factor Δ_N.

N \ q	0	1	2	3	4	5	6	7	8	9	Δ_N EXACT	Δ_N APPROXIMATE
0	1										1	0.6645913
1	1	1									2	2.136209
2	1	4	2								7	6.866459
3	1	7	11	3							22	22.07100
4	1	10	29	26	5						71	70.94324
5	1	13	56	94	56	8					228	228.03425
6	1	16	92	234	263	114	13				733	732.97400
7	1	19	137	473	815	667	223	21			2356	2356.0160
8	1	22	191	838	1982	2504	1577	424	34		7573	7572.9888
9	1	25	254	1356	4115	7191	7018	3538	789	55	24342	24342.007

$$R_j(x) = \frac{1+2x}{3} + \frac{2}{3}[4x^2+7x+1]^{1/2} \times\cos\left[\frac{\phi(x)}{3} + (j-1)\frac{2\pi}{3}\right], \quad j=1,2,3, \tag{4.15}$$

where

$$\cos\phi(x) = -\frac{(11x^3-42x^2-21x-2)}{2(4x^2+7x+1)^{3/2}}. \tag{4.16}$$

The advantage is that one is able to obtain the explicit closed form expression of the partition function

$$\Xi(x) = [R_1(x)]^{1/2},$$

where $R_1(x)$ is the largest of the three roots given in (4.15). In particular, one recovers the results obtained with the HP 34C when the activity x is set equal to unity.

V. THE $1\times3\times N$ LATTICE

Here, the situation is a bit more complicated than in the previous two examples. According to Table I, we have two sublattices B_1 and B_2 belonging to class B, and two sublattices C_1 and C_2 belonging to class C. These topologically distinct sublattices are represented in Fig. 3. Figures 3A, $3B_1$, $3B_2$, $3C_1$, and $3C_2$ give the diagrams associated with the generation of the recurrence relations according to MQH. In these figures, only the last 3 rows of compartments are shown, namely rows $N-2$, $N-1$, and N, respectively. The coupled recurrence relations that follow from the analysis of these figures are

$$\begin{aligned} A(q,N) &= A(q,N-1) + 2A(q-1,N-1) + A(q-3,N-2) + 2B_1(q-1,N-1) \\ &\quad + 2B_1(q-2,N-1) + B_2(q-1,N-1) + 2C_1(q-2,N-1) + C_2(q-2,N-1), \\ B_1(q,N) &= A(q,N-1) + A(q-1,N-1) + B_1(q-1,N-1) + B_2(q-1,N-1) + C_1(q-2,N-1), \\ B_2(q,N) &= A(q,N-1) + 2B_1(q-1,N-1) + C_2(q-2,N-1), \\ C_1(q,N) &= A(q,N-1) + B_1(q-1,N-1), \\ C_2(q,N) &= A(q,N-1) + B_2(q-1,N-1). \end{aligned} \tag{5.1}$$

The coefficients $c_{nm}[T_i,T_j]$ are easily identified from Eq. (5.1) and the matrix equation (2.3) follows

$$\begin{pmatrix} 1-y(1+2x)-x^3y^2 & -2xy(1+x) & -xy & -2x^2y & -x^2y \\ -y(1+x) & 1-xy & -xy & -x^2y & 0 \\ -y & -2xy & 1 & 0 & -x^2y \\ -y & -xy & 0 & 1 & 0 \\ -y & 0 & -xy & 0 & 1 \end{pmatrix} \begin{pmatrix} G_A \\ G_{B_1} \\ G_{B_2} \\ G_{C_1} \\ G_{C_2} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{5.2}$$

The cofactor of the first diagonal element of the square matrix is

$$H(x,y) = 1 - xy - 2x^2y^2(1+x) + x^4y^3 + x^6y^4 \tag{5.3}$$

and the determinant of the square matrix is

$$D(x,y) = 1-(1+3x)y - xy^2(2+7x+5x^2) + x^2y^3(2x^2-x-1) + x^4y^4(2+3x+5x^2) - x^6y^5(1-x) - x^9y^6. \tag{5.4}$$

The decoupled recurrence relation for $A(q,N)$ is obtained following the rule stated in Sec. IV:

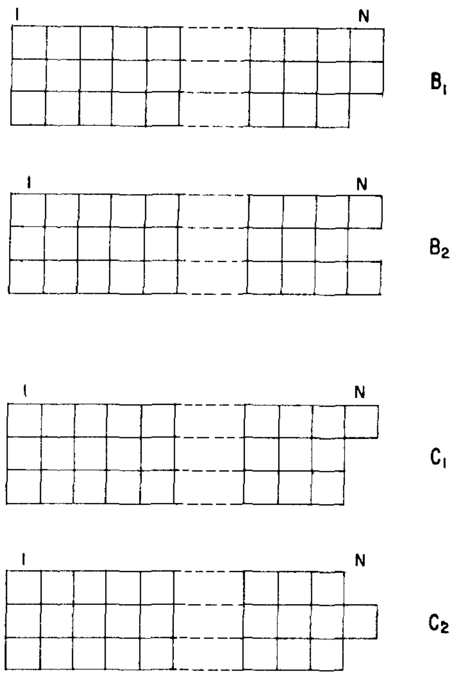


FIG. 3. B- and C-class of sublattices for the $1\times3\times N$-lattice problem.

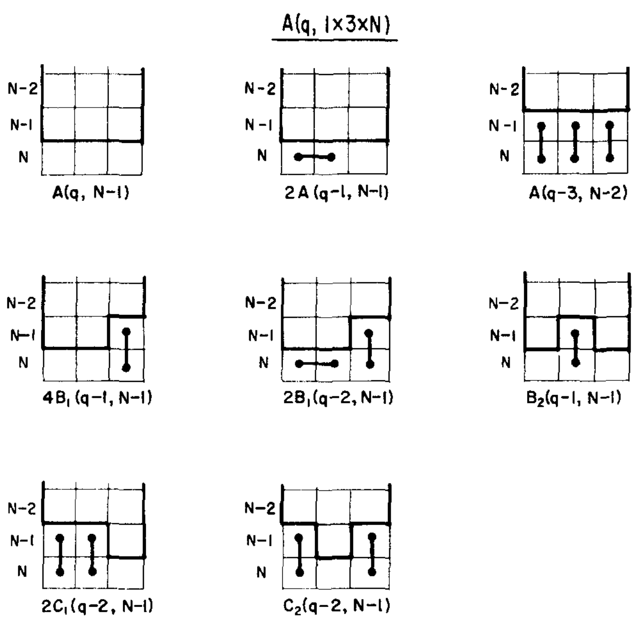


FIG. 3A. Generating a recurrence relation for the number of arrangements $A(q,N)$ of q dumbbells on a $1\times3\times N$-lattice.

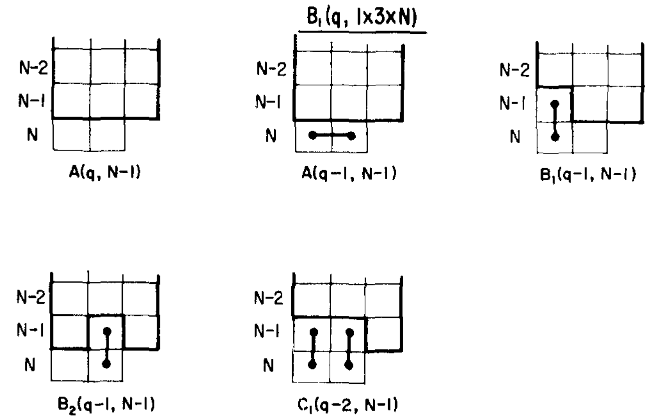


FIG. 3B$_1$. Generating a recurrence relation for the number of arrangements $B(q,N)$ of q dumbbells on a truncated $1\times3\times N$-lattice of type B_1.

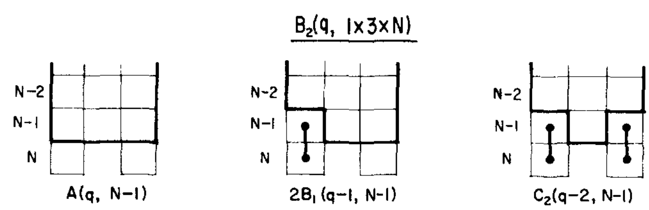


FIG. 3B$_2$. Generating a recurrence relation for the number of arrangements $B(q,N)$ of q dumbbells on a truncated $1\times3\times N$-lattice of type B_2.

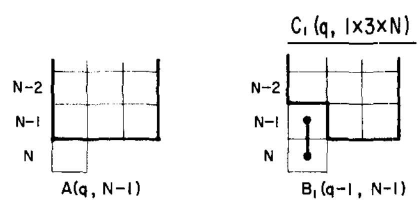


FIG. 3C$_1$. Generating a recurrence relation for the number of arrangements $C_1(q,N)$ of q dumbbells on a truncated $1\times3\times N$-lattice of type C_1.

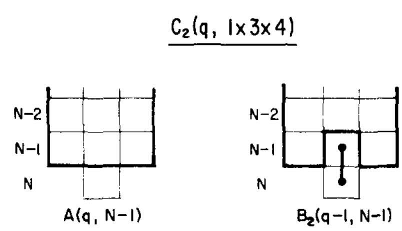


FIG. 3C$_2$. Generating a recurrence relation for the number of arrangements $C_2(q,N)$ of q dumbbells on a truncated $1\times3\times N$-lattice of type C_2.

$$\begin{aligned}A(q,N) &- A(q,N-1) - 3A(q-1,N-1) - 2A(q-1,N-2) - 7A(q-2,N-2) - 5A(q-3,N-2)\\ &+ 2A(q-4,N-3) - A(q-3,N-3) - A(q-2,N-3) + 2A(q-4,N-4) + 3A(q-5,N-4)\\ &+ 5A(q-6,N-4) - A(q-6,N-5) + A(q-7,N-5) - A(q-9,N-6) = 0. \qquad (5.5)\end{aligned}$$

This recurrence relation should satisfy the initial values of $A(q,N)$ to be computed by hand for $q = 0$ to 8 and $N = 0$ to 6. We explicitly see now why our formalism is so much easier to handle than the procedure involved in MQH's paper,[1] and the long derivations of MQL.[1]

The roots of $D(1,1/z)$ are easily obtained using the HP-34C pocket calculator. They are found to be

$R_1 = 6.212\,070\,250, \quad R_4 = -0.342\,320\,858,$

$R_2 = 0.698\,204\,338, \quad R_5 = -1,$

$R_3 = 0.350\,983\,683, \quad R_6 = -1.918\,937\,413.$

TABLE IV. Arrangements of q dumbbells on a $1\times3\times N$ lattice and the normalization factor Δ_N.

N \ q	0	1	2	3	4	5	6	7	8	9	10	Δ_N EXACT	Δ_N APPROXIMATE
0	1											1	0.551273070
1	1	2										3	3.424547038
2	1	7	11	3								22	21.27352677
3	1	12	44	56	18							131	132.1526428
4	1	17	102	267	302	123	11					823	820.9415008
5	1	22	185	748	1597	1670	757	106				5086	5099.746274
6	1	27	293	1644	5236	9503	9401	4603	908	41		31657	31679.98211
7	1	32											

All roots are simple and real, and the one of largest modulus is positive. We finally calculate the various first- and second-order derivatives of $D(x,y)$ and evaluate them at $x=1$ and $y=1/R_1$:

$$\frac{\partial D(1,1/R_1)}{\partial x} = -3y - y^2(2+14x+15x^2) + xy^3(8x^2-3x-2) + x^3y^4(8+15x+30x^2) + x^5y^5(7x-6) - 9x^8y^6 = -1.238\,195\,378,$$

$$\frac{\partial^2 D(1,1/R_1)}{\partial x^2} = -y^2(14+30x) + y^3(24x^2-6x-2) + x^2y^4(24+60x+150x^2) + x^4y^5(42x-30) = -0.915\,022\,105,$$

$$\frac{\partial^2 D(1,1/R_1)}{\partial x\,\partial y} = -3 - 2y(2+14x+15x^2) + 3xy^2(8x^2-3x-2) + 4x^3y^3(8+15x+30x^2) + 5x^5y^4(7x-6) - 54x^8y^5 = -11.872\,187\,10,$$

$$\frac{\partial D(1,1/R_1)}{\partial y} = -(1+3x) - 2xy(2+7x+5x^2) + 3x^2y^2(2x^2-x-1) + 4x^4y^3(2+3x+5x^2) - 5x^6y^4(1-x) - 6x^9y^5 = -8.341\,143\,147,$$

$$\frac{\partial^2 D(1,1/R_1)}{\partial y^2} = -2x(2+7x+5x^2) + 6x^2y(2x^2-x-1) + 12x^4y^2(2+3x+5x^2) - 20x^6y^3(1-x) - 30x^9y^4 = -24.910\,516\,74,$$

$$H(1,1/R_1) = 0.740\,211\,782.$$

These results are used to obtain

$$\Delta_N = (0.551\,273\,070)(6.212\,070\,250)^N, \tag{5.6a}$$

$$\langle\theta\rangle_\infty = 0.614\,764\,390, \tag{5.6b}$$

$$\sigma_N^2 = 0.105\,661\,426(1/N). \tag{5.6c}$$

Table IV lists the exact values of Δ_N and the values computed using the approximation formula (5.6a). This time the error is about 0.25 percent at $N=4$.

VI. THE $1\times4\times N$ LATTICE

According to Table I, there are two distinct sublattices of the B- and D-class, respectively, and four of the C-class. These sublattices are identified in Fig. 4. The drawings in Fig. 4 represent the last three rows, the $(N-2)$, $(N-1)$, and Nth row, having four compartments each. Figures 4A, $4B_1$, $4B_2$, $4C_1$, $4C_2$, $4C_3$, $4C_4$, $4D_1$, and $4D_2$ give the diagrams associated with the generation of the coupled recurrence relations among the number of arrangements $A(q,N)$, $B_i(q,N)$, $C_i(q,N)$, and $D_i(q,N)$ of q dumbbells on the associated sublattices. The coupled relations that follow from the analysis of these figures are

$$A(q,N) = A(q,N-1) + 3A(q-1,N-1) + A(q-2,N-1) + A(q-4,N-2) + 2B_1(q-1,N-1) + 4B_1(q-2,N-1) + 2B_2(q-1,N-1) + 2B_2(q-2,N-1) + 2C_1(q-2,N-1) + 2C_1(q-3,N-1) + C_2(q-2,N-1) + C_3(q-2,N-1) + C_3(q-3,N-1) + 2C_4(q-2,N-1) + 2D_1(q-3,N-1) + 2D_2(q-3,N-1),$$

$$
\begin{aligned}
B_1(q,N) = {} & A(q,N-1) + 2A(q-1,N-1) + B_1(q-1,N-1) + 2B_2(q-1,N-1) + B_1(q-2,N-1) \\
& + B_2(q-2,N-1) + C_1(q-2,N-1) + C_2(q-2,N-1) + C_4(q-2,N-1) + D_1(q-3,N-1), \\
B_2(q,N) = {} & A(q,N-1) + A(q-1,N-1) + 2B_1(q-1,N-1) + B_1(q-2,N-1) + B_2(q-1,N-1) \\
& + C_1(q-2,N-1) + C_3(q-2,N-1) + C_4(q-2,N-1) + D_2(q-3,N-1), \\
C_1(q,N) = {} & A(q,N-1) + A(q-1,N-1) + B_1(q-1,N-1) + B_2(q-1,N-1) + C_1(q-2,N-1), \\
C_2(q,N) = {} & A(q,N-1) + 2B_1(q-1,N-1) + C_3(q-2,N-1) \\
C_3(q,N) = {} & A(q,N-1) + A(q-1,N-1) + 2B_2(q-1,N-1) + C_2(q-2,N-1), \\
C_4(q,N) = {} & A(q,N-1) + B_1(q-1,N-1) + B_2(q-1,N-1) + C_4(q-2,N-1), \\
D_1(q,N) = {} & A(q,N-1) + B_1(q-1,N-1), \\
D_2(q,N) = {} & A(q,N-1) + B_2(q-1,N-1).
\end{aligned}
$$

The above nine coupled recurrence relations lead to a system of nine linear equations that give the closed form expressions of the bivariant generating functions. The square matrix associated with this system of linear equation is

$$
\begin{pmatrix}
[1-y(1+3x+x^2)-x^4y^2] & -2xy(1+2x) & -2xy(1+x) & -2x^2y(1+x) & -x^2y & -x^2y(1+x) & -2x^2y & -2x^3y & -2x^3y \\
-y(1+2x) & 1-xy(1+x) & -xy(2+x) & -x^2y & -x^2y & 0 & -x^2y & -x^3y & 0 \\
-y(1+x) & -xy(2+x) & 1-xy & -x^2y & 0 & -x^2y & -x^2y & 0 & -x^3y \\
-y(1+x) & -xy & -xy & 1-x^2y & 0 & 0 & 0 & 0 & 0 \\
-y & -2xy & 0 & 0 & 1 & -x^2y & 0 & 0 & 0 \\
-y(1+x) & 0 & -2xy & 0 & -x^2y & 1 & 0 & 0 & 0 \\
-y & -xy & -xy & 0 & 0 & 0 & 1-x^2y & 0 & 0 \\
-y & -xy & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
-y & 0 & -xy & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}.
$$

The determinant of this 9×9 matrix is not so difficult to calculate, since so many of its elements are zeros. Of course, the same holds for the cofactor of the first diagonal element. One finds

$$
\begin{aligned}
D(x,y) = {} & -x^{28}y^{14} + x^{24}y^{13} + x^{21}y^{12}(1+5x+12x^2+12x^3) - x^{18}y^{11}(1+6x+9x^2+17x^3-4x^4) \\
& - x^{15}y^{10}(1+12x+32x^2+64x^3+103x^4+47x^5) + x^{13}y^9(-2+2x+43x^2+94x^3+34x^4-20x^5) \\
& + x^{11}y^8(7+56x+175x^2+285x^3+273x^4+86x^5) - x^{10}y^7(16+70x+124x^2-14x^3-38x^4) \\
& - x^7y^6(7+60x+227x^2+429x^3+297x^4+82x^5) + x^5y^5(2+14x+21x-36x^3-57x^4-34x^5) \\
& + x^3y^4(1+16x+84x^2+176x^3+143x^4+41x^5) + x^2y^3(1+12x+39x^2+31x^3+14x^4) \\
& - xy^2(1+9x+28x^2+10x^3) - y(1+6x+2x^2) + 1.
\end{aligned}
$$

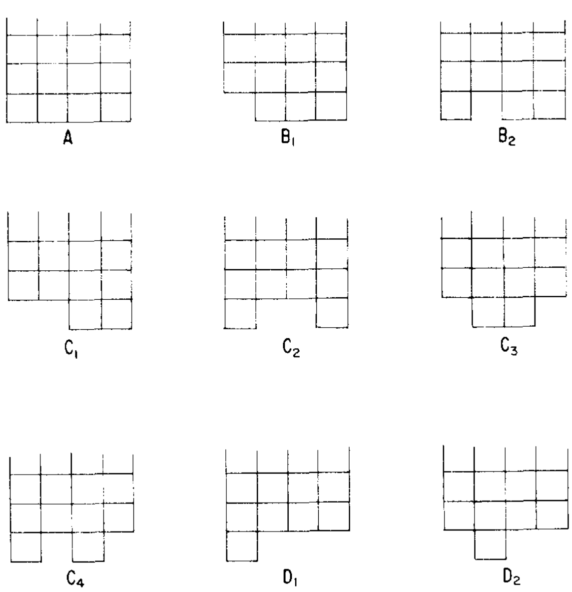


FIG. 4. A-, B-, C-, and D-class of lattices for the $1\times 4\times N$ problem.

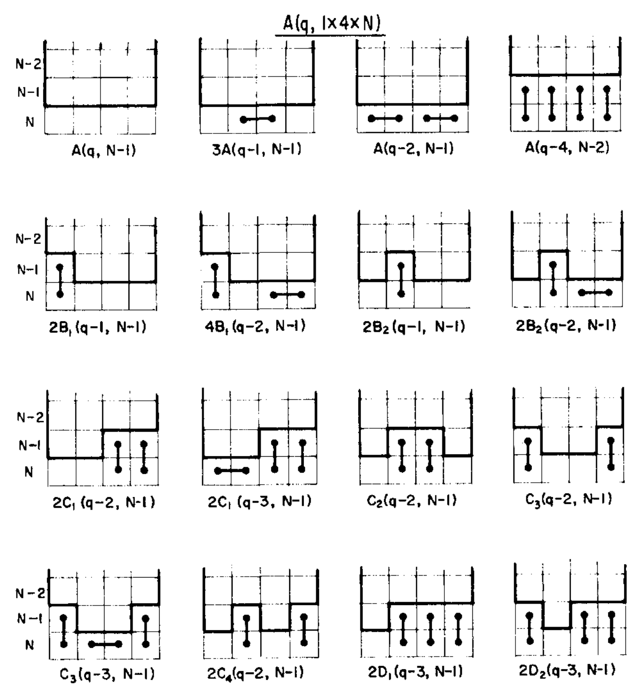


FIG. 4A. Generating a recurrence relation for the number of arrangements $A(q,N)$ of q dumbbells on a $1\times 4\times N$-lattice.

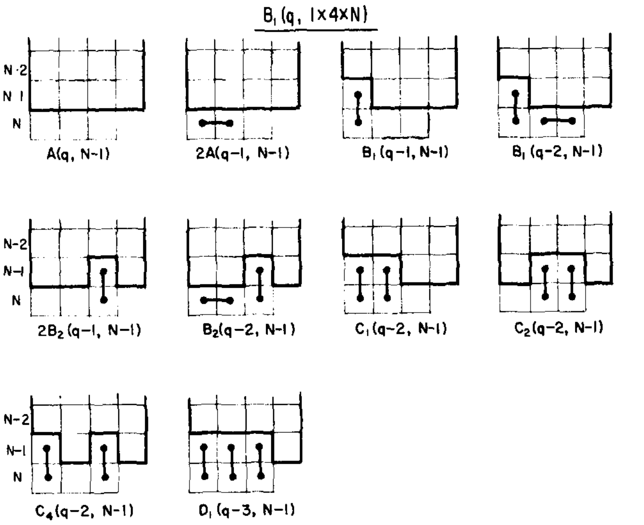


FIG. $4B_1$. Generating a recurrence relation for the number of arrangements $B_1(q,N)$ of q dumbbells on a truncated $1\times 1\times N$-lattice of type B_1.

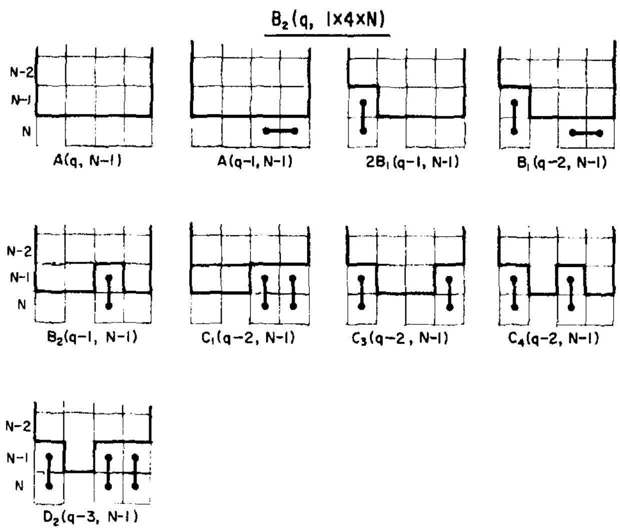


FIG. $4B_1$. Generating a recurrence relation for the number of arrangements $B_1(q,N)$ of q dumbbells on a truncated $1\times 1\times N$-lattice of type B_1.

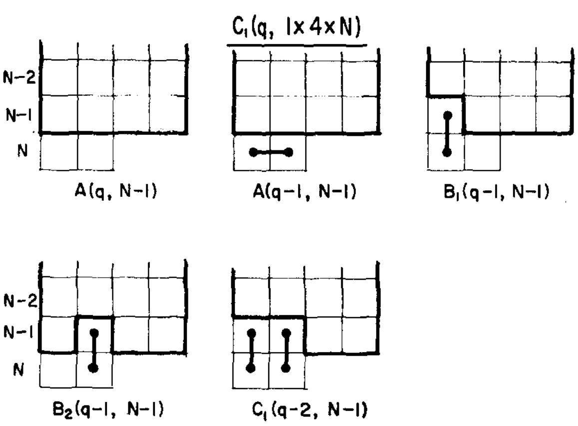


FIG. $4C_1$. Generating a recurrence relation for the number of arrangements $C_1(q,N)$ of q dumbbells on a truncated $1\times 4\times N$-lattice of type C_1.

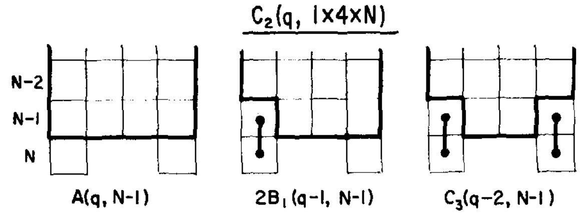


FIG. $4C_2$. Generating a recurrence relation for the number of arrangements $C_2(q,N)$ of q dumbbells on a truncated $1\times 4\times N$-lattice of type C_2.

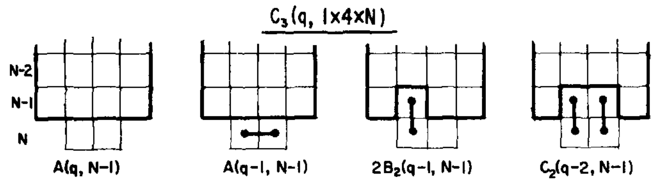


FIG. $4C_3$. Generating a recurrence relation for the number of arrangements $C_3(q,N)$ of q dumbbells on a truncated $1\times 4\times N$-lattice of type C_3.

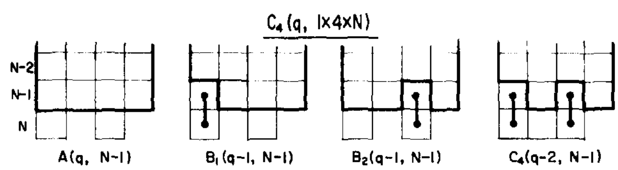


FIG. $4C_4$. Generating a recurrence relation for the number of arrangements $C_4(q,N)$ of q dumbbells on a truncated $1\times 4\times N$-lattice of type C_4.

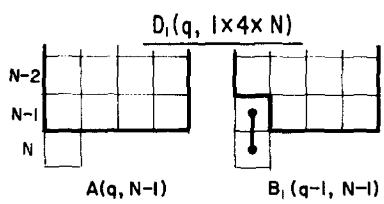


FIG. $4D_1$. Generating a recurrence relation for the number of arrangements $D_1(q,N)$ of q dumbbells on a truncated $1\times 4\times N$-lattice of type D_1.

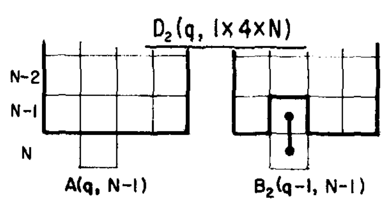


FIG. $4D_2$. Generating a recurrence relation for the number of arrangements $D_2(q,N)$ of q dumbbells on a truncated $1\times 4\times N$-lattice of type D_2.

TABLE V. Arrangements of q dumbbells on a $1\times4\times N$ lattice and the normalization factor Δ_N.

N \ q	0	1	2	3	4	5	6	7	8	9	10	Δ_N EXACT	Δ_N APPROXIMATE
0	1											1	0.472589371
1	1	3	1									5	5.699001236
2	1	10	29	26	5							71	68.72481076
3	1	17	102	267	302	123	11					823	828.7591840
4	1	24	224	1065	2602	3401	2129	509	31			9986	9994.087688
5	1	31	395	2717	10341	25974	36748	29339	11814	1935	78	119373	120519.6765

$$
\begin{aligned}
H(x,y) = {} & x^{24}y^{12} - x^{21}y^{11}(3-x) - x^{18}y^{10}(1+10x+7x^2) + x^{15}y^9(3+13x+8x^2-5x^3) + x^{13}y^8(12+32x+48x^2+19x^3) \\
& - x^{11}y^7(3+19x-5x^2-10x^3) - x^9y^6(24+86x+80x^2+26x^3) - x^7y^5(3+13x+19x^2+10x^3) \\
& + x^5y^4(12+56x+56x^2+19x^3) + x^3y^3(3+19x+14x^2+5x^3) - x^2y^2(1+14x+7x^2) - xy(3+x) + 1.
\end{aligned}
$$

$D(x,y)$ is a polynomial of 14th degree. The z-roots of $D(1,1/z)$ are all real and were found to be

$R_1 = 12.059\,097\,36,\quad R_2 = 3.043\,737\,925,\quad R_3 = 1.330\,009\,783,\quad R_4 = 0.826\,568\,481,$

$R_5 = 0.439\,218\,637,\quad R_6 = 0.328\,543\,398,\quad R_7 = 0.251\,442\,710,\quad R_8 = -0.227\,516\,685,\quad R_9 = -0.484\,178\,764,$

$R_{10} = -0.548\,230\,369,\quad R_{11} = -1,\quad R_{12} = -1.826\,568\,481,\quad R_{13} = -1.858\,555\,021,\quad R_{14} = -3.336\,086\,508.$

Again, the root of highest modulus is positive and we evaluate H and the first and second derivatives of D at $x = 1$ and $y = 1/R_1$

$$
\frac{\partial D(1,1/R_1)}{\partial x} = -1.473\,035\,364,\quad \frac{\partial^2 D(1,1/R_1)}{\partial x^2} = -0.846\,894\,881,
$$

$$
\frac{\partial D(1,1/R_1)}{\partial y} = -13.955\,575\,52,\quad \frac{\partial^2 D(1,1/R_1)}{\partial y^2} = -12.296\,075\,00,
$$

$$
\frac{\partial^2 D(1,1/R_1)}{\partial x\,\partial y} = -18.708\,414\,12,\quad H(1,1/R_1) = 0.546\,911\,310.
$$

These results will then lead to the statistical quantities for $x = 1$

$$\Delta_N = (0.475\,589\,371)(12.059\,097\,36)^N, \tag{6.1a}$$

$$\langle \theta \rangle_\infty = 0.620\,875\,482, \tag{6.1b}$$

$$\sigma_N^2 = 0.074\,714\,541/N. \tag{6.1c}$$

Again, as a check of how good the approximate relation (6.1a) is, we list in Table V the number of arrangements of q dumbbells on a $1\times4\times N$ lattice and compare the exact value of Δ_N with the approximate value computed from (6.1a). Consistent with previous results for $N = 4$ the approximation is accurate to better than 0.1 percent.

VII. SUMMARY AND CONCLUSION

The results obtained for the $1\times M\times N$ lattices with $M = 1, 2, 3$, and 4 are summarized in Table VI. We found for each of these lattices that all roots of $D(1,1/z)$ were simple and real and we also found that the root of largest modulus was positive. This may not be a surprise when all roots are found to be real. However, we are not able to prove that this should always be the case for all higher values of M. A closer analysis of the data shows that it is possible to make certain extrapolations which lead to a number of interesting predictions.

(1) Equation (2.30) shows that, for a fixed set of values L and M the first-order correction of $\langle \theta \rangle_N$ in the large-N limit is of order $1/N$, namely $(2/LMN)(k_1'/k_1)$. It is appropriate to

TABLE VI. Summary of the numerical results obtained for $1\times M\times N$ lattices with $M = 1$, 2, 3, and 4.

M	R_1	k_1	$M\lim_{N\to\infty}[N\sigma_N^2]$	$\langle\theta\rangle_\infty$
1	1.618033989	0.723606798	0.357770876	0.552786405
2	3.214319743	0.664591384	0.334201672	0.606492711
3	6.212070250	0.551273070	0.316984278	0.614764390
4	12.05909736	0.472589371	0.298858164	0.620875482

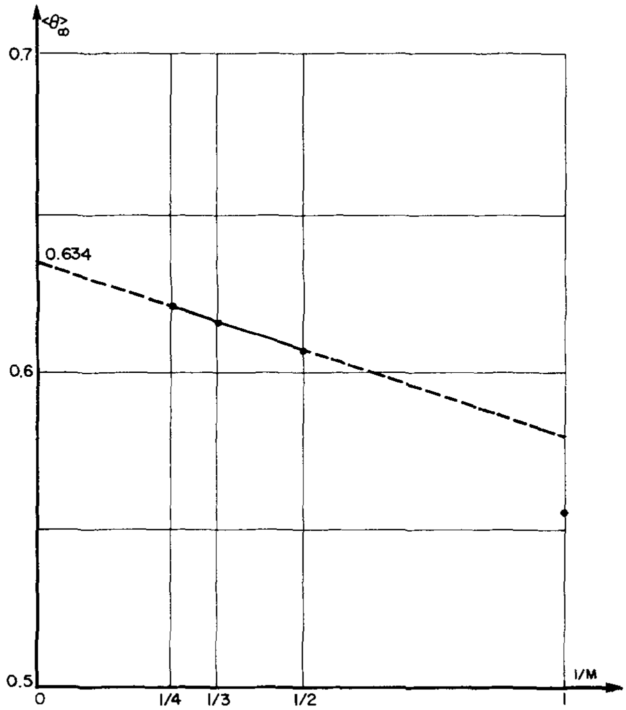


FIG. 5. The large-N limit of the average occupation of dumbbells on a $1 \times M \times N$ lattice vs $(1/M)$.

add two extra labels to the expectation $\langle \theta \rangle_\infty$ referring to L and M. In other words, we really obtained the value of $\langle \theta \rangle_{LM\infty}$ for $L = 1$ and $M = 1, 2, 3$, and 4. Because of the symmetry that exists between L, M, and N, it is obvious that the first-order correction to $\langle \theta \rangle_{LMN}$ in the large-M limit is of order $(1/M)$. It therefore makes sense to plot $\langle \theta \rangle_{1M\infty}$ versus

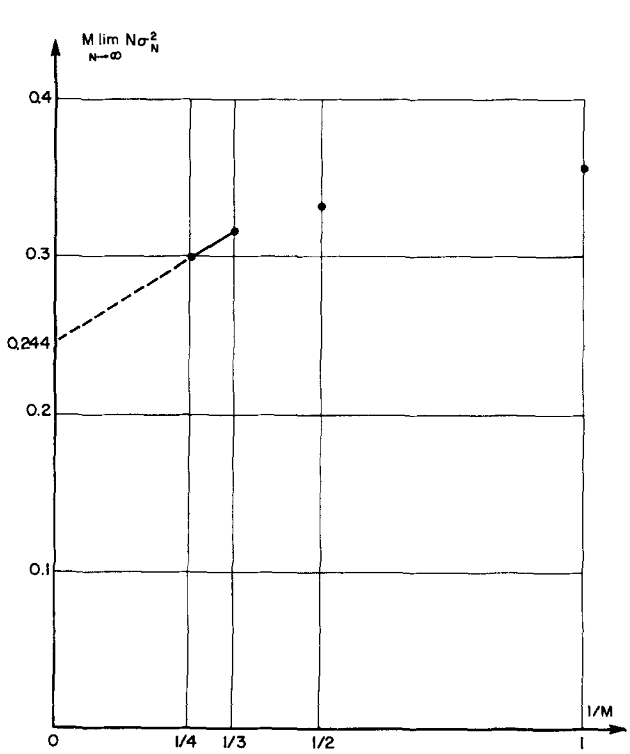


FIG. 6. The large-N limit of $MN\sigma_N^2$ vs $(1/M)$.

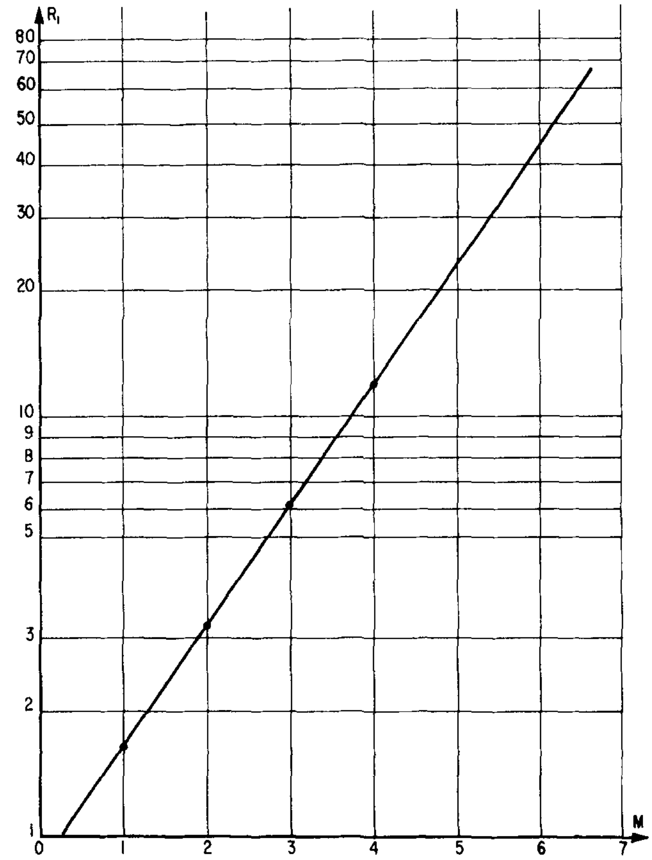


FIG. 7. A semilogarithmic plot of the largest root $R_1(1)$ calculated for $1 \times M \times N$-lattices vs M.

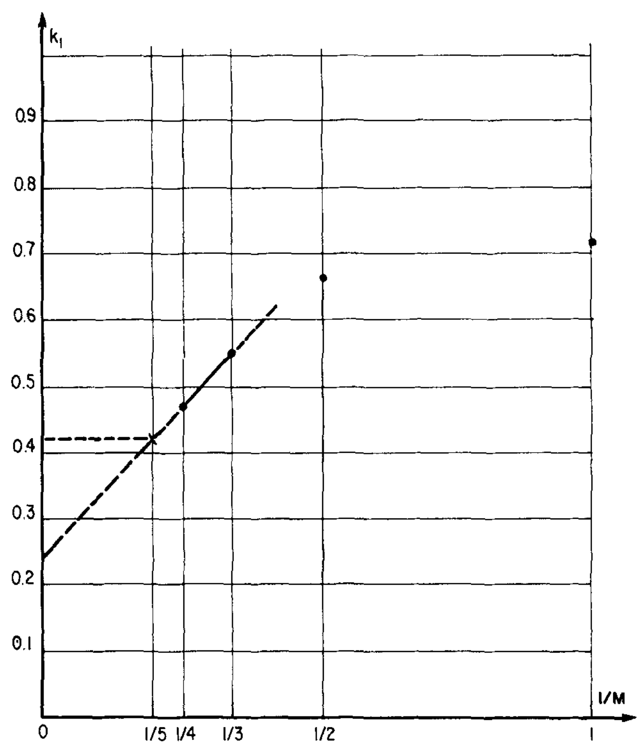


FIG. 8. The large-N limit of Δ_N / R_1^N, or $k_1(1)$, calcualted for $1 \times M \times N$ lattices vs $1/M$.

TABLE VII. Arrangements of q dumbbells on a $1 \times 5 \times N$ lattice and the normalization factor Δ_N.

N \ q	0	1	2	3	4	5	6	7	8	9	10	Δ_N EXACT	Δ_N APPROXIMATE
0	1											1	0.425379
1	1	4	3									7	10.0511
2	1	13	56	94	56	8						228	237.494
3	1	22	185	748	1597	1670	757	106				5086	5611.64
4	1	31	395	2717	10341	25974	36748	29339	11814	1935	78	119373	132595.

$(1/M)$ and follow its behavior as M increases from 1 to 4. Clearly, an almost linear graph is expected in the region of low values of $(1/M)$. Figure 5 confirms our theoretical expectation. Apparently, linearity holds most perfectly for the data points corresponding to $M = 2$, 3, and 4. A least-square fit of these three points gives

$$\langle \theta \rangle_{1M\infty} = -0.0564(1/M) + 0.6344, \quad M \geqslant 2. \tag{7.1}$$

The intercept gives the average occupation of dumbbells on a planar lattice of infinite extent, namely

$$\langle \theta \rangle_{1\infty\infty} = 63.4\% \simeq 1 - e^{-1}. \tag{7.2}$$

(2) Equation (2.40) shows that, for a fixed set of values L and M the first-order correction to $N\sigma_N^2$ in the large-N limit is also of the order of $(1/N)$, namely $(4/NL^2M^2)(k_1'/k_1) \times [1 + (k_1''/k_1')(k_1'/k_1)]$. Here again, it is appropriate to make explicit the L and M dependence and add these two labels to σ_N^2, thus giving σ_{LMN}^2. The symmetry mentioned above also requires that the first-order correction to

$$M\{\lim_{N\to\infty} N\sigma_{LMN}^2\} \tag{7.3}$$

is of order $(1/M)$. Therefore, it makes sense to plot this quantity versus $(1/M)$ (Fig. 6). Linearity is not achieved as quickly as was the case for $\langle \theta \rangle_{1M\infty}$. However, it is possible to obtain an upper bound on the quantity

$$\lim_{\substack{N\to\infty \\ M\to\infty}} MN\sigma_{1MN}^2. \tag{7.4}$$

A linear extrapolation from the last two data points, corresponding to $M = 3$ and $M = 4$, gives an intercept of 0.244. This is the upper bound we are looking for

$$\lim_{\substack{N\to\infty \\ M\to\infty}} MN\sigma_{1MN}^2 < 0.244. \tag{7.5}$$

(3) The largest root $R_1(1)$ seems to follow an exponential behavior in M. This is very obvious from the semilogarithmic graph (Fig. 7). A least-square fit of the four data points gives

$$R_1(1;M) = (0.8354)(1.9512)^M \tag{7.6}$$

and predicts a largest root of 23.6286 for $M = 5$. It also predicts that in the thermodynamic limit $\Xi(1) = \lim_{M\to\infty} [R_1(1;M)]^{1/M} = 1.9512$. Thus the error made in calculating $\Xi(1)$ for a two-dimensional lattice of infinite extent by just taking $M = 4$ is 4.5% since $[R_1(1;M=4)]^{1/4} = (12.0591)^{1/4} = 1.863$.

(4) Concerning $k_1(1)$, which represents the limit of $[\Delta_N/R_1^N]$, as N becomes infinitely large, Fig. 8 makes apparent its decreasing behavior in terms of increasing values of M. Here, with no theoretical motivation, we plot for convenience $k_1(1)$ versus $1/M$. A linear extrapolation from the last two data points corresponding to $M = 3$ and $M = 4$ gives the value

$$k_1(1;M=5) = 0.425\,379. \tag{7.7}$$

We, therefore, obtain an approximate expression for the normalization Δ_N for the $1 \times 5 \times N$-lattice problem, namely

$$\Delta_N \sim 0.425\,379(23.628\,6)^N, \quad M = 5. \tag{7.8}$$

This formula, obtained by extrapolation, works well for low values of N, as one might expect. Indeed, Δ_N is an exponential function; a small error in determining the largest root R_1 would produce a large error in computing Δ_N as N becomes large. The root R_1 should be calculated to many more significant figures than the six given by the linear fit of the data. Table VII shows how the values of Δ_N computed from (7.8) start to diverge rapidly from the exact results. In this case, the estimate on the root R_1 is a little bit too large. Nevertheless, Eq. (7.6) may be used effectively should one choose to solve explicitly the occupation statistics on higher-order planar lattices.

ACKNOWLEDGMENTS

The author would like to acknowledge useful conversations on various aspects of this problem with Dr. John D. Bruno and Dr. Jan H. Busch.

[1] D. Lichtman and R. B. McQuistan, J. Math. Phys. **8**, 2441 (1967); R. B. McQuistan, Nuovo Cimento B **58**, 86 (1968); R. B. McQuistan and S. J. Lichtman, J. Math. Phys. **11**, 3095 (1970); R. B. McQuistan and J. L. Hock, Fibonacci Quart. **21**, 196 (August 1983); J. L. Hock and R. B. McQuistan, J. Math. Phys. **24**, 1859 (1983).

[2] P. W. Kasteleyn, Physica **27**, 1209 (1961); H. N. V. Temperly and M. E. Fisher, Philos. Mag. **6**, 1061 (1961).

[3] E. H. Lieb, J. Math. Phys. **8**, 2339 (1967).

[4] See, for example, D. S. Gaunt, Phys. Rev. **179**, 174 (1969).

[5] A. J. Phares, D. E. Shaw, and F. J. Wunderlich, "Closed form analytical solutions of the monomer–timer problem on a planar rectangular lattice. II," J. Math. Phys. (to be published).

[6] A. J. Phares, "Decoupling of a System of Partial Difference Equations With Constant Coefficients," J. Math. Phys. **25**, 2167 (1984).

Weak solutions of a quasistatic model of plasmas

Zensho Yoshida, Nobuyuki Inoue, and Taijiro Uchida
Department of Nuclear Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan

(Received 30 November 1983; accepted for publication 27 January 1984)

We study an analytical structure of a quasistatic model of magnetically confined plasmas. Applying the fixed point theorem, we construct global-in-time weak solutions.

PACS numbers: 02.90. + p, 03.40.Gc, 52.30. + r, 52.55. − s

I. INTRODUCTION

This paper studies equations that model the quasistatic evolution of a plasma. The quasistatics is the study of evolution of equilibrium driven by the slow change of internal and external parameters. The quasistatic equations are given by dropping $\rho(\partial_t \mathbf{v} + (\mathbf{v}\cdot\nabla)\mathbf{v})$ (ρ is the mass density, $\mathbf{v}$ is the velocity) in the magnetohydrodynamic equations.[1] There exist systematic studies[2–6] on this subject: formulating the problem, constructing explicit numercial schemes, and analyzing mathematical structures. In this paper, however, we study a simplified model in order to understand an analytical structure of the problem from the viewpoint of partial-differential-equation theory.

Our goal in this paper is to show that the model equations have global-in-time solutions in a weak sense. Here, roughly speaking, a weak solution is a solution without regularity and uniqueness (see Sec. III). In many quasistatic problems, we see some pathological behavior of solutions such as bifurcations and discontinuous evolution, that is, the so-called catastrophic jump. Therefore, it seems a natural approach that we discuss weak solutions which include a quite general class of solutions enough to permit bifurcations and singularities.

We consider initial-boundary value problems of a system

$$\partial_t u = \Delta u + f(t) + g, \quad u - P(u)u = 0,$$

where the unknown u corresponds to the current density flowing in the plasma, $f(t)$ is a given function, g is an unknown function implicit to u and $f(t)$, and $P(u)$ is an orthogonal projector onto a closed convex set $W(u)$ of functions, which is a linear subspace dependent upon u.

Let us briefly review some analogous problems from a mathematical point of view. If the linear subspace W is independent of u, the corresponding problem is linear. A simple example of this category is the Stokes equations[7]:

$$\partial_t \mathbf{v} = \Delta \mathbf{v} + \mathbf{f}(t) - \nabla p,$$

$$\mathbf{v} - P_\sigma \mathbf{v} = 0,$$

where P_σ is the projector onto a linear subspace W_σ of divergence-free vector fields, $\mathbf{v}$ is the fluid velocity, and p is the pressure.

When the fixed set W is a general closed convex set, the corresponding problem is nonlinear and it is shown to be equivalent to a penalized equation[8]:

$$\partial_t u \in -\partial\Psi(u) + f(t),$$

$$\Psi(u) = \begin{cases} (\nabla u, \nabla u)/2, & \text{if } u \in W \cap H_0^1(\Omega) \\ +\infty, & \text{otherwise,} \end{cases}$$

$$\partial\Psi(u) = \{\xi\,;\, \Psi(v) \geqslant \Psi(u) + (\xi, v - u), \text{ for all } v \in L^2(\Omega)\},$$

where (u,v) denote the inner product of a Hilbert space $L^2(\Omega)$; see Sec. III for notation of function spaces. Making appropriate assumptions for $f(t)$, we get a unique regular global-in-time solution for this problem (see Brézis[8]).

Allowing time dependence for W gives a generalization of the above penalized problem. This temporally inhomogeneous equation has a unique regular global-in-time solution under appropriate assumptions for $f(t)$ and $W(t)$ (see Watanabe[9]).

In the present work, the convex set $W(u)$ is a linear subspace, but it depends upon u, so the problem is strongly nonlinear. In Sec. II, we discuss the physical background of the problem, and in Sec. III, we give a mathematical formulation. In Sec. IV, we prove the existence of weak solutions applying the Schauder fixed point theorem (cf. Nirenberg[10]).

II. A QUASISTATIC MODEL

The diffusive evolution of the magnetic flux density $\mathbf{B}$ is dominated by Ohm's law[1]

$$\mathbf{E} = \eta\mathbf{j} - \mathbf{v}\times\mathbf{B},$$

where $\mathbf{E}$ is the electric field, $\mathbf{j}$ is the current density, $\mathbf{v}$ is the fluid velocity, and η is the resistivity. In this work, we assume η is a constant. When we combine Ohm's law with Faraday's law

$$\partial_t \mathbf{B} = -\nabla\times\mathbf{E},$$

we get

$$\partial_t \mathbf{B} = (\eta/\mu_0)\Delta\,\mathbf{B} + \nabla\times(\mathbf{v}\times\mathbf{B}), \tag{1}$$

where we use Ampere's law with the displacement current neglected:

$$\mathbf{j} = \nabla\times\mathbf{B}/\mu_0.$$

Take the rotation of the both sides of (1) to get an evolution equation for $\mathbf{j}$:

$$\partial_t\, \mathbf{j} = (\eta/\mu_0)\Delta\mathbf{j} + \nabla\times(\nabla\times(\mathbf{v}\times\mathbf{B})). \tag{2}$$

In the quasistatic model, the nonlinear term $\nabla\times(\nabla\times(\mathbf{v}\times\mathbf{B}))$ is implicitly determined by the equilibrium condition, instead of solving an evolution equation for $\mathbf{v}$.[11] The macroscopic equilibrium condition is

$$0 = \mathbf{j}\times\mathbf{B} - \nabla p, \qquad (3)$$

where p is the pressure. To determine p, we need equations of mass and entropy, which make analysis very complicated. Therefore, instead of using (3) directly, we employ here the following reduced condition of equilibrium. Let us now consider a two-dimensional system for simplicity. A generalization to three-dimensional systems is discussed in Sec. V. The reader is referred to Grad and Rubin[12] for the following standard technique in the static theory of plasmas. Let z denote the ignorable coordinate. We denote by u the z-component of the current density $\mathbf{j}$. In a two-dimensional system, we can introduce the flux function ψ that is the z component of the vector potential; so it is related to u by a potential equation. The flux function defines flux surfaces, viz., a surface $\nabla z\times(\psi$-contour) is a flux surface which magnetic field lines are lying on. According to the equilibrium theory,[12] in a two-dimensional system, u is a function of only ψ, viz., $u = u(\psi)$. In other words, u should be constant on each ψ-contour. This *necessary* condition[13] for equilibrium is formally written as follows (see Sec. III for a more mathematical formulation). Let $W(u)$ denote a set of functions that are constant on ψ-contours, where ψ is related to u by a potential equation (see Sec. III). We denote by $P(u)$ the orthogonal projector onto $W(u)$. If and only if u is contained in $W(u)$, we have

$$u - P(u)u = 0\,. \qquad (4)$$

This equation (4) is the formal expression of the above-mentioned equilibrium condition.

In this work, we model the quasistatics using the reduced condition (4). Let us write the z-component of Eq. (2) as

$$\partial_t\, u = \Delta u + N, \qquad (5)$$

where N is the z component of the nonlinear term $\nabla\times(\nabla\times(\mathbf{v}\times\mathbf{B}))$, and we replaced the constant (η/μ_0) by unity to simplify notation. To close (5) by the necessary condition of equilibrium (4), we write $N = f(t) + g$, where $f(t)$ is a given function and g is implicit to u and $f(t)$. Note that, when we use the necessary and sufficient condition of equilibrium,[13] we can consider N itself implicit to u, and $f(t) = 0$.

We now have a reduced quasistatic model:

$$\partial_t u = \Delta u + f(t) + g, \qquad (6)$$
$$u - P(u)u = 0,$$

where unknowns are u and g. We consider the above system in a bounded domain Ω that is contained in $\mathbb{R}^2$, with a boundary condition $u = 0$ on the boundary $\partial\Omega$. In the next section, we give a mathematical formulation of $W(u)$, $P(u)$, and the system (6).

III. MATHEMATICAL FORMULATION

First, we list notation used in this paper. Let Ω be a smoothly bounded domain in $\mathbb{R}^2$. We denote by Q_T a cylindrical region $(0,T)\times\Omega$, and by Q'_T a region $(-\alpha,T)\times\Omega$ for a fixed positive number α. We use the following function spaces (cf. Lions and Magenes[14]). We denote by $L^2(\Omega)$ the space of functions $u(x)$ which are measurable and such that the norm $\|u\| = (\int_\Omega u^2\,dx)^{1/2}$ is finite. It is well known that $L^2(\Omega)$ is a Hilbert space for the scalar product

$$(u,v) = \int_\Omega u(x)v(x)dx.$$

Let $H^m(\Omega)$ denote the Sobolev space that is a Hilbert space of functions which are in $L^2(\Omega)$ together with all their derivatives of order $\leqslant m$. Let $H^{-m}(\Omega)$ be the dual space of $H^m(\Omega)$. We denote by $H^m_0(\Omega)$ the closure of $\mathscr{D}(\Omega)$ in $H^m(\Omega)$, where $\mathscr{D}(\Omega)$ is the set of infinitely differentiable functions with compact support in Ω. We define $\mathscr{D}'(\Omega) = \textit{dual of}\ \mathscr{D}(\Omega)$, which is the space of distributions on Ω. We denote by $C(\Omega)$ the space of continuous functions on Ω.

Let $L^2(0,T;V)$, for a Hilbert space V, denote a space of functions on $(0,T)$ such that the norm $(\int_0^T|u|_V^2\,dt)^{1/2}$ is finite. We denote by $((u,v))_T$ the inner product of the Hilbert space $L^2(0,T;L^2(\Omega))$, viz.,

$$((u,v))_T = \int_0^T\left(\int_\Omega u(x,t)v(x,t)dx\right)dt.$$

Let $C([0,T];V)$ denote the space of continuous functions on $[0,T]$ with values in a Hilbert space V.

We now give the mathematical expression of the flux function ψ. The flux function is related to $u\in L^2(\Omega)$, in Ω, by

$$\psi = Ku,$$

where K is a linear compact operator on $L^2(\Omega)$ such that K maps $L^2(\Omega)$ into $H^2(\Omega)$. In view of the Sobolev lemma,[15] $\psi = Ku\in C(\Omega)$ for $u\in L^2(\Omega)$. Explicit definition of K is the solution operator for the following linear elliptic problem

$$-A\psi = u,$$

where A is the Laplacian with the domain $D(A) = H^1_0(\Omega_0)\cap H^2(\Omega_0)$, where Ω_0 is a certain two-dimensional domain including Ω.

Let $W(u)$, for $u\in L^2(\Omega)$, be the closure in $L^2(\Omega)$ of smooth functions that are constant on each $\psi = $ const contour, where $\psi = Ku$. The set $W(u)$ is a subspace of $L^2(\Omega)$. We denote by $P(u)$ the orthogonal projector onto $W(u)$. We define $P(u(t))$ for $u(t)\in L^2(0,T;L^2(\Omega))$ taking, for example, $P(u(t_s))v = v$ at the discontinuous point t_s of $u(t)$.[16]

Let us interpret (6) as an evolution equation in $L^2(\Omega)$:

$$\frac{du}{dt} = \Delta u + f(t) + g, \qquad (7)$$
$$u - P(u)u = 0,$$
$$u(0) = u_0,$$

where Δ is the Laplacian with domain

$$D(\Delta) = H^1_0(\Omega)\cap H^2(\Omega),$$

du/dt is the strong derivative of u, $f(t)$ is a given function in $C([0,\infty);L^2(\Omega))$, and u_0 is an initial value in $L^2(\Omega)$.

Next, we define *a weak solution* of the system (7). We call a function $u(x,t)$ a weak solution of (7) on $(0,T)$, if there exists $g\in\mathscr{D}'(Q'_T)$ and u satisfies

$$-((u,\partial_t\phi))_T = ((u,\Delta\phi))_T + ((f,\phi))_T + \langle\langle g,\phi\rangle\rangle_T + (u_0,\phi(\cdot,0)), \qquad (8)$$
$$((u - P(u)u,\phi))_T = 0,$$

for all $\phi\in\mathscr{D}(Q'_T)$, where $\langle\langle g,\cdot\rangle\rangle_T$ represents a continuous linear functional on $\mathscr{D}(Q'_T)$. Formally, we get (8) when we multiply $\phi \in \mathscr{D}(Q'_T)$ to Eqs. (7), and integrate them over Q_T.

IV. WEAK SOLUTIONS

We will prove the existence of a weak solution of (7). The existence theorem is an application of the Schauder fixed point theorem. We first fix a function w and solve a preliminary problem:

$$-((u,\partial_t \phi))_T = ((u,\Delta\phi))_T + ((f,\phi))_T + \langle\langle g,\phi\rangle\rangle_T + (u_0, \phi(\cdot,0)), \tag{9}$$
$$((u - P(w)u, \phi))_T = 0,$$

for all $\phi \in \mathscr{D}(Q'_T)$, where $P(u)$ in (8) is here replaced by $P(w)$. Let $U(\cdot)$ denote a mapping that maps w to u, the solution of (9), viz., $U(w) = u$. We then find a fixed point of the mapping $U(\cdot)$. Let us begin with the following lemma.

Lemma 1: We fix a function $w(t) \in L^2(0,\infty;L^2(\Omega))$, and consider a parabolic equation

$$\frac{du}{dt} = \Delta u + f(t) + \frac{[P(w(t))u - u]}{\epsilon}, \tag{10}$$
$$u(0) = u_0,$$

where ϵ is an arbitrary fixed positive number, $f(t) \in C([0,\infty);L^2(\Omega))$, and $u_0 \in L^2(\Omega)$. Then, (10) admits a unique solution in

$$V(0,\infty) = \Big\{u\,;\, u \in L^2(0,\infty;H_0^1(\Omega)),\ \frac{du}{dt} \in L^2(0,\infty;H^{-1}(\Omega))\Big\}.$$

Proof: First, we recall that $V(0,\infty) \subset C([0,\infty);L^2(\Omega))$ (see Lions and Magenes[14]).

(*First step.*) To prove this lemma, we apply the fixed point theorem in $C([0,\infty);L^2(\Omega))$. We first fix a function $u \in C([0,\infty);L^2(\Omega))$, and consider the following problem:

$$\frac{dy}{dt} = \Delta y + F^u(t), \tag{11}$$
$$F^u(t) = f(t) + [P(w(t))u(t) - u(t)]/\epsilon.$$

It is well known that (11) has a unique solution $y \in V(0,\infty)$ (see Lions[17]), and $y(t)$ can be written

$$y(t) = e^{t\Delta}u_0 + \int_0^t e^{(t-s)\Delta}F^u(s)ds, \tag{12}$$

where $e^{t\Delta}$ is the analytical semigroup[15] generated by Δ. Let us write $y = Y(u)$, where y is a solution of (11) with a fixed function $u \in C([0,\infty);L^2(\Omega))$. We next find a fixed point of $Y(\cdot)$.

(*Second step.*) We begin with the existence of a local-in-time solution. We define

$$G(0,t) = \{u \in C([0,t];L^2(\Omega));\ \|u(s)\| \leqslant \|u_0\| + 1\ \ (t \geqslant s \geqslant 0)\},$$

which is a closed bounded subset of $C([0,t];L^2(\Omega))$. We prove that $Y(\cdot)$ maps $G(0,t)$ into itself for sufficiently small t. In view of (12), we have

$$\|Y(u)_{(t)}\| \leqslant \|e^{t\Delta}u_0\| + \int_0^t \|e^{(t-s)\Delta}F^u(s)\|ds,$$

where $Y(u)_{(t)}$ is the value of $Y(u)$ at t. Since $e^{t\Delta}$ is a contraction mapping for all $t \geqslant 0$, we have

$$\|Y(u)_{(t)}\| \leqslant \|u_0\| + t \operatorname*{ess\,sup}_{0<s<t} \|F^u(s)\|.$$

Because the projector $P(w(t))$ is contractive, we get

$$\|Y(u)_{(t)}\| \leqslant \|u_0\| + t\Big[\sup_{0<s<t}(\|f(s)\| + 2\|u(s)\|/\epsilon)\Big]$$
$$\leqslant \|u_0\| + t\,[M + 2(\|u_0\| + 1)/\epsilon],$$

where $M = \sup_{0<s<\infty} \|f(s)\|$. Therefore, for $t \leqslant [M + 2(\|u_0\| + 1)/\epsilon]^{-1}$, we conclude that $T(\cdot)$ maps $G(0,t)$ into $G(0,t)$. Next, we prove that $Y(\cdot)$ is a contraction mapping on $G(0,t)$, for sufficiently small t. Using (12), we estimate, for $u, v \in G(0,t)$,

$$\|Y(u)_{(t)} - Y(v)_{(t)}\| \leqslant \int_0^t \|e^{(t-s)\Delta}[F^u(s) - F^v(s)]\|ds$$
$$\leqslant t(2/\epsilon) \sup_{0<s<t} \|u(s) - v(s)\|.$$

This implies

$$\sup_{0<s<t} \|Y(u)_{(s)} - Y(v)_{(s)}\| \leqslant \sup_{0<s<t} \|u(s) - v(s)\|,$$

for $t \leqslant \epsilon/2$. So, for

$$t \leqslant T_0 = \min\{[M + (2/\epsilon)(\|u_0\| + 1)]^{-1}, \epsilon/2\}$$
$$= [M + (2/\epsilon)(\|u_0\| + 1)]^{-1},$$

$Y(\cdot)$ is a contraction mapping on $G(0,t)$. Therefore $Y(\cdot)$, for $t \leqslant T_0$, has a unique fixed point u in $G(0,T_0)$, viz., $Y(u) = u$, which implies that u solves (10) in $t \leqslant T_0$.

(*Third step.*) Finally, we prove the existence of a global-in-time solution. We construct the global solution by the continuation of local solutions. We have found, in the second step, a solution $u(t)$ in $[0,T_0]$, which satisfies $\|u(T_0)\| \leqslant \|u_0\| + 1$. Considering $u(T_0) = u_1$, the new initial value, we apply the same procedure of the second step to find a solution in $[0,T_0 + T_1]$, where

$$T_1 = [M + (2/\epsilon)(\|u_1\| + 1)]^{-1}$$
$$\geqslant [M + (2/\epsilon)(\|u_0\| + 2)]^{-1}.$$

This continuation of the local solutions gives the desired global solution, since $\Sigma_{i=0}^{\infty} T_i = \infty$. This completes the proof of Lemma 1.

We are planning to pass the limit of $\epsilon \to 0$ in Eq. (10). This technique is the so-called penalty method (cf. Lions[18]). We prepare the following *a priori* estimate that is independent of ϵ.

Lemma 2: For the solution of (10), we have an *a priori* estimate

$$\frac{d}{2\,dt}\|u(t)\|^2 \leqslant -\|\nabla u(t)\|^2 + (f(t),u(t)), \tag{13}$$

for all $t > 0$. This especially implies

$$\int_0^t \|\nabla u(s)\|^2\,ds \leqslant \tfrac{1}{2}\|u_0\|^2 + M^2t^2 + Mt\|u_0\|, \tag{14}$$

where $M = \sup_{0<s<\infty} \|f(s)\|$.

Proof: When we multiply the solution $u(t)$ of (10) to the both sides of (10), and integrate them over Ω, we get

$$\frac{d}{2dt}\|u(t)\|^2 = -\|\nabla u(t)\|^2 + (f(t),u(t)) + (P(w(t))u(t) - u(t),u(t))/\epsilon.$$

This immediately gives (13), since $P(w(t))$ is contractive. Integrating (13) over $(0,t)$ gives (14).

We now prove the existence of a solution to a modified problem (9).

Lemma 3: Let T be an arbitrary positive number, and let $w(t)$ be given in $L^2(0,T;L^2(\Omega))$. Then, there exist $u \in L^2(0,T;H^1_0(\Omega))$ and $g \in \mathscr{D}'(Q'_T)$ that satisfy (9):

$$-((u,\partial_t \phi))_T = ((u,\Delta\phi))_T + ((f,\phi))_T + \langle\langle g,\phi\rangle\rangle_T + (u_0,\phi(\cdot,0)),$$

$$((u - P(w)u,\phi))_T = 0,$$

for every $\phi \in \mathscr{D}(Q'_T)$, where the initial value u_0 is in $L^2(\Omega)$.

Proof: In what follows, we denote by u^ϵ the solution of (10), in order to clarify that u^ϵ is a solution for the parameter ϵ. For any $\phi \in \mathscr{D}(Q'_T)$, we have

$$-((u^\epsilon,\partial_t \phi))_T = ((u^\epsilon,\Delta\phi))_T + ((f,\phi))_T + ((\{[P(w)u^\epsilon - u^\epsilon]/\epsilon\},\phi))_T + (u_0,\phi(\cdot,0)).$$

In view of the estimate (14), we can select a subsequence $\{u^\delta\}$ of the sequence $\{u^\epsilon\}$ that converges to u in the weak topology of $L^2(0,T;H^1_0(\Omega))$. For such a subsequence $\{u^\delta\}$, we have, as $\delta \to 0$,

$$((u^\delta,\partial_t \phi))_T \to ((u,\partial_t\phi))_T,$$

$$((u^\delta,\Delta\phi))_T \to ((u,\Delta\phi))_T,$$

for every $\phi \in \mathscr{D}(Q'_T)$. Therefore $\{((u^\delta - P(w)u^\delta,\phi))_T/\delta\}$ converges to a finite number $g(\phi)$. On the other hand, using $((P(w)u,\phi))_T = ((u,P(w)\phi))_T$, we get $((u^\delta - P(w)u^\delta,\phi))_T \to ((u - P(w)u,\phi))_T$, as $\delta \to 0$. Consequently, we have $((u - P(w)u,\phi))_T = 0$. Writing $g(\phi) = \langle\langle g,\phi\rangle\rangle$, we get the desired result.

Finally, we prove the existence of a weak solution of (7).

Theorem: Equation (7) permits a global-in-time weak solution, viz., there exists at least one couple of $u \in L^2(0,T;H^1_0(\Omega))$ and $g \in \mathscr{D}'(Q'_T)$ that satisfies, for every $T>0$,

$$-((u,\partial_t \phi))_T = ((u,\Delta\phi))_T + ((f,\phi))_T + \langle\langle g,\phi\rangle\rangle_T + (u_0,\phi(\cdot,0)),$$

$$((u - P(u)u,\phi))_T = 0,$$

for all $\phi \in \mathscr{D}(Q'_T)$.

Proof: Let $U(w) = u$, where u is the solution, given in Lemma 3, of the problem (9) with a fixed function w. In view of the lower semicontinuity of the norm $(\int_0^T \|\nabla u(s)\|^2\,ds)^{1/2}$ in the weak topology of $L^2(0,T;H^1_0(\Omega))$, we have the same estimate (14) also for the solution $u(t)$ of (9). Let us define

$$H(0,T) = \left\{ u \in L^2(0,T;H^1_0(\Omega)); \int_0^T \|\nabla u\|^2 ds \leqslant \frac{1}{2}\|u_0\|^2 + M^2T^2 + MT\|u_0\| \right\},$$

which is a weakly closed, bounded, convex subset of $L^2(0,T;H^1_0(\Omega))$. Because of the above-mentioned estimate, the function $U(\cdot)$ maps $H(0,T)$ into $H(0,T)$, and it is compact in the weak topology of $L^2(0,T;H^1_0(\Omega))$. Applying the Schauder fixed point theorem,[10] we conclude that $U(\cdot)$ has a fixed point in the weak topology. This completes the proof.

V. DISCUSSION

Below we list some additional remarks.

Remark 1: There are four points to be made.

(i) The regularity of the weak solution can be slightly improved as follows, which is obvious from the proof of Lemma 3. The weak solution u of (7) satisfies

$$-((u,\partial_t v))_T = -((\nabla u,\nabla v))_T + ((f,v))_T + ((g,v))_T + (u_0,v(\cdot,0)),$$

$$((u - P(u)u,v))_T = 0,$$

for every $v \in H^1(0,T;H^1_0(\Omega))$, and g is in the dual space of $H^1(0,T;H^1_0(\Omega))$.

(ii) A weak solution may have discontinuity in time.

(iii) The concept of a weak solution has been introduced to the theory of the Navier–Stokes system by Leray[19] (see Temam[7]).

(iv) The projector $P(u)$ gives the flux surface average of functions. In the present work, however, we have used no specific property of the flux surface average besides the fact that $P(u)$ is linear and contractive in $L^2(\Omega)$. Behavior of $P(u)$ is unfavorable for the temporal regularity. For example, in (11) of Lemma 1, we get no improvement in regularity for its solution, even if we assume higher regularity for $w(t)$.

Remark 2 (free boundary): A solution of the quasistatic model has a free boundary in the following sense. Let Ω_t^+ denote the support of the solution $u(t)$ at time t. The region Ω_t^+ is defined almost everywhere in $[0,\infty)$. Generally, Ω_t^+ is strictly contained in Ω, and the plasma has a free boundary $\partial\Omega_t^+$; this is obvious when we consider the definition of $W(u)$ and the boundary condition $u(t) \in H^1_0(\Omega)$ (a.e., $t \geqslant 0$).

Remark 3 (three-dimensional plasmas): Let us consider a three-dimensional domain Ω that is surrounded by a simple torus $\partial\Omega$. We assume that the plasma has flux surfaces everywhere in Ω, viz., every magnetic field line lies on a corresponding surface and twists the long way around the torus. Then, we can introduce a flux coordinate (ψ,θ,ζ)[1] such that θ is an angle the short way around, ζ is an angle the long way around, and $\psi = \int_\Gamma \mathbf{A}\cdot d\mathbf{x}$, where $\mathbf{A}$ is the vector potential and the integral is taken over a curve Γ that is the intersection of the flux surface and a poloidal cut.

A necessary condition for equilibrium is that the current density $\mathbf{j}$ satisfies

$$\mathbf{j} = j^\theta(\psi)\nabla\zeta \times \nabla\psi + j^\zeta(\psi)\nabla\psi\times\nabla\theta.$$

Let $\mathbf{W}(\mathbf{j})$ be the closure in $(L^2(\Omega))^3$ of smooth functions that satisfy the above condition and let $P(\mathbf{j})$ denote the projector onto $\mathbf{W}(\mathbf{j})$. We consider the following quasistatic model:

$$\partial_t \mathbf{j} = \Delta\mathbf{j} + \mathbf{f}(t) + \mathbf{g},$$

$$\mathbf{j} - P(\mathbf{j})\mathbf{j} = \mathbf{0}.$$

Then, we conclude that the above system has a weak solution provided that the plasma has flux surfaces everywhere; the argument is similar to that of the two-dimensional problem.

ACKNOWLEDGMENTS

The authors are grateful to Professor H. Grad for valuable comments and encouragement. The authors are also grateful to Professor Y. Giga and Professor Y. Konishi for valuable discussions and suggestions. Thanks are also due to the staffs of the Magneto-Fluid Dynamics Division of the Courant Institute.

[1]G. Bateman, *MHD Instabilities* (MIT Press, Cambridge, MA, 1978).

[2]H. Grad, Proc. Congr. Math., Nice **3**, 105 (1970).

[3]H. Grad and J. Hogan, Phys. Rev. Lett. **24**, 1377 (1970).

[4]H. Grad, P. N. Hu, and D. C. Stevens, Proc. Nat. Acad. Sci. USA **72**, 3789 (1975).

[5]G. Vigufsson, Bull. Am. Math. Soc. **85**, 778 (1979).

[6]R. Temam, in *Proceedings of the International Meeting on Recent Methods in Nonlinear Analysis*, Rome 1978 (Pitagora Editrice, Bologana, 1979), p. 83.

[7]R. Temam, *Navier–Stokes Equations* (North-Holland, Amsterdam, 1977).

[8]H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert* (North-Holland, Amsterdam, 1973).

[9]J. Watanabe, in *Functional Analysis and Numerical Analysis, Japan–France Seminar*, Tokyo and Kyoto (Japan Soc. for Promotion of Science, Tokyo, 1978).

[10]L. Nirenberg, *Topics in Nonlinear Functional Analysis* (Courant Institute of Mathematical Sciences, N. Y. U., New York, 1974).

[11]Z. Yoshida and Y. Giga, J. Math. Phys. **24**, 2860 (1983).

[12]H. Grad and H. Rubin, in Proc. 2nd U. N. Int. Conf. Peaceful Uses of Energy **31**, 190 (1958).

[13]The necessary and sufficient condition for equilibrium is the Grad–Shafranov equation[1,12]; $u = u(\psi) = dp(\psi)/d\psi + dI^2(\psi)/d\psi$, where $p(\psi)$ and $I(\psi)$ denote the pressure and the current flux, respectively.

[14]J-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications* (Dunod, Paris, 1968), Vol. 1.

[15]K. Yosida, *Functional Analysis* (Springer-Verlag, Berlin, 1978).

[16]The measure of the discontinuous points $\{t_s\}$ is zero. For a weak solution the value of $P(u(t_s))$ is not essential.

[17]J-L. Lions, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles* (Dunod, Paris, 1968).

[18]J-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires* (Dunod, Paris, 1969).

[19]J. Leray, Acta Math. **63**, 193 (1934).

First-order equivalent Lagrangians and conservation laws

Sergio Hojman and Javier Gómez
Centro de Estudios Nucleares, Universidad Nacional Autónoma de México, Circuito Exterior, C.U., 04510 México, D.F., Mexico

(Received 23 May 1983; accepted for publication 27 December 1983)

We present a theorem for (first-order) Lagrangian theories which associates several conserved quantities to *one* (s-equivalence) symmetry transformation.

PACS numbers: 03.20. + i

I. INTRODUCTION

In 1918, Noether established the well-known theorem, which bears her name, relating conserved quantities to invariances of the action principle.[1,2] In this theorem *each* symmetry of the action is associated with *one* conserved quantity, in a very well-defined fashion, which makes it useful to construct the conserved quantities when a Lagrangian and one of its symmetries are known for a given problem. It is, of course, unnecessary to emphasize the tremendous importance that this theorem has had for mathematical physics over the years.

In this work we present a similar theorem, which is based on a different definition of symmetry from the one stated above, that associates *many* conserved quantities with *one* given symmetry of the problem in consideration. The theorem may also be of use since the construction of the constants of the motion is simple and explicit once a Lagrangian and a symmetry of the problem are known.

In order to be more precise, let us briefly review the basic concepts involved in the Noether theorem and present the new elements which appear in the theorem we prove.

Consider a Lagrangian $L(q^i, \dot{q}^i, t)$ for a system with N degrees of freedom. The coordinate transformation

$$\bar{q}^i = \bar{q}^i(q,t), \qquad \bar{t} = \bar{t}(q,t) \tag{1}$$

induces a new Lagrangian $\bar{L}(\bar{q}^i, \dot{\bar{q}}^i, \bar{t})$, where

$$\bar{L}(\bar{q}, \dot{\bar{q}}, \bar{t}) = L(q, \dot{q}, t)\frac{dt}{d\bar{t}}. \tag{2}$$

Noether's theorem essentially asserts that if the transformation given by Eq. (1) is a symmetry transformation for the Lagrangian (in the sense that the equations of motion derived from $\bar{L}$ are the *same* as the ones derived from L), i.e., if

$$\bar{L} = L + \frac{df(q,t)}{dt}, \tag{3}$$

then one can construct a constant of the motion associated with the symmetry. Strictly speaking Noether's theorem is concerned with infinitesimal transformations

$$\bar{q}^i = q^i + \delta q^i(q,t), \qquad \bar{t} = t + \delta t(q,t), \tag{4}$$

and the symmetry requirement may be stated as follows.

The transformation (4) is a symmetry transformation for the Lagrangian $L(q^i, \dot{q}^i, t)$ if there exists a function $\delta f(q^i,t)$ such that

$$\delta L \equiv \frac{\partial L}{\partial q^i}\delta q^i + \frac{\partial L}{\partial \dot{q}^i}(\delta q^i)^{\cdot} + \frac{\partial L}{\partial t}\delta t + \Big(L - \frac{\partial L}{\partial \dot{q}^i}\dot{q}^i\Big)(\delta t)^{\cdot} \equiv -\frac{d}{dt}(\delta f), \tag{5}$$

and the theorem asserts that if condition (5) is met, then

$$\delta G = \frac{\partial L}{\partial \dot{q}^i}\delta q^i + \Big(L - \frac{\partial L}{\partial \dot{q}^i}\dot{q}^i\Big)\delta t + \delta f \tag{6}$$

is a constant of the motion. The proof is straightforward when one uses Eq. (6), condition (5), and the equations of motion for the problem

$$E_i L \equiv \frac{d}{dt}\Big(\frac{\partial L}{\partial \dot{q}^i}\Big) - \frac{\partial L}{\partial q^i} = 0. \tag{7}$$

The conventional treatment may be found in many books (see, for instance, Refs. 2 and 3).

The purpose of this paper is to derive a similar theorem when a more general definition of symmetry is used.

The idea in Noether's theorem is to look for transformations which change the Lagrangian at most by a total time derivative leaving, therefore, the equations of motion invariant.

One can consider a wider class of transformations that will alter the equations of motion without changing their solution. This amounts to considering transformations that will change the equations of motion covariantly leaving the solution space invariant. In other words, we are looking for transformations which will take a solution curve into another solution curve of the same problem.

As a matter of fact, the idea just described has been discussed recently[4–6] and its relationship to conservation laws has been found. We will briefly outline the main definitions and results in what follows.

Two Lagrangians $L(q^i, \dot{q}^i, t)$ and $\bar{L}(q^i, \dot{q}^i, t)$ are called s(olution)-equivalent iff their Euler–Lagrange equations have the same set of solutions.

It is not difficult to prove that if L and $\bar{L}$ are s-equivalent, then[4]

$$E_i \bar{L} = \Lambda_i^{\ j}(q, \dot{q}, t) E_j L, \tag{8}$$

and

$$(\mathrm{tr}(\Lambda^k))^{\cdot} = 0 \quad \text{for integer } k > 0, \tag{9}$$

where we have assumed that L and $\bar{L}$ are nonsingular, i.e.,

$$\det W_{ij} \equiv \det \frac{\partial^2 L}{\partial \dot{q}^i \partial \dot{q}^j} \neq 0, \tag{10}$$

$$\det \bar{W}_{ij} \equiv \det \frac{\partial^2 \bar{L}}{\partial \dot{q}^i \partial \dot{q}^j} \neq 0. \tag{11}$$

Due to the Cayley–Hamilton theorem, at most N of the constants appearing in Eq. (9) are functionally independent.

The matrix Λ can be written as

$$\Lambda = \overline{W}W^{-1} \quad (\det \Lambda \neq 0). \tag{12}$$

González-Gascón[7] and Lutzky[8] have suggested ways to use the result stated above to obtain conserved quantities from symmetries related to s-equivalence within the context of a second-order formalism (see also Refs. 4–6).

In this paper we use the counterpart of the result given in Eq. (9) for a first-order formulation of the dynamical problem.[5] The advantages of using a first-order formalism are essentially two.

(i) For a first-order system of differential equations, there are always infinitely many Lagrangians associated with it, which makes s-equivalence an interesting possibility.[5,6] For second-order systems, there may be essentially one Lagrangian (up to the addition of a total time derivative) or none at all, which makes s-equivalence trivial or nonexistent.[9,10]

(ii) Due to the fact that the number of equations of motion for a first-order system is twice as many as the number of equations of the second-order system from which it originated, one may, in principle, get as many as twice the number of functionally independent constants of the motion.

In Sec. II we briefly review the results obtained in the first-order formulation of the inverse problem of the calculus of variations. Section III contains the theorem which links *several* constants of the motion to *one* s-equivalence relation. In Sec. IV we present one example to illustrate the theorem and Sec. V contains the conclusions.

II. REVIEW OF FIRST-ORDER LAGRANGIAN FORMALISM

The contents of this section may be viewed as a summary of some aspects of Ref. 5 relevant to the subject of this paper.

Consider a second-order differential system

$$\ddot{q}^i = F^i(q^j, \dot{q}^j, t), \quad i,j = 1,\dots, N. \tag{13}$$

It is well known that any such system can be brought to a first-order form by introducing suitable new variables. For instance, let us define $2N$ variables x^a $(a = 1,\dots,2N)$

$$x^i = q^i, \quad i = 1,\dots, N, \qquad x^{N+j} = \dot{q}^j, \quad j = 1,\dots, N. \tag{14}$$

Therefore, Eq. (13) and definitions (14) can be written as

$$\dot{x}^a = f^a(x^b, t), \quad a, b = 1,\dots,2N, \tag{15}$$

where

$$f^i = x^{i+N}, \quad i = 1,\dots,N, \qquad f^{N+j} = F^j(x^k, x^{k+N}, t), \quad j, k = 1,\dots, N. \tag{16}$$

It will be useful to introduce

$$x^0 = t, \qquad f^0 = 1 \tag{17}$$

to write Eq. (15) as

$$\dot{x}^\mu = f^\mu(x^\nu), \quad \mu, \nu = 0,1,\dots,2N. \tag{18}$$

It is easy to see that a Lagrangian $L(x^\mu, \dot{x}^\mu)$ for the first-order systems (18) must be linear in the velocities $\dot{x}^\mu$ and can, therefore, be written as

$$L = l_\mu(x^\nu)\dot{x}^\mu. \tag{19}$$

In order for the Lagrangian (19) to give rise to equations of motion equivalent to (18) it is necessary and sufficient to have

$$l_\mu = l_{(a)}(C^{(b)}) \frac{\partial C^{(a)}}{\partial x^\mu} + \frac{\partial \Lambda}{\partial x^\mu}, \tag{20}$$

where $l_{(a)}$ are arbitrary functions of the $2N$ functionally independent constants of the motion $C^{(a)}$ which satisfy

$$\det \eta_{(a)(b)} \equiv \det \left(\frac{\partial l_{(a)}}{\partial C^{(b)}} - \frac{\partial l_{(b)}}{\partial C^{(a)}} \right) \neq 0, \tag{21}$$

where the $C^{(a)}$ are such that

$$\frac{\partial C^{(a)}}{\partial x^\mu} f^\mu \equiv 0, \tag{22}$$

and

$$\det \frac{\partial C^{(a)}}{\partial x^b} \neq 0. \tag{23}$$

Here Λ is an arbitrary function of x^μ. Therefore L can be written as

$$L = l_{(a)}(C^{(b)}) \frac{\partial C^{(a)}}{\partial x^\mu} \dot{x}^\mu + \frac{d\Lambda}{dt} \equiv L^{(0)} + \frac{d\Lambda}{dt}. \tag{24}$$

It is worth mentioning that $L^{(0)}$ vanishes when the equations of motion are satisfied, i.e.,

$$L^{(0)}\big|_{\dot{x}^\mu = f^\mu} \equiv 0. \tag{25}$$

Note that the change of variables

$$C^{(\alpha)} = C^{(\alpha)}(x^\mu) \quad (C^{(0)} \equiv x^0 = t) \tag{26}$$

is well-defined because of Eq. (23).

The equations of motion obtained from the Lagrangian (24) are

$$M_{\mu\nu}\dot{x}^\nu = 0, \tag{27}$$

where

$$M_{\mu\nu} \equiv \frac{\partial l_\mu}{\partial x^\nu} - \frac{\partial l_\nu}{\partial x^\mu} = -M_{\nu\mu}. \tag{28}$$

Obviously, the $2N+1$ equations (27) are not independent due to the fact that M is singular (because it is an antisymmetric matrix of odd dimensionality). The number of independent equations is $2N$ because of relations (21) and (23). Consider the matrix m_{ab} defined by

$$m_{ab} = M_{ab}, \quad a, b = 1,\dots,2N. \tag{29}$$

It is not difficult to realize that

$$\det m \neq 0. \tag{30}$$

Assume now that $\overline{L}$ is a Lagrangian s-equivalent to L. It is straightforward to prove that

$$(\mathrm{tr}(\Lambda^k))^{\cdot} = 0, \quad \text{for any integer } k > 0, \tag{31}$$

with

$$\Lambda = \overline{m}m^{-1}. \tag{32}$$

This is the first-order counterpart of the theorem stated in the Introduction [Eqs. (9)]. To end this section, we should emphasize that condition (21) can be met in infinitely many ways so that any first-order system has infinitely many s-equivalent Lagrangians.

III. s-EQUIVALENCE AND CONSERVED QUANTITIES

We will now prove the theorem which constitutes the main contribution of this paper. Given a Lagrangian $L(x^\mu, \dot{x}^\mu)$ for a first-order differential system, define P,

$$P \equiv \frac{\partial L}{\partial x^\mu}\delta x^\mu + \frac{\partial L}{\partial \dot{x}^\mu}(\delta x^\mu)^{\cdot} = l_{\nu,\mu}\dot{x}^\nu \delta x^\mu + (\delta x^\mu)_{,\nu}\dot{x}^\nu l_\mu, \tag{33}$$

where

$$\delta x^\mu = \delta x^\mu(x^\nu). \tag{34}$$

The transformation (34) is a symmetry transformation (related to s-equivalence) if

$$E_\nu P|_{\dot{x}^\mu = f^\mu} \equiv \left(\frac{d}{dt}\frac{\partial}{\partial \dot{x}^\nu} - \frac{\partial}{\partial x^\nu}\right)P|_{\dot{x}^\mu = f^\mu} = 0. \tag{35}$$

The conserved quantities associated with this symmetry transformation will be called non-Noetherian.

The non-Noetherian conserved quantities I_k are

$$I_k = \operatorname{tr} \Lambda^k, \tag{36}$$

with

$$\Lambda = \overline{m}m^{-1}, \tag{37}$$

where $\overline{m}$ is obtained from $\overline{L}$ defined by

$$\overline{L} = L + P. \tag{38}$$

Now for the proof. First note that P can be written as

$$P = \Pi_\nu(x_\rho)\dot{x}^\nu, \tag{39}$$

with

$$\Pi_\nu = l_{\nu,\mu}\delta x^\mu + (\delta x^\mu)_{,\nu} l_\mu. \tag{40}$$

Therefore, performing the change of coordinates (26)

$$\Pi_\nu = \Pi_{(\alpha)}(C^{(\beta)})\frac{\partial C^{(\alpha)}}{\partial x^\nu}, \tag{41}$$

or

$$\Pi_\nu = p_{(a)}(C^{(b)}, C^{(0)})\frac{\partial C^{(a)}}{\partial x^\nu} + \frac{\partial \Lambda}{\partial C^{(\alpha)}}\frac{\partial C^{(\alpha)}}{\partial x^\nu} \equiv p_\nu + \Lambda_{,\nu}, \tag{42}$$

where we have chosen $\Lambda(C^{(\alpha)})$ such that

$$\frac{\partial \Lambda(C^{(\alpha)})}{\partial C^{(0)}} = \Pi_{(0)}(C^{(\alpha)}). \tag{43}$$

Now, the equations of motion for P are

$$E_\mu P \equiv (\Pi_{\mu,\nu} - \Pi_{\nu,\mu})\dot{x}^\mu = (p_{\mu,\nu} - p_{\nu,\mu})\dot{x}^\nu = 0. \tag{44}$$

We now assume that Eq. (35) holds. It is straightforward to prove that when the equations of motion for L hold, i.e., Eq. (22), and because of Eqs. (23) and (26), one gets

$$\frac{\partial p_{(a)}}{\partial C^{(0)}} = 0, \tag{45}$$

and therefore

$$p_{(a)} = p_{(a)}(C^{(b)}). \tag{46}$$

For $\overline{L}$ we get

$$\overline{L} = \overline{l}_\mu \dot{x}^\mu = (l_\mu + \Pi_\mu)\dot{x}^\mu, \tag{47}$$

or

$$L = [l_{(a)}(C^{(b)}) + p_{(a)}(C^{(b)})]\frac{\partial C^{(a)}}{\partial x^\mu}\dot{x}^\mu + \frac{d\Lambda}{dt}. \tag{48}$$

It is straightforward to prove now that I_k are constants of the motion with definitions (37), (29), and (49). It may happen that

$$\det \overline{m} = 0, \tag{49}$$

and therefore not even the first $2N$ I_k's are functionally independent. In general there are less than $2N$ independent equations derived from $\overline{L}$. If

$$\det \overline{m} \neq 0, \tag{50}$$

then $\overline{L}$ and L are s-equivalent and there may be as much $2N$ functionally independent I_k's. We have then proven that considering the definition (35) of symmetry in first-order systems one may get up to $2N$ non-Noetherian constants of motion. On the other hand, in the second-order formalism one may find, at most, N non-Noetherian constants of motion when equivalent Lagrangians exist.

IV. EXAMPLE

The example we present in what follows intends to prove that several functionally independent constants of motion of a given problem may, in principle, be obtained from *one* transformation of s-equivalence, thus showing the power of the theorem proved in this paper.

Consider the two-dimensional free-particle Lagrangian

$$L = \tfrac{1}{2}(x_1\dot{x}_2 - x_2\dot{x}_1 + x_2^2 + x_3\dot{x}_4 - x_4\dot{x}_3 + x_4^2). \tag{51}$$

The Euler–Lagrange equations are

$$\begin{aligned} &\dot{x}_2 = 0, \\ &-\dot{x}_1 + x_2 = 0, \\ &\dot{x}_4 = 0, \\ &-\dot{x}_3 + x_4 = 0. \end{aligned} \tag{52}$$

Consider the transformation

$$\begin{aligned} \delta x_1 &= (x_1 - x_2 t)x_4 + (x_3 - x_4 t)x_2 t, \\ \delta x_2 &= (x_3 - x_4 t)x_2, \\ \delta x_3 &= (x_3 - x_4 t)x_2 + (x_1 - x_2 t)x_4 t, \\ \delta x_4 &= (x_1 - x_2 t)x_4. \end{aligned} \tag{53}$$

It is straightforward to construct P:

$$\begin{aligned} P = {} & \dot{x}_1[-\tfrac{1}{2}x_2x_4 - \tfrac{1}{2}(x_3 - x_4 t)x_2 - \tfrac{1}{2}x_4^2 t + \tfrac{1}{2}x_3x_4] + \dot{x}_2[\tfrac{1}{2}x_1x_4 + \tfrac{1}{2}x_1(x_3 - x_4 t) - \tfrac{1}{2}x_4(x_3 - x_4 t) + \tfrac{1}{2}x_4^2 t^2 - \tfrac{1}{2}x_3x_4 t] \\ & + \dot{x}_3[-\tfrac{1}{2}x_2^2 t + \tfrac{1}{2}x_1x_2 - \tfrac{1}{2}x_4x_2 - \tfrac{1}{2}(x_1 - x_2 t)x_4] + \dot{x}_4[-\tfrac{1}{2}x_2(x_1 - x_2 t) + \tfrac{1}{2}x_2^2 t^2 - \tfrac{1}{2}x_1x_2 t + \tfrac{1}{2}x_3x_2 \\ & + \tfrac{1}{2}x_3(x_1 - x_2 t)] + \tfrac{1}{2}x_2^2x_4 + \tfrac{1}{2}x_2^2x_4 t + \tfrac{1}{2}x_2^2(x_3 - x_4 t) - \tfrac{1}{2}x_1x_2x_4 + \tfrac{1}{2}x_4^2x_2 + \tfrac{1}{2}x_4^2x_2 t + \tfrac{1}{2}x_4^2(x_1 - x_2 t) - \tfrac{1}{2}x_2x_3x_4. \end{aligned} \tag{54}$$

It is not difficult to prove that Eq. (35) is satisfied, and the matrix $\Omega \equiv \Lambda - 1$ is

$$\Omega = \begin{bmatrix} (x_3 - x_4t) + x_4 & 0 & -(x_2 - x_4)t & -(x_2 - x_4) \\ 0 & (x_3 - x_4t) + x_4 & (x_3 - x_4t) - (x_1 - x_2t) + (x_2 - x_4)t^2 & (x_2 - x_4)t \\ (x_2 - x_4)t & x_2 - x_4 & (x_1 - x_2t) + x_2 & 0 \\ (x_1 - x_2t) - (x_3 - x_4t) - (x_2 - x_4)t^2 & -(x_2 - x_4)t & 0 & (x_1 - x_2t) + x_2 \end{bmatrix}. \tag{55}$$

We adopt the notation

$$\begin{aligned} C_1 &= x_1 - x_2t, \\ C_2 &= x_2, \\ C_3 &= x_3 - x_4t, \\ C_4 &= x_4. \end{aligned} \tag{56}$$

The traces of powers of Ω are

$$\operatorname{tr} \Omega = 2(C_1 + C_2 + C_3 + C_4), \tag{57}$$

$$\operatorname{tr} \Omega^2 = 2[(C_1 + C_2)^2 + (C_3 + C_4)^2 - 2(C_1 - C_3)(C_2 - C_4)], \tag{58}$$

$$\operatorname{tr} \Omega^3 = 2[(C_1 + C_2)^3 + (C_3 + C_4)^3 - 3(C_1 + C_2)(C_1 - C_3)(C_2 - C_4) - 3(C_3 + C_4)(C_1 - C_3)(C_2 - C_4)], \tag{59}$$

$$\operatorname{tr} \Omega^4 = 2\{[(C_1 + C_2)^2 - (C_1 - C_3)(C_2 - C_4)]^2 + [(C_3 + C_4)^2 - (C_1 - C_3)(C_2 - C_4)]^2 - 2(C_1 + C_2 + C_3 + C_4)^2(C_1 - C_3)(C_2 - C_4)\}. \tag{60}$$

Equations (57)–(60) and (36) imply that there are two functionally independent constants of motion which are written as functions of C_1, C_2, C_3, C_4. These constants are

$$A = C_1 + C_2 + C_3 + C_4, \tag{61}$$

$$B = (C_1 + C_2)(C_3 + C_4) + (C_1 - C_3)(C_2 - C_4), \tag{62}$$

and the traces of powers of Ω are functions of them only. We have got *two* functionally independent constants of motion from *one* s-equivalence symmetry transformation (53).

V. CONCLUSIONS

We have presented a theorem which associates to *one* (s-equivalence) symmetry transformation *several* constants of motion. The construction of the constants involves no integration and it is algebraic in character. This theorem is based on a previous one obtained for second-order Lagrangians.[4] The fact that we have a first-order formulation implies that it is always possible to find s-equivalence transformations and to get several constants of motion for one problem as we did in the example.

The results of this paper are, strictly speaking, useful for constants of motion which are defined globally.

It may be interesting to try to generalize the methods presented here to systems for which some of the constants of motion are defined only locally.

The relationship of s-equivalence to symmetries of the equations of motion will be given in a forthcoming paper.[11]

[1]E. Noether, Nachr. Kön. Wiss. Math. Phys. Kl., 235 (1918).
[2]E. L. Hill, Rev. Mod. Phys. **23**, 253 (1951).
[3]D. Lovelock and H. Rund, *Tensors, Differential Forms and Variational Principles* (Wiley–Interscience, New York, 1975).
[4]S. Hojman and H. Harleston, J. Math. Phys. **22**, 1414 (1981).
[5]S. Hojman and L. F. Urrutia, J. Math. Phys. **22**, 1896 (1981).
[6]S. Hojman and L. Shepley, Rev. Mex. Fís. **28**, 149 (1982).
[7]F. González-Gascón, J. Phys. A **13**, 1965 (1980).
[8]M. Lutzky (private communication).
[9]J. Douglas, Trans. Am. Math. Soc. **50**, 71 (1941).
[10]M. Henneaux, Ann. Phys. (N.Y.) **140**, 45 (1982).
[11]S. Hojman and F. Zertuche (in preparation).

The scattering of an obliquely incident surface wave by a submerged fixed vertical plate

Birendranath Mandal
Department of Applied Mathematics, University College of Science, 92, A.P.C. Road, Calcutta-700 009, India
Sudip Kumar Goswami
Department of Mathematics, Presidency College, Calcutta-700 073, India

(Received 29 June 1983; accepted for publication 4 November 1983)

The problem of scattering of surface waves obliquely incident on a submerged fixed vertical plate is solved approximately for a small angle of incidence by reducing it to the solution of an integral equation. The correction to the reflection and transmission coefficients over their normal incidence values for a small angle of incidence are obtained. For different values of the incident angle these coefficients are evaluated numerically, taking particular values of the wave number and the depth of the plate, and represented graphically.

PACS numbers: 03.40.Kf, 47.35. + i, 02.30.Rz

1. INTRODUCTION

Dean[1] and Ursell[2] first considered the problems of scattering of surface waves normally incident on submerged and partially immersed fixed vertical plane barriers in deep water. These problems are subsequently studied by Williams[3], Goswami[4], and others by employing different mathematical techniques, e.g., reduction method, integral equation method, etc. The problem of scattering of surface waves normally incident on a submerged fixed vertical plate in water of finite depth was considered by Goswami,[5] the deep water case being earlier considered by Evans.[6]

The scattering of surface waves obliquely incident on a partially immersed or completely submerged vertical barrier in deep water was studied by Faulkner,[7,8] Jarvis and Taylor,[9] Evans and Morris.[10] In the present paper the problem of scattering of surface waves obliquely incident on a submerged fixed vertical plate in deep water is solved by reducing it to the solution of an integral equation involving the unknown difference of velocity potentials across the plate by a simple use of Green's integral theorem in the fluid medium. The kernel of this integral equation is expanded in a series involving different orders of the sine of the angle of incidence which is assumed to be small. This expansion of the kernel suggests the corresponding form of the expansion of the unknown function in the integral equation and this is then used to solve the integral equation approximately. A somewhat similar type of technique of solving the integral equation approximately was successfully used by Goswami[5,11,12] and Mandal and Goswami.[13]

2. STATEMENT OF THE PROBLEM

We consider the scattering of surface waves by a submerged fixed vertical plate in deep water and use a coordinate system in which the y axis is taken to be vertically downwards, the mean-free surface is the plane $y = 0$, and the position of the plate is given by $x = 0, a \leqslant y \leqslant b, -\infty < z < \infty$. Assuming the fluid to be inviscid and incompressible and the motion to be irrotational and simple harmonic in time with circular frequency σ and small amplitude, a velocity potential exists and it may be taken to be the real part of $\chi(x, y, z)\, e^{-i\sigma t}$ satisfying the equations

$$\nabla^2 \chi = 0, \text{ in the fluid region,}$$

$$\frac{\partial \chi}{\partial y} + K\chi = 0, \text{ on } y = 0,$$

$$\frac{\partial \chi}{\partial x} = 0, \text{ on } x = 0,\ a < y < b,$$

where $K = \sigma^2/g$.

A wave represented by $\chi_0 = \exp(-Ky + i\mu x + i\nu z)$, where $\mu = K\cos\alpha$, $\nu = K \sin\alpha$ is assumed to be incident at an angle α to the normal of the plate from negative infinity. Such a wave will be partially reflected and transmited by the plate, and in view of the geometry of the plate it is reasonable to assume $\chi(x, y, z) = \Phi(x, y) e^{i\nu z}$. Then Φ must satisfy

$$\frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2} - \nu^2 \Phi = 0, \quad y \geqslant 0, \tag{2.1}$$

$$\frac{\partial \Phi}{\partial y} + K\Phi = 0, \quad y = 0, \tag{2.2}$$

$$\frac{\partial \Phi}{\partial x} = 0, \quad x = 0, \quad a < y < b, \tag{2.3}$$

and it and its derivatives are continuous everywhere except possibly across $x = 0, a < y < b$. We also require that Φ and its first derivatives are bounded everywhere away from the lines $x = 0, y = a$ and $x = 0, y = b$ and that near these lines $\{x^2 + (y-a)^2\}^{1/2}$ grad Φ and $\{x^2 + (y-b)^2\}^{1/2}$ grad Φ, respectively, are bounded. This is called the edge condition. Finally we assume that as $|x| \to \infty$, $\Phi(x, y)$ has the asymptotic forms

$$\begin{aligned} \Phi(x, y) &\sim \exp(i\mu x - Ky) + R \exp(-i\mu x - Ky) \quad (x \to -\infty), \\ \Phi(x, y) &\sim T \exp(i\mu x - Ky) \quad (x \to +\infty), \end{aligned} \tag{2.4}$$

where R and T are the (complex) reflection and transmission coefficients, respectively. Let us write $\Phi(x, y) = \varphi_0(x, y) + \varphi(x, y)$, where $\varphi_0(x, y) = \exp(-Ky + i\mu x)$ and $\varphi(x, y)e^{i\nu z}$ is the scattered velocity potential. $\varphi(x, y)$ also satisfies (2.1), (2.2), and the edge conditions stated above, and by the conditions (2.4) $\varphi(x, y)$ represents an outgoing wave at infinity.

3. REDUCTION TO AN INTEGRAL EQUATION

Following Levine,[14] the generalized Green's function satisfying (2.1), (2.2), and G, grad G being bounded at a large distance, and G representing an outgoing wave at infinity, may be obtained as

$$G(x, y;\xi,\eta) = K_0(\nu\rho) - K_0(\nu\rho^*) + i\frac{2\pi K}{(K^2-\nu^2)^{1/2}}\exp\{-K(y+\eta) + i(K^2-\nu^2)^{1/2}|x-\xi|\}$$
$$+2\int_\nu^\infty \frac{(k^2-\nu^2)^{1/2}\cos[(k^2-\nu^2)^{1/2}(y+\eta)] - K\sin[(k^2-\nu^2)^{1/2}(y+\eta)]}{K^2+k^2-\nu^2}\exp\{-k|x-\xi|\}dk, \quad (3.1)$$

where $\rho^2, \rho^{*2} = (x-\xi)^2 + (y\mp\eta)^2$.

Now applying Green's theorem to $\varphi(x, y)$ and $G(x, y;\xi,\eta)$ in the fluid region we obtain

$$2\pi\varphi(\xi,\eta) = \int_a^b f(y)\frac{\partial G}{\partial x}(0,y;\xi,\eta)dy, \quad (3.2)$$

where

$$f(y) \equiv \varphi(+0, y) - \varphi(-0, y).$$

By (2.3) and (3.2), we have

$$-i2\pi K\cos\alpha e^{-K\eta} = \int_a^b f(y)\frac{\partial^2 G}{\partial\xi\partial x}(0,y;0,\eta)dy,$$
$$a<\eta<b. \quad (3.3)$$

Writing $\epsilon = \sin\alpha$ and following Mandal and Goswami[13] we can obtain

$$\frac{\partial^2 G}{\partial\xi\partial x}(0, y;0,\eta) = \frac{\partial^2 G_0}{\partial\eta^2}(0, y;0,\eta) - \frac{1}{2}K^2\epsilon^2 G_0(0, y;0,\eta) + O(\epsilon^4\ln\epsilon,\epsilon^4),$$
$$a<y, \ \eta<b, \quad (3.4)$$

where

$$G_0(0, y;0,\eta) = -\ln\left|\frac{y-\eta}{y+\eta}\right| - 2\int_0^\infty \frac{K\sin k(y+\eta) - k\cos k(y+\eta)}{K^2+k^2}dk + i2\pi\exp\{-K(y+\eta)\},$$

and $G_0(x, y;\xi,\eta)$ is the expression (3.1) when $\alpha = 0$. From the expansion (3.4) it is reasonable to expand $f(y)$ in the following form:

$$f(y) = f_0(y) + \epsilon^2 f_1(y) + O(\epsilon^4\ln\epsilon,\epsilon^4). \quad (3.5)$$

By (3.3), (3.4), and (3.5), and by noting the coefficients of terms involving a different order of ϵ, we obtain

$$-i2\pi K e^{-K\eta} = \frac{d^2}{d\eta^2}\int_a^b f_0(y)G_0(0, y;0,\eta)dy, \quad a<\eta<b, \quad (3.6)$$

and

$$i\pi K e^{-K\eta} = \frac{d^2}{d\eta^2}\int_a^b f_1(y)G_0(0, y;0,\eta)dy - \frac{K^2}{2}\int_a^b f_0(y)G_0(0,y;0,\eta)dy,$$
$$a<\eta<b. \quad (3.7)$$

Let us now define a function $\psi(y)$ by

$$\psi(y) = Kf(y) + f'(y)$$

so that

$$f(y) = e^{-Ky}\int_a^y e^{Ku}\psi(u)du. \quad (3.8)$$

Then $\psi(y)$ may be expanded as

$$\psi(y) = \psi_0(y) + \epsilon^2\psi_1(y) + O(\epsilon^4\ln\epsilon,\epsilon^4). \quad (3.9)$$

Now integrating both sides of (3.6) with respect to η, we have

$$A_0 + i2\pi e^{-K\eta} = \frac{d}{d\eta}\int_a^b f_0(y)G_0(0, y;0,\eta)dy, \quad a<\eta<b, \quad (3.6a)$$

and by (3.6) and (3.6a), we obtain

$$\int_a^b \psi_0(y)\frac{2y\,dy}{y^2-\eta^2} = KA_0. \quad (3.10)$$

The singular integral equation (3.10) is a Cauchy-type one and following Mikhlin,[15]

$$\psi_0(y) = \frac{D_0(d_0^2 - y^2)}{(y^2-a^2)^{1/2}(b^2-y^2)^{1/2}}, \quad a<y<b, \quad (3.11)$$

where $D_0(= -KA_0/\pi)$, d_0^2 are constants to be determined. Since $f(b) = 0$, by (3.8)

$$\int_a^b \psi(u)e^{Ku}\,du = 0,$$

and therefore,

$$\int_a^b \psi_0(u)e^{Ku}\,du = 0, \quad \int_a^b \psi_1(u)e^{Ku}\,du = 0, \text{ etc.}, \quad (3.12)$$

so that d_0^2 may be obtained from

$$\int_a^b \frac{(d_0^2 - u^2)e^{Ku}\,du}{(u^2-a^2)^{1/2}(b^2-u^2)^{1/2}} = 0. \quad (3.13)$$

By (3.6) and (3.8), we obtain

$$\int_a^b \psi_0(y)\left[2\int_0^\infty \frac{\sin ky(K\sin k\eta - k\cos k\eta)}{K^2+k^2}k\,dk - i\pi K\exp\{-K(y+\eta)\}\right]dy = i2\pi e^{-K\eta},$$
$$a<\eta<b. \quad (3.14)$$

Now by (3.11) and (3.14), we obtain

$$D_0 = \frac{2i}{p_0 - q_0 - ir_0}, \quad (3.15)$$

where

$$p_0 = \int_{-a}^a \frac{(d_0^2 - u^2)e^{-Ku}\,du}{(a^2-u^2)^{1/2}(b^2-u^2)^{1/2}},$$

$$q_0 = \int_b^\infty \frac{(d_0^2 - u^2)e^{-Ku}\,du}{(u^2-a^2)^{1/2}(u^2-b^2)^{1/2}},$$

and

$$r_0 = \int_a^b \frac{(d_0^2 - u^2)e^{-Ku}\,du}{(u^2-a^2)^{1/2}(b^2-u^2)^{1/2}}.$$

Similarly from (3.7) and (3.11), we obtain

$$\psi_1(y) = \frac{K^2 D_0 (d_1^4 - y^4)}{4(y^2-a^2)^{1/2}(b^2-y^2)^{1/2}} + i\left(p_1 - q_1 - ir_1 + \frac{2e^{Ka}}{K^3}(1-Ka)\right) \times \frac{K^2 D_0^2 (d_0^2 - y^2)}{8(y^2-a^2)^{1/2}(b^2-y^2)^{1/2}}, \quad a<y<b, \qquad (3.16)$$

where

$$p_1 = \int_{-a}^{a} \frac{(d_1^4 - u^4)e^{-Ku}\,du}{(a^2-u^2)^{1/2}(b^2-u^2)^{1/2}},$$

$$q_1 = \int_b^{\infty} \frac{(d_1^4 - u^4)e^{-Ku}\,du}{(u^2-a^2)^{1/2}(u^2-b^2)^{1/2}},$$

and

$$r_1 = \int_a^b \frac{(d_1^4 - u^4)e^{-Ku}\,du}{(u^2-a^2)^{1/2}(b^2-u^2)^{1/2}}$$

and d_1^4 is given by

$$\int_a^b \frac{(d_1^4 - u^4)e^{Ku}\,du}{(u^2-a^2)^{1/2}(b^2-u^2)^{1/2}} = 0.$$

4. REFLECTION AND TRANSMISSION COEFFICIENTS

As $\xi \to -\infty$ and $+\infty$, respectively, we have by (3.2) and (3.8) the complex reflection and transmission coefficients R and T as

$$R = -\frac{1}{2}\int_a^b \psi(y)e^{-Ky}\,dy,$$
$$T = 1 + \frac{1}{2}\int_a^b \psi(y)e^{-Ky}\,dy, \qquad (4.1)$$

so that $T = 1 - R$ and hence

$$T_0 = 1 - R_0, \quad T_1 = -R_1, \quad \text{etc.},$$

where

$$R = R_0 + \epsilon^2 R_1 + O(\epsilon^4 \ln \epsilon, \epsilon^4),$$
$$T = T_0 + \epsilon^2 T_1 + O(\epsilon^4 \ln \epsilon, \epsilon^4). \qquad (4.2)$$

Now by (4.1), (4.2), (3.11), (3.15), and (3.16), we have

$$R_0 = -\tfrac{1}{2} D_0 r_0,$$
$$R_1 = -i\tfrac{1}{16} K^2 D_0^2 \left(r_0(p_1 - q_1) - r_1(p_0 - q_0) + \frac{2}{K^3}(1-Ka) r_0 e^{Ka} \right), \qquad (4.3)$$

etc.

5. DISCUSSION

The reflection and transmission coefficients obtained here are valid for all angles of incidence ($0° \leqslant \alpha < 90°$) of the surface wave and for all wavelengths other than short ones. We have calculated these coefficients numerically for different values of α in the range $0° \leqslant \alpha \leqslant 15°$ thus enabling us to retain only up to two terms in the different analytical approximations. However, by taking an appropriate number of terms in these approximations, these coefficients can be calculated for values of α beyond 15°.

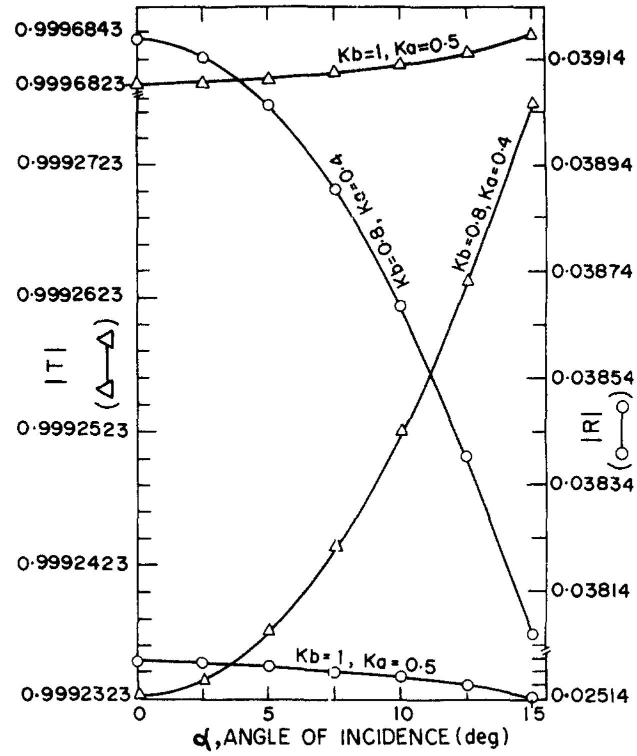


FIG. 1. $|T|$ and $|R|$ against α(deg).

Thus the integral equation method is found to be simple and straightforward and gives equally good results in all the cases of partially immersed as well as submerged fixed vertical barrier and plate in case of oblique incidence of a surface wave train. In contrast to this, Evans and Morris[10] found by using the method of variational principle that the results are not so good in the case of submerged fixed vertical barrier in comparison to those found in the case of a partially immersed fixed vertical barrier.

It should be noted that $|R_0|$, $|T_0|$ give the corresponding results for a normally incident wave train and are in complete agreement with those obtained earlier by Evans.[6] Taking $Kb = 2Ka = 0.8$ and $Kb = 2Ka = 1.0$, respectively, $|R|$ and $|T|$ are calculated for $\alpha = 0°, 2.5°, 5°, 7.5°, 10°, 12.5°$, and 15°, and plotted in Fig. 1.

It appears from the figure that for fixed Ka (and hence, Kb), transmission coefficient increases and reflection coefficient decreases with α, which is similar to the situation arising in the case of a partially immersed barrier (cf. Evans and Morris,[10] Mandal and Goswami[13]).

ACKNOWLEDGMENT

One of us (S.K.G.) thanks the U.G.C., New Delhi, for the award of a Teaching Fellowship under the Faculty Improvement Program. We also take this opportunity to thank the referee for his comments in revising the paper.

[1]W. R. Dean, Proc. Cambridge Philos. Soc. **41**, 231 (1945).
[2]F. Ursell, Proc. Cambridge Philos. Soc. **43**, 374 (1947).
[3]W. E. Williams, Proc. Cambridge Philos. Soc. **62**, 507 (1966).
[4]S. K. Goswami, ZAMM **62**, 637 (1982).
[5]S. K. Goswami, J. Indian Inst. Sci. **64**, 79 (1983).
[6]D. V. Evans, J. Fluid Mech. **40**, 433 (1970).
[7]T. R. Faulkner, Proc. Cambridge Philos. Soc. **62**, 829 (1966).
[8]T. R. Faulkner, ZAMP **17**, 699 (1966).
[9]R. J. Jarvis and B. S. Taylor, Proc. Cambridge Philos. Soc. **66**, 417 (1969).
[10]D. V. Evans and C. A. N. Morris, J. Inst. Math. Appl. **9**, 198 (1972).
[11]S. K. Goswami, Bull. Calcutta Math. Soc. **74**, 75 (1982).
[12]S. K. Goswami, Bull. Calcutta Math. Soc. **74**, 92 (1982).
[13]B. N. Mandal and S. K. Goswami, Appl. Sci. Res. **40** (1983) (to appear).
[14]H. Levine, J. Math. Phys. **6**, 1231 (1965).
[15]S. G. Mikhlin, *Integral Equation* (Pergamon, New York, 1964), pp. 131.

Variational principle for electromagnetic problems in a linear, static, inhomogeneous anisotropic medium

K. R. Brownstein
Department of Physics and Astronomy, University of Maine, Orono, Maine 04469

(Received 30 November 1983; accepted for publication 10 February 1984)

A variational principle $\delta \int\int L\, dV\, dt = 0$ is formulated for electromagnetic problems in the time domain for a medium which is linear, static, inhomogeneous, and anisotropic. The principle involves all four electromagnetic field vectors (**E**,**H**,**D**,**B**) and no "adjoint-type" fields. Both the two "curl" Maxwell equations as well as the two constitutive equations emerge as the four associated Euler–Lagrange equations. The problem of possible variations in the field functions at the temporal boundaries is solved by using the constitutive equations as a constraint at one of the temporal boundaries.

PACS numbers: 03.50.De, 02.70. + d

I. INTRODUCTION

Recently there has been an increasing interest in the solution of time-dependent electromagnetics problems. Since an exact solution is usually unattainable, one is led to seek an approximation method. Variational methods for approximately solving such problems offer not only elegance but also the possibility of computational efficiency.

Variational formulations for electromagnetics problems in the frequency domain have been proposed, e.g., by Morishita and Kumagai[1] and Mohsen.[2] A criticism of the former is that it involves varying the potentials (**A** and ϕ) rather than the electromagnetic fields (e.g., **E** and **H**). The latter reference is formulated purely in terms of the electromagnetic fields (**E** and **H**). However, it also employs two "adjoint" electromagnetic fields ($\mathbf{E}^a$ and $\mathbf{H}^a$); thus one is forced to use four electromagnetic field vectors only two of which have direct relevance to the actual problem.

Recently Mohsen[3] formulated a variational principle $\delta \int\int_{t_1}^{t_2} L\, dV\, dt = 0$ for electromagnetics problems directly in the time domain. This principle also employs the two electromagnetic fields (**E** and **H**) as well as the two adjoint electromagnetic fields ($\mathbf{E}^a$ and $\mathbf{H}^a$). Here the adjoint electromagnetic fields satisfy Maxwell's equations with the permittivity and permeability tensors ($\tilde{\epsilon}$ and $\tilde{\mu}$) replaced by their transposes.

Before proceeding, we raise two points concerning Mohsen's principle.

(i) In the time domain, the permittivity and permeability tensors must be symmetric. Thus there is no distinction between the electromagnetic fields (**E** and **H**) and their adjoints. Of course, all four of these fields must be regarded as independent for variational purposes.

(ii) In the derivation of the equivalence of Maxwell's equations and the variational principle, the fields are assumed to have no variation at $t = t_1$ and $t = t_2$. We may regard t_1 as the initial time; using the (presumed given) initial fields at $t = t_1$ assures no variation in the fields at this time. However, we cannot fix the fields at their correct values at $t = t_2$ since these are generally unknown. This seems to cast doubt on the ability to demonstrate the equivalence of Maxwell's equations and the variational principle.

The purpose of this article is to develop a variational principle for electromagnetics problems which (i) is in terms of the electromagnetic fields and not the potentials, (ii) is equivalent to Maxwell's equations, (iii) employs no "adjoint-type" fields, and (iv) does not suffer from the above difficulty at $t = t_2$.

II. FIELD EQUATIONS

The four (unknown) electromagnetic field vectors (**E**,**H**,**D**,**B**) are functions of position **r** and time t. These field vectors obey Maxwell's equations

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} - \mathbf{K}, \tag{1a}$$

$$\nabla \times \mathbf{H} = +\frac{\partial \mathbf{D}}{\partial t} + \mathbf{J}, \tag{1b}$$

$$\nabla \cdot \mathbf{D} = \rho, \tag{2a}$$

$$\nabla \cdot \mathbf{B} = m. \tag{2b}$$

Here $\mathbf{J}(\mathbf{r},t)$ and $\mathbf{K}(\mathbf{r},t)$ are the (given) electric and magnetic current densities; $\rho(\mathbf{r},t)$ and $m(\mathbf{r},t)$ are the corresponding electric and magnetic charge densities. These are related by equations of continuity,

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0, \tag{3a}$$

$$\nabla \cdot \mathbf{K} + \frac{\partial m}{\partial t} = 0. \tag{3b}$$

The introduction of fictitious magnetic current and charge densities preserves the theory under duality transformations.[4]

In view of Eqs. (1) and (3) the second pair of Maxwell's equations (2) are automatically satisifed if they are obeyed merely at any single value of time, say $t = t_1$. Therefore, we shall concentrate on satisfying the first pair of Maxwell's equations (1); the second pair of Maxwell's equations (2) are regarded as having a secondary role.

The two electric field vectors (**E** and **D**) are related through a constitutive equation

$$\mathbf{D} = \tilde{\epsilon} \cdot \mathbf{E}. \tag{4a}$$

Similarly, the two magnetic field vectors (**H** and **B**) are related through a constitutive equation

$$\mathbf{B} = \tilde{\mu} \cdot \mathbf{H}. \tag{4b}$$

 0022-2488/84/061784-03$02.50

The permittivity $\tilde{\epsilon}$ and permeability $\tilde{\mu}$ are assumed to be nonsingular tensors which are functions only of the position **r** and not of the time t:

$$\tilde{\epsilon} = \tilde{\epsilon}(\mathbf{r}) \quad \text{and} \quad \tilde{\mu} = \tilde{\mu}(\mathbf{r}). \tag{5}$$

It can readily be shown by a thermodynamic argument (Panofsky and Phillips[5]) that these two tensors must be symmetric; this symmetry will be assumed in what follows.

In summary, we shall be concerned with solving Maxwell's equations (1) for given electric and magnetic current densities (**J** and **K**). The medium under consideration can be classified as being linear, static (i.e., time-independent), inhomogeneous, anisotropic, and nondissipative.

III. VARIATIONAL PRINCIPLE

The variational principle which will be used is of the form

$$\delta \int_{t_1}^{t_2} \int_V L\left(\Psi_j, \frac{\partial \Psi_j}{\partial q_i}, q_i\right) dV\,dt = 0. \tag{6}$$

Here the Ψ_j are the field functions (the twelve components of the four electromagnetic field vectors), the q_i stand for (x,y,z,t) and $dV = dx\,dy\,dz$. The Euler–Lagrange partial differential equations which correspond to (6) are

$$\sum_{i=1}^{4} \frac{\partial}{\partial q_i}\left(\frac{\partial L}{\partial(\partial \Psi_j/\partial q_i)}\right) - \frac{\partial L}{\partial \Psi_j} = 0 \quad (j = 1,2,\ldots,12). \tag{7}$$

The equivalence of the variational principle (6) and the corresponding Euler–Lagrange equations (7) requires a certain behavior of the field functions at the (spatial and temporal) boundaries of the region of integration; this matter will be dealt with in Sec. IV.

The "Lagrangian density" L is now chosen to be

$$L = (\mathbf{H} - \mathbf{B}\cdot\tilde{\mu}^{-1})\cdot\left(\nabla\times\mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} + \mathbf{K}\right) - (\mathbf{E} - \mathbf{D}\cdot\tilde{\epsilon}^{-1})\cdot\left(\nabla\times\mathbf{H} - \frac{\partial \mathbf{D}}{\partial t} - \mathbf{J}\right). \tag{8}$$

In passing we note that (i) L is bilinear in the field functions and their first partial derivatives, (ii) L depends on **r** (but not t) via the inverse permittivity and permeability tensors, and (iii) L depends on **r** and t via the electric and magnetic current densities.

The twelve Euler-Lagrange equations (7) may be combined into four vector equations associated with the variation of each of the four electromagnetic field vectors. These four partial differential equations (and their associated electromagnetic field vector) are

$$\text{for } \mathbf{E},\ -\nabla\times(\mathbf{H} - \mathbf{B}\cdot\tilde{\mu}^{-1}) + \left(\nabla\times\mathbf{H} - \frac{\partial \mathbf{D}}{\partial t} - \mathbf{J}\right) = 0; \tag{9a}$$

$$\text{for } \mathbf{H},\ +\nabla\times(\mathbf{E} - \mathbf{D}\cdot\tilde{\epsilon}^{-1}) - \left(\nabla\times\mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} + \mathbf{K}\right) = 0; \tag{9b}$$

$$\text{for } \mathbf{D},\ +\frac{\partial(\mathbf{E} - \mathbf{D}\cdot\tilde{\epsilon}^{-1})}{\partial t} - \tilde{\epsilon}^{-1}\cdot\left(\nabla\times\mathbf{H} - \frac{\partial \mathbf{D}}{\partial t} - \mathbf{J}\right) = 0; \tag{9c}$$

$$\text{for } \mathbf{B},\ +\frac{\partial(\mathbf{H} - \mathbf{B}\cdot\tilde{\mu}^{-1})}{\partial t} + \tilde{\mu}^{-1}\cdot\left(\nabla\times\mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} + \mathbf{K}\right) = 0. \tag{9d}$$

Introduce four auxiliary electromagnetic field vectors ($\mathscr{E}$, $\mathscr{H}$, $\mathscr{D}$, $\mathscr{B}$) defined by

$$\mathscr{E} = \mathbf{E} - \mathbf{D}\cdot\tilde{\epsilon}^{-1}, \tag{10a}$$

$$\mathscr{H} = \mathbf{H} - \mathbf{B}\cdot\tilde{\mu}^{-1}, \tag{10b}$$

$$\mathscr{D} = \tilde{\epsilon}\cdot\mathbf{E} - \mathbf{D}, \tag{10c}$$

$$\mathscr{B} = \tilde{\mu}\cdot\mathbf{H} - \mathbf{B}. \tag{10d}$$

As a consequence of these definitions, the auxiliary electromagnetic field vectors ($\mathscr{E}$, $\mathscr{H}$, $\mathscr{D}$, $\mathscr{B}$) satisfy the constitutive equations

$$\mathscr{D} = \tilde{\epsilon}\cdot\mathscr{E}, \tag{11a}$$

$$\mathscr{B} = \tilde{\mu}\cdot\mathscr{H}, \tag{11b}$$

regardless of whether the constitutive equations are obeyed by (**E**,**H**,**D**,**B**). Using the definitions (10), Eqs. (9b) and (9d) can be combined to yield

$$\nabla\times\mathscr{E} = -\frac{\partial \mathscr{B}}{\partial t}; \tag{12a}$$

similarly, Eqs. (9a) and (9c) yield

$$\nabla\times\mathscr{H} = +\frac{\partial \mathscr{D}}{\partial t}. \tag{12b}$$

Therefore the auxiliary electromagnetic field vectors satisfy (i) the source-free Maxwell's equations (12), and (ii) the constitutive equations (11). Assuming that the auxiliary electromagnetic field vectors vanish at some single value of time, say $t = t_1$ [i.e., that the constitutive equations (4) are satisfied at $t = t_1$], and that the spatial boundary conditions cause no difficulty (see Sec. IV), the unique solution of Eqs. (12), (11), and (5) is

$$\mathscr{E} = 0, \quad \mathscr{H} = 0, \quad \mathscr{D} = 0, \quad \mathscr{B} = 0. \tag{13}$$

Thus, using (10), it follows that the constitutive equations (4) are satisfied for all t. Finally, substitution of (13) into either (9b) or (9d) gives the Maxwell equation (1a); similarly, substitution of (13) into either (9a) or (9c) gives the Maxwell equation (1b).

Thus, assuming correct initial and boundary conditions, the variational principle (6) together with the Lagrangian density (8) implies that the two Maxwell equations

$$\nabla\times\mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} - \mathbf{K}, \tag{1a}$$

$$\nabla\times\mathbf{H} = +\frac{\partial \mathbf{D}}{\partial t} + \mathbf{J}, \tag{1b}$$

as well as the two constitutive equations

$$\mathbf{D} = \tilde{\epsilon}\cdot\mathbf{E}, \tag{4a}$$

$$\mathbf{B} = \tilde{\mu}\cdot\mathbf{H} \tag{4b}$$

are all satisfied.

IV. SPATIAL AND TEMPORAL BOUNDARY CONDITIONS

The equivalence of the variational principle (6) and the corresponding Euler–Lagrange equations (7) is based upon the neglect of certain terms which arise from (spatial or temporal) integration by parts. Those terms which arise from spatial integration by parts (i.e., Gauss' theorem) are

$$-\int_{t_1}^{t_2}\int_S \{(\mathbf{H}-\mathbf{B}\cdot\tilde{\mu}^{-1})\cdot(\mathbf{n}\times\delta\mathbf{E}) - (\mathbf{E}-\mathbf{D}\cdot\tilde{\epsilon}^{-1})\cdot(\mathbf{n}\times\delta\mathbf{H})\}\,dS\,dt. \tag{14}$$

Here dS is an element of the surface area S which bounds the original volume V, and $\mathbf{n}$ is a unit outward normal vector at this bounding surface. Those terms which arise from temporal integration by parts are

$$\int_V \{(\mathbf{H}-\mathbf{B}\cdot\tilde{\mu}^{-1})\cdot(\delta\mathbf{B}) + (\mathbf{E}-\mathbf{D}\cdot\tilde{\epsilon}^{-1})\cdot(\delta\mathbf{D})\}\big|_{t=t_1}^{t=t_2}\,dV. \tag{15}$$

We must have that both (14) and (15) vanish in order to obtain the equivalence of (6) and (7).

The vanishing of the integrand in (14) at the spatial boundaries can be achieved in the usual manner. For example, if we were dealing with a wave packet in an unbounded region then the appropriate spatial boundary condition is the vanishing of all electromagnetic field vectors at large distances; if we were dealing with an electromagnetic field in a cavity with perfectly conducting walls then the appropriate spatial boundary condition is the vanishing of the tangential component of $\mathbf{E}$ (as well as $\mathbf{D}\cdot\tilde{\epsilon}^{-1}$). In either of these two examples the "natural" spatial boundary conditions are sufficient to make (14) vanish.

In the variational principle (6), we shall regard the time $t = t_1$ as the initial time. Thus we shall consider the four electromagnetic field vectors $(\mathbf{E},\mathbf{H},\mathbf{D},\mathbf{B})$ as being given at their argument $(\mathbf{r},t_1)$. The vanishing of the integrand in (15) at $t = t_1$ then occurs automatically if one imposes these initial conditions: for given initial fields there can be no variation $\delta\,\mathbf{B}$ and $\delta\,\mathbf{D}$ at $t = t_1$. The remaining condition which is required is the vanishing of the integrand in (15) at $t = t_2$. This cannot be done by fixing the fields when $t = t_2$ since these fields are unknown; in fact this is the very problem—we are seeking to predict the future fields ($t = t_2$) from the initial fields ($t = t_1$). Instead, we choose to invoke the (reasonable) *constraint* that the four electromagnetic field vectors obey the constitutive equations (4) when $t = t_2$. Clearly this is sufficient to make the integrand in (15) vanish at $t = t_2$.

In summary, the boundary and initial conditions which we propose are

(a) the usual "natural" spatial boundary conditions [e.g., the vanishing of the tangential component of $\mathbf{E}$ (and $\mathbf{D}\cdot\tilde{\epsilon}^{-1}$) for the case of perfectly conducting walls]; (16a)

(b) the usual initial conditions on the four electromagnetic field vectors $(\mathbf{E},\mathbf{H},\mathbf{D},\mathbf{B})$ at $t = t_1$; (16b)

(c) the constraint that the four electromagnetic field vectors obey the two constitutive equations ($\mathbf{D} = \tilde{\epsilon}\cdot\mathbf{E}$ and $\mathbf{B} = \tilde{\mu}\cdot\mathbf{H}$) at $t = t_2$. (16c)

V. SUMMARY

A variational principle for electromagnetic problems in a linear, static, inhomogeneous, and anisotropic medium has been formulated. This variational principle is

$$\delta\int_{t_1}^{t_2}\int_V L\,dV\,dt = 0, \tag{6}$$

where

$$L = (\mathbf{H}-\mathbf{B}\cdot\tilde{\mu}^{-1})\cdot\left(\nabla\times\mathbf{E} + \frac{\partial\mathbf{B}}{\partial t} + \mathbf{K}\right) - (\mathbf{E}-\mathbf{D}\cdot\tilde{\epsilon}^{-1})\cdot\left(\nabla\times\mathbf{H} - \frac{\partial\mathbf{D}}{\partial t} - \mathbf{J}\right). \tag{8}$$

All four electromagnetic field vectors $(\mathbf{E},\mathbf{H},\mathbf{D},\mathbf{B})$ are treated as independent for variational purposes. This leads to four vector Euler–Lagrange equations. Two of them are the two "curl" Maxwell equations:

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} - \mathbf{K}, \tag{1a}$$

$$\nabla\times\mathbf{H} = +\frac{\partial\mathbf{D}}{\partial t} + \mathbf{J}; \tag{1b}$$

the other two are the two constitutive equations:

$$\mathbf{D} = \tilde{\epsilon}\cdot\mathbf{E}, \tag{4a}$$

$$\mathbf{B} = \tilde{\mu}\cdot\mathbf{H}. \tag{4b}$$

The usual spatial boundary conditions (16a) and the usual initial conditions at $t = t_1$ (16b) are assumed. In addition, the four electromagnetic field vectors are assumed to satisfy the constraint of obeying the two constitutive equations at $t = t_2$ (16c).

Two features of this variational principle are worthy of note.

(i) All four electromagnetic field vectors $(\mathbf{E},\mathbf{H},\mathbf{D},\mathbf{B})$ are treated on an equal footing for variational purposes. No additional (artificial) "adjoint-type" fields are needed. Thus all four variational fields have a direct relevance in the actual problem.

(ii) Both the two "curl" Maxwell equations and the two constitutive equations arise as a result of the variational principle; in particular, the constitutive equations are treated on an equal footing with the Maxwell equations and are not invoked as an *a priori* constraint. The following attractive possibility then exists: by relaxing the requirement that the constitutive equations be exactly satisfied, the optimized trial electromagnetic fields may be able to satisfy more closely the Maxwell equations.

[1]K. Morishita and N. Kumagai, IEEE Trans. Microwave Theory Tech. **MTT-25**, 34 (1977).

[2]A. Mohsen, J. Phys. A **11**, 1681 (1978).

[3]A. Mohsen, J. Phys. A **15**, 1001 (1982).

[4]The reader who is not concerned with these fictitious quantities need only delete the symbol "**K**" in what follows.

[5]W. K. H. Panofsky and M. Phillips, *Classical Electricity and Magnetism* (Addison-Wesley, Reading, MA, 1955).

Lorentz transformation of Debye potentials

David K. Monroe
Department of Physics and Astronomy, Appalachian State University, Boone, North Carolina 28608

(Received 30 June 1983; accepted for publication 6 January 1984)

The Debye potential formalism is a simple and elegant method of representing and analyzing radiation fields in classical electromagnetism. Despite this, there are a number of unanswered questions about the formalism. One of these, the transformation properties of the potentials under Lorentz boosts, is dealt with in this article.

PACS numbers: 03.50.De

I. INTRODUCTION

Of all the commonly used formalisms for solving the source-free Maxwell equations, few can approach the Debye potential formalism[1] in the twin categories of economy and simplicity. The two Debye potentials concisely contain the two degrees of freedom per space-time point of the electromagnetic field. There are no extra components, no difficult vector decompositions, and no vector harmonics. All of the gauge freedom normally associated with the usual potentials $\mathbf{A}$ and Φ is conveniently "frozen out."

The Debye potentials are commonly referred to as "scalar" potentials. This is understandable because they transform as scalars under continuous rotations in 3-space, and they satisfy the scalar wave equation.[2] From the last property one might erroneously conclude that the potentials behave as 4-scalars under Lorentz boosts. Their actual behavior under these transformations is quite complex and seems not to have been clarified in any of the literature on electromagnetic radiation. The pupose of this paper is to address this issue.

In order to establish notational conventions, Sec. II will provide a brief review of the Debye potential formalism. Important identities needed later in this article will be derived. Section III contains the derivations of the transformation properties of the Debye potentials under Lorentz boosts. These will be presented in 3-vector notation. Because of the way in which the Debye potentials are defined, little notational simplicity is obtained by casting these results in 4-vector formalism.

II. THE DEBYE POTENTIAL FORMALISM

Although the Debye potential approach is applicable to electrodynamics in all situations, it is simplest when sources and matter are absent. These conditions are normally obtained when treating the radiation fields of antennas at points far from the localized charge and current distributions producing the radiation. Assuming no sources, dielectrics, or magnetic materials, Maxwell's equations in CGS units reduce to[3]

$$\nabla\cdot\mathbf{E} = 0, \tag{1}$$

$$\nabla\cdot\mathbf{B} = 0, \tag{2}$$

$$\nabla\times\mathbf{E} + \partial_0\mathbf{B} = 0, \tag{3}$$

$$\nabla\times\mathbf{B} - \partial_0\mathbf{E} = 0, \tag{4}$$

where $\partial_0 = c^{-1}\partial_t$. The Debye potential formalism consists of expressing the electric and magnetic fields in the form[4]

$$\mathbf{E} = \mathbf{L}\partial_0\phi + \mathbf{M}\psi, \tag{5}$$

$$\mathbf{B} = \mathbf{L}\partial_0\psi - \mathbf{M}\phi, \tag{6}$$

where $\mathbf{L}$ and $\mathbf{M}$ are vector operations defined by

$$\mathbf{L} = \nabla\times\mathbf{r} = -\mathbf{r}\times\nabla, \tag{7}$$

$$\mathbf{M} = \nabla\times\mathbf{L}. \tag{8}$$

The functions ϕ and ψ are the Debye potentials.

The operator $\mathbf{L}$ is proportional to the angular momentum operator in quantum mechanics, and so many of its properties are familiar. It behaves as a 3-vector under continuous rotations in 3-space, but exposes its true pseudovector nature under reflections.[5] Because the electric field $\mathbf{E}$ is a 3-vector, it is apparent from Eq. (5) that ϕ must be a pseudoscalar under transformations in 3-space. From Eq. (7) it is easily shown that

$$\nabla\cdot\mathbf{L} = 0, \tag{9}$$

$$\mathbf{r}\cdot\mathbf{L} = 0. \tag{10}$$

The second operator $\mathbf{M}$ is the curl of $\mathbf{L}$ but can be expressed in other convenient forms as well, for example,

$$\mathbf{M} = -\mathbf{r}\Delta + \nabla(\mathbf{r}\cdot\nabla) + \nabla, \tag{11}$$

where $\Delta = \nabla\cdot\nabla$ is the Laplacian. Under rotations and reflections in 3-space, $\mathbf{M}$ transforms as an ordinary 3-vector. Since $\mathbf{E}$ is a 3-vector, one sees from Eq. (5) that ψ is a 3-scalar. Useful identities satisfied by $\mathbf{M}$ are

$$\nabla\cdot\mathbf{M} = 0, \tag{12}$$

$$\mathbf{r}\cdot\mathbf{M} = -\mathbf{L}\cdot\mathbf{L}, \tag{13}$$

$$\mathbf{L}\cdot\mathbf{M} = 0. \tag{14}$$

The elegance of the Debye potential formalism is apparent with the observation that $\mathbf{E}$ and $\mathbf{B}$ in Eqs. (5) and (6) satisfy the source-free Maxwell equations if and only if the Debye potentials ϕ and ψ satisfy the scalar wave equations:

$$\Box\phi = 0, \tag{15}$$

$$\Box\psi = 0, \tag{16}$$

where $\Box = -\partial_0^2 + \Delta$ is the d'Alembertian. These simple results are immediate consequences of the fact that $\mathbf{L}$ and $\mathbf{M}$ commute with the Laplacian:[6]

$$\mathbf{L}\Delta = \Delta\mathbf{L}, \tag{17}$$

$$\mathbf{M}\Delta = \Delta\mathbf{M}. \tag{18}$$

A very important joint property of the operators $\mathbf{L}$ and

 0022-2488/84/061787-04$02.50

$\mathbf{M}$ is that they are orthogonal. For arbitrary functions F and G

$$\int_V (\mathbf{L}F)\cdot(\mathbf{M}G)\, dv = 0, \tag{19}$$

where V includes all of 3-space. The validity of this result is a direct consequence of the divergence theorem, which allows the volume integral to be written as

$$\oint_S (F\mathbf{r})\times(\mathbf{M}G)\cdot\hat{n}\, da,$$

with the help of Eq. (10). Taking S to be a spherical surface expanding to spatial infinity, one immediately obtains Eq. (19) since $\hat{n} = \hat{r}$.

III. TRANSFORMATION OF THE DEBYE POTENTIALS UNDER BOOSTS

The Lorentz transformation properties of the electric and magnetic fields are, of course, well known, so one should in principle be able to derive the transformation equations for the Debye potentials directly for Eqs. (5) and (6). In practice this approach is difficult because the transformation equations for $\mathbf{M}$ are unwieldy. A simpler strategy begins by using Eqs. (10) and (13) to express the Debye potentials in terms of simple dot products:

$$\mathbf{r}\cdot\mathbf{E} = -L^2\psi, \tag{20}$$

$$\mathbf{r}\cdot\mathbf{B} = +L^2\phi. \tag{21}$$

The scalar operator $L^2 = \mathbf{L}\cdot\mathbf{L}$ will occur so frequently in conjunction with the Debye potentials that it is useful to define alternate potentials:

$$\hat{\phi} \equiv L^2\phi, \tag{22}$$

$$\hat{\psi} \equiv L^2\psi. \tag{23}$$

We must now relate unprimed quantities measured in the lab frame $\mathscr{F}$ to corresponding quantities in the moving frame $\mathscr{F}'$. If $\mathscr{F}'$ moves with velocity $\mathbf{v} = \boldsymbol{\beta}c = \beta c\hat{i}$ so that the x' axis is conlinear with the x axis and the origins coincide at $t' = t = 0$, then[7]

$$x^0 = \gamma(x^{0'} + \beta x^{1'}), \tag{24}$$

$$x^1 = \gamma(x^{1'} + \beta x^{0'}), \tag{25}$$

$$x^2 = x^{2'}, \tag{26}$$

$$x^3 = x^{3'}, \tag{27}$$

where $x^0 = ct$, $x^1 = x$, $x^2 = y$, and $x^3 = z$. The fields $\mathbf{E}$ and $\mathbf{B}$ then transform according to:

$$E_1 = E_{1'}, \tag{28}$$

$$E_2 = \gamma(E_{2'} + \beta B_{3'}), \tag{29}$$

$$E_3 = \gamma(E_{3'} - \beta B_{2'}), \tag{30}$$

$$B_1 = B_{1'}, \tag{31}$$

$$B_2 = \gamma(B_{2'} - \beta E_{3'}), \tag{32}$$

$$B_3 = \gamma(B_{3'} + \beta E_{2'}). \tag{33}$$

With these well-known results we can now obtain expressions for $\hat{\phi}$ and $\hat{\psi}$ in terms of primed quantities. Using Eqs. (25)–(27) and Eqs. (31)–(33), we find that Eq. (22) becomes

$$\hat{\phi} = \gamma[(x^{1'} + \beta x^{0'})B_{1'} + x^{2'}(B_{2'} - \beta E_{3'}) + x^{3'}(B_{3'} + \beta E_{2'}),$$

$$\hat{\phi} = \gamma[\mathbf{r}'\cdot\mathbf{B}' + \beta(x^{0'}B_{1'} - x^{2'}E_{3'} + x^{3'}E_{2'})], \tag{34}$$

$$\hat{\phi} = \gamma[\hat{\phi}' + \boldsymbol{\beta}\cdot(x^{0'}\mathbf{B}' - \mathbf{r}'\times\mathbf{E}')],$$

where $\hat{\phi}' \equiv \mathbf{L}'\cdot\mathbf{L}'\phi'$ is defined in the primed frame just as $\hat{\phi}$ is defined in Eq. (22). Likewise one obtains for $\hat{\psi}$

$$\hat{\psi} = \gamma[\hat{\psi}' - \boldsymbol{\beta}\cdot(x^{0'}\mathbf{E}' + \mathbf{r}'\times\mathbf{B}')], \tag{35}$$

where $\hat{\psi}' \equiv \mathbf{L}'\cdot\mathbf{L}'\psi'$ is defined as in Eq. (23).

If observers in $\mathscr{F}'$ express the fields $\mathbf{E}'$ and $\mathbf{B}'$ in terms of Debye potentials ϕ' and ψ', then we have in accordance with Eqs. (5) and (6)

$$\mathbf{E}' = \mathbf{L}'\partial_{0'}\phi' + \mathbf{M}'\psi', \tag{36}$$

$$\mathbf{B}' = \mathbf{L}'\partial_{0'}\psi' - \mathbf{M}'\phi'. \tag{37}$$

The operators $\mathbf{L}'$ and $\mathbf{M}'$ are defined similar to Eqs. (7) and (8). We can now eliminate $\mathbf{E}'$ and $\mathbf{B}'$ from Eqs. (34) to obtain

$$\hat{\phi} = \gamma\{\hat{\phi}' - \boldsymbol{\beta}\cdot[(x^{0'}\mathbf{M}' + \mathbf{r}'\times\mathbf{L}'\partial_{0'})\phi' - (x^{0'}\mathbf{L}'\partial_{0'} - \mathbf{r}'\times\mathbf{M}')\psi']\},$$

$$\hat{\phi} = \gamma\{\hat{\phi}' - \boldsymbol{\beta}\cdot[(x^{0'}\boldsymbol{\nabla}' + \partial_{0'}\mathbf{r}')\times\mathbf{L}'\phi' - (x^{\alpha'}\partial_{\alpha'} + 1)\mathbf{L}'\psi']\}, \tag{38}$$

where, of course,

$$x^{\alpha'}\partial_{\alpha'} \equiv x^{0'}\partial_{0'} + x^{a'}\partial_{a'}.$$

In like fashion the expression for $\hat{\psi}$ in Eq. (35) can be reexpressed as

$$\hat{\psi} = \gamma\{\hat{\psi}' - \boldsymbol{\beta}\cdot[(x^{0'}\partial_{0'}\mathbf{L}' - \mathbf{r}'\times\mathbf{M}')\phi' + (x^{0'}\mathbf{M}' + \mathbf{r}'\times\mathbf{L}'\partial_{0'})\psi']\}, \tag{39}$$

$$\hat{\psi} = \gamma\{\hat{\psi}' - \boldsymbol{\beta}\cdot[(x^{0'}\boldsymbol{\nabla}' + \partial_{0'}\mathbf{r}')\times\mathbf{L}'\psi' + (x^{\alpha'}\partial_{\alpha'} + 1)\mathbf{L}'\phi']\}.$$

Note that Eqs. (38) and (39) are equivalent under the exchange transformation:

$$\phi\to\psi, \quad \psi\to-\phi. \tag{40}$$

This property can be traced back to a similar exchange symmetry with regard to $\mathbf{E}$ and $\mathbf{B}$ in Maxwell's equations without sources.

One may regard Eqs. (38) and (39) as the transformation equations for the Debye potentials. Given the potentials ϕ' and ψ' determined in $\mathscr{F}'$, we can calculate $\hat{\phi}'$ and $\hat{\psi}'$ using Eqs. (22) and (23) expressed in primed coordinates, and then obtain $\hat{\phi}$ and $\hat{\psi}$ from the equations above. The alternate potentials $\hat{\phi}$ and $\hat{\psi}$ contain all the information carried by ϕ and ψ, except for additive functions of $r = |\mathbf{r}|$, which are annihilated by the operator $L^2 = \mathbf{L}\cdot\mathbf{L}$. However, these same functions would be annihilated by $\mathbf{L}$ and $\mathbf{M}$ as well, so they cannot contribute to the fields $\mathbf{E}$ and $\mathbf{B}$. In the language of multipole analysis, these additive functions are pure monopole terms which do not pertain to radiation fields anyway.

The transformations given by Eqs. (38) and (39) clearly indicate that the Debye potentials are not Lorentz scalars. Their complex transformation properties under boosts can, of course, be traced back to their definitions in terms of the operators $\mathbf{L}$ and $\mathbf{M}$, and the central role played by the three-vector $\mathbf{r}$ in each of these vectors. No significant simplification can be achieved, even with the elimination of the superfluous operator $L^2 = \mathbf{L}\cdot\mathbf{L}$ contained in the potentials $\hat{\phi}$ and $\hat{\psi}$, as will now be shown.

The exchange symmetry in Eqs. (40) is of great utility at this point. Any reduction of Eq. (38) will be imitated by a similar reduction of Eq. (39), but with ϕ replaced by ψ and ψ replaced by $-\phi$, simultaneously. Clearly, then, it suffices to reduce only one of these, say Eq. (38). In component form this equation becomes

$$\begin{aligned}\hat{\phi} &= \gamma[L_{1'} L_{1'}\phi' + \beta(x^{\alpha'}\partial_{\alpha'} + 1)L_{1'}\psi'] \\ &\quad + \gamma[L_{2'} + \beta(\partial_{0'} x^{3'} + \partial_{3'} x^{0'})]L_{2'}\phi' \\ &\quad + \gamma[L_{3'} - \beta(\partial_{0'} x^{2'} + \partial_{2'} x^{0'})]L_{3'}\phi'.\end{aligned}$$

Because $\mathbf{L}'$ and $x^{\alpha'}\partial_{\alpha'}$ commute, we get

$$\begin{aligned}\hat{\phi} &= \gamma L_{1'}[L_{1'}\phi' + \beta(x^{\alpha'}\partial_{\alpha'} + 1)\psi'] \\ &\quad + \gamma[L_{2'} + \beta(\partial_{0'} x^{3'} + \partial_{3'} x^{0'})]L_{2'}\phi' \\ &\quad + \gamma[L_{3'} - \beta(\partial_{0'} x^{2'} + \partial_{2'} x^{0'})]L_{3'}\phi'. \end{aligned} \tag{41}$$

We now need the transformation equations for the components of $\mathbf{L}$ under Lorentz boosts. Using Eq. (7) and Eqs. (25)–(27) we readily find that[8]

$$L_1 = L_{1'}, \tag{42}$$

$$L_2 = \gamma[L_{2'} + \beta(\partial_{0'} x^{3'} + \partial_{3'} x^{0'})], \tag{43}$$

$$L_3 = \gamma[L_{3'} - \beta(\partial_{0'} x^{2'} + \partial_{2'} x^{0'})]. \tag{44}$$

We can then reduce Eq. (41) to

$$\begin{aligned}\hat{\phi} &= L_1\gamma[L_1\phi' + \beta(x^{\alpha}\partial_{\alpha} + 1)\psi'] \\ &\quad + L_2 L_{2'}\phi' + L_3 L_{3'}\phi'.\end{aligned} \tag{45}$$

Utilizing Eq. (22), we get

$$\begin{aligned}0 &= L_1[L_1\phi - \gamma L_1\phi' - \beta\gamma(x^{\alpha}\partial_{\alpha} + 1)\psi'] \\ &\quad + L_2[L_2\phi - L_{2'}\phi'] + L_3[L_3\phi - L_{3'}\phi'].\end{aligned}$$

Again, applying Eqs. (43) and (44),

$$\begin{aligned}0 &= L_1[L_1(\phi - \gamma\phi') - \beta\gamma(x^{\alpha}\partial_{\alpha} + 1)\psi'] \\ &\quad + L_2[L_2(\phi - \gamma\phi') + \beta\gamma(\partial_0 x^3 + \partial_3 x^0)\phi'] \\ &\quad + L_3[L_3(\phi - \gamma\phi') - \beta\gamma(\partial_0 x^2 + \partial_2 x^0)\phi']\end{aligned}$$

or

$$\begin{aligned}0 &= L^2(\phi - \gamma\phi') - \gamma\boldsymbol{\beta}\cdot\mathbf{L}(x^{\alpha}\partial_{\alpha} + 1)\psi' \\ &\quad + \gamma\boldsymbol{\beta}\times\mathbf{L}\cdot(\partial_0\mathbf{r} + x^0\ \nabla)\phi'.\end{aligned} \tag{46}$$

The last result may also be regarded as a transformation equation for the potential ϕ, although it is not in standard form with primed quantities on one side of the equation and unprimed ones on the other. To apply it, one must express the primed arguments of ϕ' and ψ' in terms of the unprimed ones, using Eqs. (24)–(27). Also, we observe from Eq. (46) that the scalar operator L^2 cannot be eliminated from the transformation equation without introducing its inverse elsewhere. Our last task will be to see what the result of this further reduction is.

There are three distinct terms in Eq. (46). The first presents no problem in our efforts to eliminate L^2, since it contains this operator explicitly. The second term,

$$-\gamma\boldsymbol{\beta}\cdot\mathbf{L}(x^{\alpha}\partial_{\alpha} + 1)\psi',$$

requires that we introduce L^2 and its inverse, denoted by L^{-2}, as follows:

$$-\gamma\boldsymbol{\beta}\cdot\mathbf{L}(x^{\alpha}\partial_{\alpha} + 1)L^2(L^{-2}\psi').$$

Because L^2 commutes with $x^{\alpha}\partial_{\alpha}$ and $\mathbf{L}$, we find that this last expression becomes

$$-L^2\gamma\boldsymbol{\beta}\cdot\mathbf{L}(x^{\alpha}\partial_{\alpha} + 1)(L^{-2}\psi'),$$

and so Eq. (46) becomes

$$\begin{aligned}0 &= L^2[(\phi - \gamma\phi') - \gamma\boldsymbol{\beta}\cdot\mathbf{L}(x^{\alpha}\partial_{\alpha} + 1)(L^{-2}\psi')] \\ &\quad + \gamma\boldsymbol{\beta}\times\mathbf{L}\cdot(\partial_0\mathbf{r} + x^0\ \nabla)\phi'.\end{aligned} \tag{47}$$

The last term in Eq. (47) can be dealt with in a similar fashion:

$$\gamma\boldsymbol{\beta}\times\mathbf{L}\cdot(\partial_0\mathbf{r} + x^0\ \nabla)L^2(L^{-2}\phi').$$

It is not too difficult to show that

$$\boldsymbol{\beta}\times\mathbf{L}\cdot\mathbf{r}L^2 = -L^2\boldsymbol{\beta}\times\mathbf{r}\cdot\mathbf{L} \tag{48}$$

and likewise that

$$\boldsymbol{\beta}\times\mathbf{L}\cdot\nabla L^2 = -L^2\boldsymbol{\beta}\times\nabla\cdot\mathbf{L}, \tag{49}$$

$$\boldsymbol{\beta}\times\mathbf{L}\cdot\nabla L^2 = -L^2\boldsymbol{\beta}\cdot\mathbf{M}. \tag{50}$$

These identities allow one to transform that last term of Eq. (47) into

$$-L^2\gamma\boldsymbol{\beta}\times(\partial_0\mathbf{r} + x^0\ \nabla)\cdot\mathbf{L}\phi',$$

and so Eq. (47) becomes

$$\begin{aligned}0 &= L^2[(\phi - \gamma\phi') - \gamma\boldsymbol{\beta}\cdot\mathbf{L}(x^{\alpha}\partial_{\alpha} + 1)(L^{-2}\psi') \\ &\quad - \gamma\boldsymbol{\beta}\times(\partial_0\mathbf{r} + x^0\ \nabla)\cdot\mathbf{L}(L^{-2}\phi')],\end{aligned}$$

from which we conclude that

$$\begin{aligned}\phi &= \gamma\{\phi' + \boldsymbol{\beta}\cdot[(x^{\alpha}\partial_{\alpha} + 1)\mathbf{L}(L^{-2}\psi') \\ &\quad + (\partial_0\mathbf{r} + x^0\ \nabla)\times\mathbf{L}(L^{-2}\phi')]\}.\end{aligned} \tag{51}$$

Using the exchange symmetry in Eqs. (40), we immediately obtain the corresponding equation for the potential ψ:

$$\begin{aligned}\psi &= \gamma\{\psi' - \boldsymbol{\beta}\cdot[(x^{\alpha}\partial_{\alpha} + 1)\mathbf{L}(L^{-2}\phi') \\ &\quad - (\partial_0\mathbf{r} + x^0\ \nabla)\times\mathbf{L}(L^{-2}\psi')]\}.\end{aligned}$$

Like Eq. (46), the last two results may be regarded as transformation equations for the Debye potentials, even though they are not in standard form. The biggest obstacle to applying them is dealing with the inverse operator L^{-2} which acts upon unprimed coordinates. The primed arguments of the potentials ϕ' and ψ' must be replaced by unprimed arguments in accordance with the Lorentz transformation in Eqs. (24)–(27).

IV. CONCLUSIONS

The Debye potential formalism is ideally suited to analyzing the electromagnetic radiation produced by localized sources. The potentials transform as scalars under rotations in 3-space, but display more complex behavior under Lorentz boosts. Several expressions for their transformation properties have been obtained, with those in Eqs. (38) and (39) being the most natural.

It was observed in Sec. II that if the fields $\mathbf{E}$ and $\mathbf{B}$ satisfy the source-free Maxwell equations, then the potentials must satisfy

$$\Box\phi = 0,$$

$$\Box\psi = 0.$$

The transformation equations obtained for the Debye potentials guarantee that these equations will be satisfied in all Lorentz frames if they are satisfied in one frame.

[1]See, for example, A. Nisbet, Proc. R. Soc. (London), Ser. A **231**, 250 (1955); J. Cohen and L. Kegels, Phys. Rev. D **10**, 1070 (1974).

[2]These points will be clarified in Sec. II of this article.

[3]See, for example, J. Jackson, *Classical Electrodynamics* (Wiley, New York, 1962), Chap. 7.

[4]These simple equations apply only to the case of no sources and no matter.

[5]Under reflections $\mathbf{r} \to \mathbf{r}$, the potential ϕ transforms as $\phi \to -\phi$.

[6]Equations (5) and (6) obviously satisfy the divergence conditions (1) and (2). The wave equations arise as a result of applying Eqs. (3) and (4) to $\mathbf{E}$ and $\mathbf{B}$.

[7]In what follows all indices will be manipulated with the Lorentz metric $\eta_{\mu\nu}$ whose signature is $\eta = [-1,1,1,1]$.

[8]One also needs to use the corresponding covarient transformations of the partial derivatives.

Description of the degree of nonuniqueness in inverse source problems

A. G. Ramm
Mathematics Department, Kansas State University, Manhattan, Kansas 66506

(Received 5 April 1983; accepted for publication 14 October 1983)

For time-independent and time-dependent inverse source problems the degree of nonuniqueness of solutions is characterized.

PACS numbers: 03.50.De, 03.65.Nk, 03.80. + r

I. INTRODUCTION

1. Recently a number of authors discussed the inverse source problems[1–5] (see also the bibliography in Refs. 1 and 5). The problem under discussion is as follows. Let

$$(\nabla^2 + k^2)u = h, \quad x \in \mathbb{R}^3; \quad h = 0, \quad |x| \geqslant a. \tag{1}$$

Suppose the Cauchy data is given for u near infinity:

$$u|_{|x|=R} = u_0, \quad \frac{\partial u}{\partial r}\bigg|_{|x|=R} = u_1, \quad R \gg 1. \tag{2}$$

Can one determine h from the data (2)? (3)

Most of the authors ask a slightly different but equivalent question. Namely, can one determine h from the radiation pattern $f(\theta,\phi,k)$, which is defined by the formula

$$f(\theta,\phi,k) = \lim_{r\to+\infty} \left(\frac{e^{ikr}}{4\pi r}\right)^{-1} u(x,k),$$
$$r = |x|, \quad x = (r,\theta,\phi), \tag{4}$$

where u is the (unique) solution to (1) satisfying the radiation condition? The equivalence of the data (3) and (4) comes from the fact that the asympotics of u at infinity can be differentiated. On a more formal level, the Cauchy data (3) defines the solution of the homogeneous equation (1) uniquely in the domain $|x| \geqslant a$, and the radiation pattern (4) does the same (because of the Rellich's uniqueness theorem, see, e.g., Ref. 6). The problem (3) is the time-independent inverse source problem.

2. This problem for Maxwell's equations was discussed in detail in Ref. 7 (see also Ref. 8, pp. 208–211). The problem for Maxwell equations has some features which the scalar problem (1)–(4) does not have. In particular, the radiation patterns for Maxwell's equations is a two-component vector field defined on the unit sphere $S^2 \subset \mathbb{R}^3$ and tangent to this sphere (it is two-component in the spherical coordinates: $f_r = 0$). On the other hand, the source $j(x)$ in Maxwell's equations:

$$\nabla \times E = ikH, \quad \mathrm{rot}\, H = -ikE + j, \quad \epsilon = \mu = 1,$$

is a three-component vector. The corresponding radiation pattern is

$$f = ik\,(I_\phi a_\phi + I_\theta a_\theta), \tag{5}$$

where a_r, a_θ, a_ϕ are the unit vectors of the spherical coordinates at the point $x = (r,\theta,\phi)$, $I = (I_r, I_\theta, I_\phi)$,

$$I = \int \exp\{-i(\mathbf{k}, y)\}\, j(y)dy, \quad \mathbf{k} = (k,\theta,\phi), \quad \int \equiv \int_{\mathbb{R}^3}.$$

It is clear from (5) that the radiation pattern $f(k,\theta,\phi)$ given for a fixed $k = k_0$ as a vector field on the unit sphere does not determine I uniquely, and therefore does not determine the sources j uniquely. The degree of nonuniqueness can be described as follows: in order to determine the sources uniquely one should specify: (1) one scalar function $I_r(k_0,\theta,\phi)$, and (2) three functions $f_\theta(k,\theta,\phi)$, $f_\phi(k,\theta,\phi)$, $I_r(k,\theta,\phi)$, $0 < k < \infty$, which are equal to the functions $f_\theta(k_0,\theta,\phi)$, $f_\phi(k_0,\theta,\phi)$, $I_r(k_0,\theta,\phi)$ at $k = k_0$. Here we did not impose the important requirement that the sources have compact support. If this requirement is imposed then the functions $I_r(k,\theta,\phi)$, $f_\theta(k,\theta,\phi)$, $f_\phi(k,\theta,\phi)$ should satisfy the condition that the vector function $I = I_r a_r + (1/ik) f_\phi a_\phi + (1/ik) f_\theta a_\theta$ be an entire function of $\mathbf{k}$ of exponential type, i.e., $|I| \leqslant c_1 \exp(c_2 k)$ for some constants $c_1, c_2 > 0$. This requirement comes from the Paley–Wiener theorem which says that a function $I(\mathbf{k})$ is a Fourier transform of a function j with compact support iff I is an entire function of exponential type. In addition, $j \in L^2$ if $I \in L^2(\mathbb{R}^3)$. This analysis, a computational scheme for finding j from the radiation pattern, and some examples are given in Ref. 7.

One should have in mind that in some concrete inverse source problems (e.g., synthesis of linear or spherical antennas) the uniqueness of the solution holds because of some special assumptions about the sources. There is an extensive literature on this subject (see, e.g., Ref. 9). In order to explain how the uniqueness holds, consider the linear antenna synthesis problem. The sources are currents along the line segment $-l < z < l$. The currents are defined by one scalar function $j(z)$ in this case. The radiation pattern is proportional to $\int_{-l}^{l} \exp(ik_0 z \cos\theta)\, j(z)dz \equiv f(\theta)$, $0 \leqslant \theta \leqslant \pi$. This can be written as $\int_{-l}^{l} \exp(i\lambda z)\, j(z)dz = f(\lambda)$, $-k_0 \leqslant \lambda \leqslant k_0$, $\lambda = k_0 \cos\theta$. The last equation clearly has no more than one solution. It is solvable iff $f(\lambda)$ is an entire function of exponential type $\leqslant l$ of λ.

3. In the scalar case one has

$$f = -\frac{1}{4\pi} \int \exp\{-i(\mathbf{k}, y)\}\, h(y)dy.$$

Thus, the knowledge of $f(k_0,\theta,\phi)$ does not determine the sources $h(y)$ uniquely. The degree of nonuniqueness can be described as follows: given the radiation pattern at $k = k_0$ one can fix a (arbitrary continuous in k) function $f(k,\theta,\phi)$ for $0 < k < \infty$ such that $f(k,\theta,\phi)|_{k=k_0} = f(k_0,\theta,\phi)$. These data determine the Fourier transform of the sources and there-

 0022-2488/84/061791-03$02.50

fore the sources $h(y)$ uniquely. In this argument we did not impose any *a priori* conditions on the sources. If one assumes (as is natural) that $h \in L^2$ and $h = 0$ for $|x| \geqslant a$, then the radiation pattern is an entire function of $\mathbf{k}$ of exponential type and the corresponding extension $f(k,\theta,\phi)$ should satisfy this necessary requirement. This requirement is also sufficient for $h \in L^2$ and $h = 0$, $|x| \geqslant a$ if $f \in L^2$. This is a complete description of the degree of nonuniqueness in the scalar problem of finding the sources from the radiation pattern given at a fixed frequency.

4. The reason why the problem (3) is discussed here is that the similar problem arises in time-dependent cases. Consider, for example, the following time-dependent inverse source problem. Let

$$u_{tt} - \Delta u = f(x,t), \quad -\infty < t < \infty, \quad x \in \mathbb{R}^3,$$
$$f = 0 \quad \text{of } t \notin (0,T). \tag{6}$$

Suppose that

$$u(0,x) = u_0(x), \quad u_t(0,x) = u_1(x), \tag{7}$$
$$u(T,x) = v_0(x), \quad u_t(T,x) = v_1(x). \tag{8}$$

Can one find the sources $f(x,t)$ from the data (7), (8)? (9)

Because of the uniqueness of the Cauchy problem for Eq. (6) the data (7), (8) determine the solution of (7) for $t < 0$ $(t > T)$ uniquely. In Sec. II the description of the degree of nonuniqueness of the solution to problems (3) and (9) are given.

II. DESCRIPTION OF THE DEGREE OF NONUNIQUENESS IN THE INVERSE SOURCE PROBLEMS

1. Consider first problem (9). Taking the Fourier transform in x, one obtains

$$\ddot{w} + k^2 w = g(t,\mathbf{k}),$$
$$w = \frac{1}{(2\pi)^3} \int u(x,t) \exp\{-i(\mathbf{k},x)\} dx \equiv \tilde{u}, \quad g \equiv \tilde{f}, \tag{10}$$
$$w(0,\mathbf{k}) = w_0 \equiv \tilde{u}_0, \quad \dot{w}(0,\mathbf{k}) = w_1 \equiv \tilde{u}_1, \quad \dot{w} = \frac{dw}{dt}, \tag{11}$$
$$w(T,\mathbf{k}) = p_0 \equiv \tilde{v}_0, \quad \dot{w}(T,\mathbf{k}) = p_1 \equiv \tilde{v}_1. \tag{12}$$

Here $\mathbf{k} = (k_1,k_2,k_3)$, $k^2 = k_1^2 + k_2^2 + k_3^2$, $k = |\mathbf{k}|$. The problem (10), (11) can be solved explicitly,

$$w = \int_0^t \frac{\sin\{k(t-s)\}}{k} g(s,\mathbf{k}) ds + w_0 \cos(kt) + w_1 \frac{\sin kt}{k}. \tag{13}$$

From (13) and (12) one obtains

$$p_0 = \int_0^T \frac{\sin\{k(T-s)\}}{k} g\, ds + w_0 \cos(kT) + w_1 \frac{\sin(kT)}{k},$$
$$p_1 = \int_0^T \cos\{k(T-s)\} g\, ds - k w_0 \sin(kT) + w_1 \cos kT. \tag{14}$$

This can be written as

$$\frac{\sin(kT)}{k} \int_0^T \cos(ks) g\, ds - \cos(kT) \times \int_0^T \sin(ks) g\, ds = b_0(k),$$
$$\frac{\cos(kT)}{k} \int_0^T \cos(ks) g\, ds + \sin(kT) \times \int_0^T \sin(ks) g\, ds = b_1(\mathbf{k}), \tag{15}$$

where b_0, b_1 are known explicitly:

$$b_0 = p_0 - w_0 \cos(kT) - (w_1/k) \sin kT,$$
$$b_1 = p_1 + k\omega_0 \sin(kT) - w_1 \cos(kT). \tag{16}$$

The determinant of the system (15) is $1/k > 0$. Thus, one can uniquely find from (15) the two integrals

$$\int_0^T \cos(ks) g(s,\mathbf{k}) ds = \phi_1(\mathbf{k}),$$
$$\int_0^T \sin(ks) g(s,\mathbf{k}) ds = \phi_2(\mathbf{k}), \tag{17}$$

where $\phi_1(\mathbf{k})$ and $\phi_2(\mathbf{k})$ are given explicitly.

It is now possible to describe the degree of nonuniqueness of the solution to the inverse source problem (9). Namely, the data (11), (12) determine ϕ_1 and ϕ_2 in (17). Equations (17) determine g, the Fourier transform of the sources, of the form

$$g = c_1(\mathbf{k}) \cos(ks) + c_2(\mathbf{k}) \sin(ks) + g_1(s,\mathbf{k}), \tag{18}$$

where $c_1(\mathbf{k})$ and $c_2(\mathbf{k})$ are uniquely determined by ϕ_1 and ϕ_2, while $g_1(s,k)$ is an arbitrary function orthogonal to $\cos(ks)$ and $\sin(ks)$ in $L^2([0,T])$:

$$\int_0^T \sin(ks) g_1(s,\mathbf{k}) ds = \int_0^T \cos(ks) g_1(s,\mathbf{k}) ds = 0 \quad \forall\, k. \tag{19}$$

This is a complete description of the degree of nonuniqueness of the solution to problem (9) in the case when no conditions are imposed on the sources $f(x,t)$ except the last condition (10) and, say, the mild requirement like $f(x,t) \in L^2(\mathbb{R}^3)$ for any t and is of class C^2 in time. If one requires additionally that $f(x,t) = 0$ for $|x| \geqslant a$ (the sources have compact support), then g_1 should satisfy not only Eqs. (19), but also ensure that the function g defined by formula (18) be an entire function of exponential type in the variables $\mathbf{k} = (k_1,k_2,k_3)$. Since g determines the sources f uniquely, one can say that the solution of problem (9) is defined nonuniquely up to an additive term orthogonal to a two-dimensional subspace. The method of this section can be applied without any changes to the abstract time-dependent inverse source problem. This problem is similar to the problem (6)–(9), but Eq. (6) is of the form $u_{tt} + Au = f(t)$, $-\infty < t < \infty$, where $A > 0$ is a self-adjoint operator on a Hilbert space $h, f(t) \in H$. One assumes that the eigenfunction expansion theorem for A is known and the eigenfunctions satisfy the equation $A\phi = k^2\phi$, and obtains the same equations (10)–(12) in which w is the coefficient of the expansion of u in the eigenfunctions ϕ (analog of the Fourier transform). The rest of the argument is unchanged.

2. Let us consider problem (1)–(3). The Cauchy data (2) determines uniquely the solution to Eq. (1) in the domain $|x|\geqslant a$ in which (1) is homogeneous. Let

$$u|_{r=a}=\psi_0, \quad \frac{\partial u}{\partial r}\bigg|_{r=a}=\psi_1, \tag{20}$$

where ψ_0 and ψ_1 are uniquely determined by u_0 and u_1. (Of course, as is well known, the data u_0 and u_1 for an elliptic equation cannot be given arbitrarily.) One can also find ψ_0 and ψ_1 from the knowledge of the radiation pattern. Let $G(x,y)$ be the Green's function $(\nabla^2+k^2)G=-\delta(x-y)$, $|x|\leqslant a$, $G|_{|x|=a}=0$. Then, the solution of the problem

$$(\nabla^2+k^2)u=h, \quad |x|\leqslant a, \quad u|_{|x|=a}=\psi_0 \tag{21}$$

can be found explicitly provided that k^2 is not an eigenvalue of the Dirichlet Laplacian in the domain $|x|\leqslant a$. Namely,

$$u=\int_{\mathscr{D}} G(x,y)h\,dy+\int_{\Gamma}\frac{\partial G(x,s)}{\partial N}\psi_0\,ds, \quad x\in\mathscr{D}, \tag{22}$$

where $\mathscr{D}=\{x:|x|\leqslant a\}$, $\Gamma=\{x:|x|=a\}$, $N=N_s$ is the outer unit normal to Γ at the point s. From (22) and the second condition (20) one obtains an equation for h. Since the problem is linear and one is interested in the description of nonuniqueness, one can take $\psi_0=\psi_1=0$. In this case the equation for h which follows from (20) and (22) is of the form

$$0=\int_{\mathscr{D}}\frac{\partial G(s,y)}{\partial N}h\,dy, \quad s\in\Gamma. \tag{23}$$

To describe the degree of nonuniqueness of the solution to problem (3) one should describe the set of solutions to Eq. (23). If $\psi_0=0$ and $\psi_1=0$ then $u\equiv 0$ in $|x|\geqslant a$. Therefore one describes the nonradiating sources. This question was often discussed in the physical literature.[5] The following is a description of the set of all solutions of Eq. (23). Let us multiply (23) by a smooth function $\phi(s)$ and integrate over Γ to obtain

$$0=\int_{\mathscr{D}} dy\, h(y)F(y), \tag{24}$$

where $F(y)=\int_\Gamma(\partial G(s,y)/\partial N)\phi(s)ds$ is the solution of the Dirichlet problem,

$$(\nabla^2+k^2)F=0 \quad \text{in } \mathscr{D}, \tag{25}$$

$$F|_\Gamma=-\phi. \tag{26}$$

Therefore F runs through the set of functions satisfying the homogeneous Helmholtz's equation in $\mathscr{D}$. Equations (23) and (24) are equivalent. Thus one obtains the following statement. The set of all solutions of (23) (i.e., the set of all nonradiating sources) is precisely the set of functions orthogonal to all of the solutions (in L^2 say) of Eq. (25). For example, if $\mathscr{D}$ is a ball of radius a with center at the origin, then a basis in the set of solutions of Eq. (25) forms the functions $r^{-1/2}J_{n+1/2}(kr)Y_{nm}(\theta,\phi)$, where J_n is the Bessel function and Y_{nm} are the spherical harmonics.

Remark: It seems that the first paper on antenna synthesis theory was published in 1937 by my father.[10]

[1]A. Devaney and G. Sherman, "Nonuniqueness in inverse source and scattering problems," IEEE Trans. Antennas Propag. **AP-30**, 1034–1037 (1982).

[2]A. Devaney, J. Math Phys. **19**, 1526–1531 (1978); **20**, 1687–1691 (1979).

[3]N. Bleistein and J. Cohen, J. Math. Phys. **18**, 194–201 (1977).

[4]N. Bojarski, J. Math. Phys. **22**, 1647–1650 (1981).

[5]*Inverse Source Problems in Optics*, edited by H. Bates (Springer-Verlag, New York, 1978).

[6]A. G. Ramm, J. Math. Phys. **22**, 275–276 (1981).

[7]A. G. Ramm, *On Antenna Synthesis Theory*, Antennas, No. 5 (Izdat. Svjaz, Moscow, 1969), pp. 35–46 (Russian).

[8]A. G. Ramm, *Theory and Applications of Some New Classes of Integral Equations* (Springer-Verlag, New York, 1980).

[9]B. Minkovich and V. Jakovlev, *Antenna Synthesis Theory* (Soviet Radio, Moscow, 1967) (Russian).

[10]G. S. Ramm, "On antifeding antennas," in *Collection of Scientific Papers on Electrocommunication* (Leningrad Institute of Electrical Engineering and Communication (LEEC), Leningrad, 1937), Vol. 3, p. 19. Also G. S. Ramm, Ph.D. Thesis "Antifeding antennas," LEEC 1937.

On joint lower bounds of position and momentum observables in quantum mechanics

Paul Busch
Institute for Theoretical Physics, University of Cologne, Cologne, Federal Republic of Germany

(Received 19 July 1983; accepted for publication 18 October 1983)

A new theorem on joint lower bounds of unsharp position and momentum observables is formulated and proven. It is developed as a straightforward formal generalization of a similar theorem for the spectral projections of the usual observables which has been employed to express the strict incompatibility of position and momentum. The relevance of our theorem to the problem of joint measurements in quantum mechanics has been discussed elsewhere.

PACS numbers: 03.65.Bz

I. INTRODUCTION

Usually the incompatibility of quantum mechanical position and momentum is expressed rather formally by means of the commutation relation $[\mathscr{Q},\mathscr{P}]_- = i\hbar I$. It was Jauch[1] who used a theorem on the intersections of the ranges of the spectral projections of $\mathscr{Q}$ and $\mathscr{P}$ to point out the nonexistence of a well-behaved joint probability distribution. This "Jauch theorem" (J) (which will be reported in Sec. II) represents a much more intuitive and very strong explication of the incompatibility of $\mathscr{Q}$ and $\mathscr{P}$.

On the other hand, it has been shown in recent years that by introducing "unsharp" or "fuzzy" observables Q_f, P_g a notion of joint measurements becomes possible for certain pairs of "Fourier related" Q_f, P_g, with measuring unsharpnesses obeying the uncertainty relation

$$\Delta q\cdot\Delta p \geqslant \tfrac{1}{2}\hbar \tag{1}$$

(see Sec. III).[2-4]

The purpose of the present paper is to generalize theorem (J) from ordinary observables to the full class of unsharp position and momentum observables. We arrive at a theorem $(\tilde{J})$, the proof of which will be seen to be fully independent of the uncertainty relation (1). This gives rise to a real extension of the class of pairs Q_f, P_g, which are compatible in the sense defined below.

The physical interpretation of these results has been studied in Ref. 5, where a detailed analysis of the significance of the uncertainty relation to the possibility of joint measurements is given.

II. INCOMPATIBILITY STATEMENT: JAUCH THEOREM (J)

Let $\mathscr{L}^2(\mathscr{R},dq)$ and $\mathscr{L}^2(\mathscr{R},dp)$ be the configuration and momentum space representations of the state space $\mathscr{H}$ of a (one-dimensional) quantum mechanical particle. In the standard framework, position and momentum observables are defined as projection-valued measures on the Borel sets $\mathscr{B}(\mathscr{R})$ of $\mathscr{R}$

$$\begin{aligned} &Q:\ \ \mathscr{B}(\mathscr{R})\rightarrow\mathscr{P}(\mathscr{H}),\quad X\rightarrow Q(X),\\ &(Q(X)\psi)(q)=\chi_X(q)\cdot\psi(q),\\ &P:\ \ \mathscr{B}(\mathscr{R})\rightarrow\mathscr{P}(\mathscr{H}),\quad Y\rightarrow P(Y),\\ &(P(Y)\tilde{\psi})(p)=\chi_Y(p)\cdot\tilde{\psi}(p) \end{aligned} \tag{2}$$

such that

$$\mathscr{Q}=\int_{\mathscr{R}} q\,Q(dq),\quad \mathscr{P}=\int_{\mathscr{R}} p\,P(dp) \tag{3}$$

are the usual Fourier–Plancherel equivalent position and momentum operators on $\mathscr{H}$. We denote by $\mathscr{E}(\mathscr{H})$ the set of all *effects*, i.e., of all self-adjoint operators A lying between O and I, $O\leqslant A\leqslant I$; for any two $A, B\in\mathscr{E}(\mathscr{H})$ we define the set $\mathscr{L}\mathscr{b}(A,B)$ of positive *l*ower *b*ounds as

$$\mathscr{L}\mathscr{b}(A,B)=\{E\in\mathscr{E}(\mathscr{H})|E\leqslant A \text{ and } E\leqslant B\}\subseteq\mathscr{E}(\mathscr{H}).$$

We call effects A, B *incompatible* if $\mathscr{L}\mathscr{b}(A,B)=\{0\}$; otherwise they are called *compatible*.

Finally, let $P\wedge Q$ $(P,Q\in\mathscr{P}(\mathscr{H}))$ be the projection onto the intersection $P\mathscr{H}\cap Q\mathscr{H}$ of the ranges of P and Q. Then we may formulate the following:

Theorem (J): For arbitrary measurable sets X, Y with finite Lebesgue measures, it is

$$\begin{aligned} &Q(\mathscr{R}/X)\wedge P(\mathscr{R}/Y)\neq 0,\\ &\text{i.e., } \mathscr{L}\mathscr{b}(Q(\mathscr{R}/X),P(\mathscr{R}/Y))\neq\{0\} \end{aligned} \tag{J-0}$$

$$[\text{moreover, } \dim(Q(\mathscr{R}/X)\wedge P(\mathscr{R}/Y)\mathscr{H})=\infty],$$

$$Q(X)\wedge P(Y)=0,\quad \text{i.e., } \mathscr{L}\mathscr{b}(Q(X),P(Y))=\{0\}. \tag{J-1}$$

Further, for semibounded X, Y with finite measures,

$$\begin{aligned} &Q(X)\wedge P(\mathscr{R}/Y)=0,\\ &\text{i.e., } \mathscr{L}\mathscr{b}(Q(X),P(\mathscr{R}/Y))=\{0\}, \end{aligned} \tag{J-2}$$

$$\begin{aligned} &Q(\mathscr{R}/X)\wedge P(Y)=0,\\ &\text{i.e., } \mathscr{L}\mathscr{b}(Q(\mathscr{R}/X),P(Y))=\{0\}. \end{aligned} \tag{J-3}$$

Statements (J-0)–(J-3) have been established and proven first for bounded X, Y by Lenard,[6] the extension to semibounded X, Y comes from Berthier and Jauch,[7] and the proof of (J-0) and (J-1) for arbitrary X, Y with finite measures is given in a more general context by Amrein and Berthier.[8] As an example, (J-1) means that there is no state $\psi\in\mathscr{H}$ such that both $\psi(q)$ and $\tilde{\psi}(p)$ have supports with finite measures. Physically, there is no possibility of localizing a particle within a finite phase space cell.

This result is fully independent of the uncertainty relation (1) which as a consequence *may not* be taken as an expression of position and momentum incompatibility within the standard formalism of quantum mechanics. On the contrary, as we shall see now, the UR (1) seems to make possible a sort of joint measurement of unsharp position and momentum values respecting the limitations given by (J).

 0022-2488/84/061794-04$02.50

III. JOINT DISTRIBUTIONS FOR FOURIER-RELATED OBSERVABLES

Unsharp position and momentum observables Q_f, P_g are defined as effect-valued measures

$$\begin{aligned} &Q_f: \quad \mathscr{B}(\mathscr{R}) \to \mathscr{E}(\mathscr{H}), \quad X \to Q_f(X), \\ &(Q_f(X)\psi)(q) = \nu_X^f(q) \cdot \psi(q), \\ &P_g: \quad \mathscr{B}(\mathscr{R}) \to \mathscr{E}(\mathscr{H}), \quad Y \to P_g(Y), \\ &(P_g(Y)\tilde{\psi})(p) = \nu_Y^g(p) \cdot \tilde{\psi}(p), \end{aligned} \tag{4}$$

where

$$\nu_X^f(q) := (\chi_X * f)(q), \quad \nu_Y^g(p) := (\chi_Y * g)(p) \tag{5}$$

are smeared-out characteristic functions which are concentrated on X and Y; f and g are normalized positive distribution functions satisfying

$$\begin{aligned} &\langle \mathscr{Q} \rangle_f := \int_{\mathscr{R}} q f(q)\, dq = 0, \\ &(\Delta_f \mathscr{Q})^2 \equiv (\Delta f)^2 := \int_{\mathscr{R}} dq\, q^2 f(q) < \infty, \\ &\langle \mathscr{P} \rangle_g := \int_{\mathscr{R}} p\, g(p)\, dp = 0, \\ &(\Delta_g \mathscr{P})^2 \equiv (\Delta g)^2 := \int_{\mathscr{R}} dp\, p^2 g(p) < \infty . \end{aligned} \tag{6}$$

From (4)–(6) we get the operators (3):

$$\mathscr{Q} = \int_{\mathscr{R}} q\, Q_f(dq), \quad \mathscr{P} = \int_{\mathscr{R}} p\, P_g(dp) \quad \text{(weakly).}$$

Clearly the standard observables (2) are recovered by setting $f(q) = \delta(q)$ and $g(p) = \delta(p)$. The unsharpnesses described by (5) and (6) reflect the finite resolutions of the position and momentum measuring instruments[3,4] arising from the quantum mechanical nature of their interaction with the observed object; this interpretation has been justified by means of a quantum mechanical theory of measurement by the present author.[9] We call observables Q_f, P_g *incompatible* if for any pair of bounded measurable sets X, Y in $\mathscr{B}(\mathscr{R})$ the effects $Q_f(X)$, $P_g(Y)$ are *incompatible*; otherwise, Q_f and P_g are called *compatible*. Ali and Prugovecki[4] have shown as a generalization of a theorem by Wigner the following:

Theorem (AP): For a given pair of positive numbers $(\Delta q, \Delta p)$, there exists a joint observable a_{fg} for some position and momentum observables Q_f, P_g with $\Delta f = \Delta q, \Delta g = \Delta p$ if and only if $\Delta q \cdot \Delta p \geqslant \hbar/2$.

A joint observable a_{fg} is a continuous effect-valued measure on phase space $\Gamma = \mathscr{R}^2$ ($\hbar = 1$)

$$\begin{aligned} &a_{fg}: \mathscr{B}(\mathscr{R}^2) \to \mathscr{E}(\mathscr{H}), \\ &\Delta \to a_{fg}(\Delta) = (2\pi)^{-1} \int_{\Delta} \mathscr{W}_{qp}\, dq\, dp, \\ &\mathscr{W}_{qp} = \exp(ip\mathscr{Q}) \exp(-iq\mathscr{P}) \cdot \mathscr{W} \cdot \exp(iq\mathscr{P}) \\ &\qquad \times \exp(-ip\mathscr{Q}), \\ &\mathscr{W} \geqslant 0, \ \mathrm{tr}(\mathscr{W}) = 1 \end{aligned} \tag{7}$$

with marginal observables Q_f, P_g:

$$a_{fg}(X \times \mathscr{R}) = Q_f(X), \quad a_{fg}(\mathscr{R} \times Y) = P_g(Y). \tag{8}$$

It has been shown within a theory of measurement that it is reasonable to regard a measurement of a_{fg} as a joint measurement of Q_f and P_g.[9] Thus (AP) states that the strict incompatibility of position and momentum (J-1) disappears if one considers "Fourier-related" observables Q_f, P_g for which $f(q) = \langle q | \mathscr{W} | q \rangle$, $g(p) = \langle p | \mathscr{W} | p \rangle$ for some state operator $\mathscr{W}$. In that case there exists some nonzero lower bound α of $Q_f(X)$ and $P_g(Y)$ [namely, $\alpha = a_{fg}(X \times Y)$]:

$$\alpha \leqslant Q_f(X) \quad \text{and} \quad \alpha \leqslant P_g(Y), \quad \alpha = \alpha(X, Y). \tag{9}$$

In the remaining part of this paper we shall investigate to which extent the class of pairs Q_f, P_g admitting (9) can be enlarged by dropping the restriction to Fourier-related pairs. It will become clear that it is not the UR which is responsible for a relaxation of position and momentum incompatibility (J-1) through (9). This will be illustrated by means of some examples.

IV. GENERALIZED JAUCH THEOREM ($\tilde{\mathrm{J}}$)

Quite surprisingly, it is the structures of the generalized "characteristic functions" ν_X^f, ν_Y^g [Eq. (5)], which are responsible for the existence or nonexistence of some positive lower bound of the corresponding $Q_f(X)$, $P_g(Y)$ [defined in (4)], just as it was the case for χ_X, χ_Y in Jauch's theorem (J). Namely, we may formulate as "generalized Jauch theorem" the following:

Theorem ($\tilde{\mathrm{J}}$): Let $X \to Q_f(X) := (\chi_X * f)(\mathscr{Q}) \equiv \nu_X^f(\mathscr{Q})$, $Y \to P_g(Y) := (\chi_Y * g)(\mathscr{P}) \equiv \nu_Y^g(\mathscr{P})$ be unsharp position and momentum observables characterized by some distribution functions f, g (cf. Sec. III). There exists a positive lower bound of $Q_f(X)$ and $P_g(Y)$ for given measurable sets X, Y, that is,

$$\mathscr{L}\ell(Q_f(X), P_g(Y)) \neq \{0\}$$

if and only if the ranges of the respective square root operators possess nontrivial intersection

$$R\,(Q_f(X)^{1/2}) \cap R\,(P_g(Y)^{1/2}) \neq \{0\} . \tag{10}$$

A necessary condition for this is the following:

$$Q\,(\mathrm{supp}\,\{\nu_X^f\}) \wedge P(\mathrm{supp}\,\{\nu_Y^g\}) \neq \{0\} . \tag{11}$$

The proof will be given in the Appendix. At present it is an open problem whether the last relation (11) is also sufficient; clearly this is the case for standard observables ($f = g = \delta$). Let us call Q_f, P_g a *second type couple* if there exist bounded measurable sets X, Y such that (10) holds true; otherwise, they shall be called a *first type couple*. Then our theorem immediately implies the following:

Corollary: Q_f, P_g are compatible if and only if they are a second type couple.

According to Theorem (AP) Fourier couples are second type couples and thus compatible. In the next section we shall show the existence of non-Fourier second type couples which means that it is possible to break the verdict (J-1) independently of the UR (1).

V. SOME APPLICATIONS AND EXAMPLES

First we present a relaxation of the verdicts (J-2) and (J-3). Take Q_f with supp $\{f\} = \mathscr{R}$, X a measurable set with finite measure. Then $\chi_{X'} * f = 1 - \chi_X * f$ is strictly positive so that $Q_f(X') = Q_f(\mathscr{R}/X)$ is positive definite and therefore $R(Q_f(R/X)^{1/2}) = \mathscr{H}$. We obtain for any P_g and arbitrary (bounded) measurable set Y

$$\mathscr{L}\mathscr{b}(Q_f(R/X), P_g(Y)) \neq \{0\}.$$

Similarly, for supp $\{g\} = \mathscr{R}$, Y with finite measure and arbitrary f, X

$$\mathscr{L}\mathscr{b}(Q_f(X), P_g(\mathscr{R}/Y)) \neq \{0\}.$$

Next the verdict (J-1) can be extended to the following subclass of first type couples Q_f, P_g. Let supp $\{f\}$, supp $\{g\}$ as well as X, Y be bounded sets; then (11) is violated so that

$$\mathscr{L}\mathscr{b}(Q_f(X), P_g(Y)) = \{0\}.$$

The incompatibility of first type couples holds independently of the value of the uncertainty product $\Delta f \cdot \Delta g$.

The next example proves the existence of compatible Q_f, P_g with arbitrarily small uncertainty product $\Delta f \cdot \Delta g$ ($\ll \hbar/2$). Let $f^\circ(q) = |\phi(q)|^2$, $g^\circ(p) = |\tilde{\phi}(p)|^2 = g(p)$ with supp $\{\phi\}$ being bounded. Take any f with supp $\{f\} = \mathscr{R}$. It is

$$P_{g^\circ}(Y) \leqslant P_g(Y).$$

Further, as we shall show, there exist bounded measurable sets X° such that

$$Q_{f^\circ}(X^\circ) \leqslant Q_f(X).$$

Since Q_{f°, P_{g° are a Fourier couple, Theorem (AP) gives

$$a_{f^\circ g^\circ}(X^\circ \times Y) \in \mathscr{L}\mathscr{b}(Q_f(X), P_g(Y)) \neq \{0\} \tag{12}$$

In particular, Q_f and P_g are compatible even for $\Delta f \cdot \Delta g < \hbar/2$. To prove the second of the above inequalities, we mention that for bounded X° the set $S^\circ = \text{supp}\{\nu_{X^\circ}^{f^\circ}\}$ is bounded. Taking smaller and smaller X° makes $M^\circ = \max\{\nu_{X^\circ}^{f^\circ}(q) | q \in S^\circ\}$ arbitrarily small. Further, for supp $\{f\} = \mathscr{R}$ the continuity of $\nu_X^f = \chi_X * f$ implies

$$0 < \min\{\nu_X^f(q) | q \in S\} = m \leqslant \min\{\nu_X^f(q) | q \in S^\circ\}$$

for any fixed bounded $S \supset S^\circ$. By taking X° such that $M^\circ \leqslant m$, we obtain

$$\nu_{X^\circ}^{f^\circ}(q) \leqslant \nu_X^f(q) \quad \text{for all } q \in \mathscr{R},$$

which is equivalent to $Q_{f^\circ}(X^\circ) \leqslant Q_f(X)$. This proves (12).

ACKNOWLEDGMENT

I should like to express my gratitude to Dr. P. J. Lahti from Turku, Finland, for drawing my attention to the problem treated in the present paper as well as for stimulating discussions during his stay at Cologne.

APPENDIX: PROOF OF THE GENERALIZED JAUCH THEOREM ($\tilde{\mathrm{J}}$)

In order to specify the necessary and sufficient conditions for a nontrivial set of lower bounds $\mathscr{L}\mathscr{b}(Q_f(X), P_g(Y)) \neq \{0\}$, we shall proceed in several steps.

(1) First we notice that the existence of some effect $a \neq 0$ with

$$a \leqslant Q_f(X) \tag{A1}$$

is equivalent to the existence of some effect $\hat{a} := \alpha P_\xi$ with

$$\hat{a} := \alpha P_\xi \leqslant Q_f(X),$$

$$0 < \alpha < 1, \xi \in \mathscr{H}, \quad \|\xi\| = 1, \quad P_\xi = |\xi\rangle\langle\xi|. \tag{A2}$$

One simply has to take the spectral projections $A(S)$ of a belonging to some subset S of the spectrum of a such that there exists a positive number $\alpha > 0$ as lower bound of S. Then, for ξ with $A(S)\xi = \xi$, we have

$$\hat{a} = \alpha P_\xi \leqslant \alpha A(S) \leqslant a.$$

(2) If an effect αP_ξ is to obey (A2), then the following restriction for α must hold:

$$0 < \alpha \leqslant \langle Q_f(X)\rangle_\xi = (\xi, Q_f(X)\xi) = \|Q_f(X)^{1/2}\xi\|^2.$$

Any vector $\psi \in \mathscr{H}$, $\|\psi\| = 1$, can be decomposed into $\psi = \kappa\xi + \lambda\xi'$ with some ξ' orthogonal to ξ ($\xi \perp \xi'$), $\|\xi'\| = 1$. If (A2) is to hold, i.e., if

$$(\psi, \alpha P_\xi \psi) = \alpha|(\psi, \xi)|^2 \leqslant (\psi, Q_f(X)\psi) \quad \text{for all } \psi \in \mathscr{H},$$

it follows (for $\xi' \perp \xi$)

$$0 \leqslant |\kappa|^2 \cdot \epsilon + |\lambda|^2 \cdot \|Q_f(X)^{1/2}\xi'\|^2 + 2\,\mathrm{Re}\,\{\kappa\lambda^*(\xi', Q_f(X)\xi)\},$$

where

$$0 \leqslant \epsilon := \|Q_f(X)^{1/2}\xi\|^2 - \alpha.$$

Since this must be true for any $\kappa \in \mathscr{R}$, $\lambda = (1 - |\kappa|^2)^{1/2} \cdot \exp(-i\tau)$, $0 \leqslant \tau < 2\pi$, we get [by taking the minimum with respect to $\delta = \cos^{-1}(\kappa)$]

$$[\mathrm{Re}\,\{\exp(i\tau) \cdot (\xi', Q_f(X)\xi)\}]^2 \leqslant \epsilon \cdot \|Q_f(X)^{1/2}\xi'\|^2$$

and from this (by maximizing with respect to τ)

$$|(\xi', Q_f(X)\xi)|^2 \leqslant \epsilon \cdot \|Q_f(X)^{1/2}\xi'\|^2.$$

Explicitly written, this gives

$$|(Q_f(X)^{1/2}\psi, Q_f(X)^{1/2}\xi)|^2 \leqslant \|Q_f(X)^{1/2}\psi\|^2 \cdot \{\|Q_f(X)^{1/2}\xi\|^2 - \alpha\}$$
$$\text{for all } \psi \text{ with } (\psi, \xi) = 0. \tag{A3}$$

(A3) is readily seen to be equivalent to (A2); for $\alpha = 0$ it reduces to an application of Schwarz's inequality.

Now let us consider the closed subspace

$$\mathscr{M}_X := \overline{Q_f(X)^{1/2}([\xi]^\perp)} \equiv Q_X \mathscr{H}$$

with projection Q_X; for $\xi_X := Q_f(X)^{1/2}\xi$ and $\psi \in \mathscr{M}_X$ we have

$$(\psi, \xi_X) = (Q_X\psi, \xi_X) = (\psi, Q_X\xi_X)$$

so Schwarz's inequality implies

$$|(\psi, \xi_X)|^2 \leqslant \|\psi\|^2 \cdot \|Q_X\xi_X\|^2 \quad \text{for all } \psi \in \mathscr{M}_X.$$

But then (A3) is seen to be equivalent to

$$\|Q_X\xi_X\|^2 \leqslant \|\xi_X\|^2 - \alpha < \|\xi_X\|^2$$

i.e.,

$$0 < \alpha \leqslant \|(I - Q_X)\xi_X\|^2. \tag{A4}$$

(3) In the final step we have to investigate the question whether for $Q_f(X)$ there exists some state $\xi \in \mathscr{H}$ such that the image ξ_X does not belong to the closure $\mathscr{M}_X$ of the image of $[\xi]^\perp$ under $Q_f(X)^{1/2}$. This is the content of relation (A4) which we shall prove now to be equivalent to

$$\xi = Q_f(X)^{1/2}\eta \quad \text{for some } \eta \in \mathscr{H},$$

$$\text{i.e.,} \quad \xi \in R(Q_f(X)^{1/2}). \tag{A5}$$

Let $A = Q_f(X)^{1/2}$. Then we have to show

$$A\xi \notin \overline{A([\xi]^\perp)} = \mathscr{M}_X \quad \text{iff} \quad \xi \in A\mathscr{H} = R(A).$$

We begin by remarking that

$$\overline{R(A)} = R(Q(S^\circ)) = :\mathscr{H}^\circ,$$

$$S^\circ := \operatorname{supp}\{\nu_X^f\}, \quad R(A) = A\mathscr{H} = A\mathscr{H}^\circ,$$

which follows from the fact that A is self-adjoint and one-to-one on $\mathscr{H}^\circ$. Both conditions (A4) as well as (A5) imply $\xi \in \mathscr{H}^\circ$: in case of (A5) this is evident. To show that $\xi \in \mathscr{H}^\circ$ is necessary for (A4), assume $\xi = \xi_0 + \xi_1 \notin \mathscr{H}^\circ$, $\xi_0 \in \mathscr{H}^\circ$, $\xi_1 \perp \mathscr{H}^\circ$, $\xi_1 \neq 0$. Then for

$$\psi = \xi_0 - \xi_1 \|\xi_0\|^2/\|\xi_1\|^2 \in [\xi]^\perp$$

we have

$$A\psi = A\xi_0 = A\xi = \xi_X,$$

which means $\xi_X \in \mathscr{M}_X = \overline{A([\xi]^\perp)}$. So we have established (11) as necessary condition for (10), and it suffices to prove the above equivalence under the assumption $\xi \in \mathscr{H}^\circ$.

Consider a complete orthonormal system in $\mathscr{H}^\circ$, $\{\psi_0 = \xi, \psi_1, \psi_2, \cdots\}$. Then the set $\{A\psi_0, A\psi_1, \cdots\}$ is total in $\mathscr{H}^\circ$: from $(\phi, A\psi_i) = 0$ for all $i, \phi \in \mathscr{H}^\circ$, it follows that $(A\phi, \psi_i) = 0$ for all i such that $A\phi = 0$ and $\phi = 0$. Now assume first $\xi = A\eta \in R(A)$, $\eta \in \mathscr{H}^\circ$. It follows that

$$0 = (\xi, \psi_i) = (\eta, A\psi_i) \quad \text{for all } i \geqslant 1,\ \eta \neq 0;$$

thus $\eta \perp \overline{A([\xi]^\perp)} = Q_X\mathscr{H}$, i.e., $(I - Q_X)\eta = \eta$. Since $(\eta, A\xi) = (\xi, \xi) \neq 0$, we obtain $0 \neq ((I - Q_X)\eta, A\xi) = (\eta, (I - Q_X)A\xi)$ and thus $A\xi \notin \overline{A([\xi]^\perp)}$. To prove the second part of the equivalence statement, assume $\xi \notin R(A)$. From $(A\psi_i, \phi) = 0$ for all $i \geqslant 1, \phi \in \mathscr{H}^\circ$, we obtain $(\psi_i, A\phi) = 0$ for all $i \geqslant 1$. But then $A\phi = \lambda\psi_0 = \lambda\xi$ for some constant λ. $\xi \notin R(A)$ implies $\lambda = 0$, thus $A\phi = 0$, and $\phi = 0$. Since $\phi \in \mathscr{H}^\circ$ is arbitrary, we have proven that already the set $\{A\psi_1, A\psi_2, \cdots\}$ is total in $\mathscr{H}^\circ$. Therefore,

$$A\psi_0 = A\xi \in \overline{[A\psi_1, A\psi_2, \cdots]} = \overline{A([\xi]^\perp)}.$$

To summarize, we have established the following statement: there exists some effect a obeying (A1) if and only if there is some nonzero $\xi \in \mathscr{H}$ obeying one of the equivalent conditions (A2)–(A5). A similar chain of equivalent conditions may be formulated for $P_g(Y)$. So we arrive at the desired result: any vector $\xi \neq 0$ lying in the intersection $\mathscr{M}$ of the ranges

$$\mathscr{M} = R(Q_f(X)^{1/2}) \cap R(P_g(Y)^{1/2})$$

will give rise to a positive lower bound of $Q_f(X)$ and $P_g(Y)$. $\mathscr{M} \neq 0$ is a necessary and sufficient condition for $\mathscr{L}b(Q_f(X), P_g(Y)) \neq \{0\}$.

[1] J. M. Jauch, Synthese **29**, 131 (1974).
[2] E. B. Davies, *Quantum Theory of Open Systems* (Academic, New York, 1976).
[3] E. Prugovecki, J. Math. Phys. **17**, 517 (1976).
[4] S. T. Ali and E. Prugovecki, J. Math. Phys. **18**, 219 (1977).
[5] P. Busch and P. J. Lahti, "On various joint measurements of position and momentum observables," preprint, University of Turku, 1983, Phys. Rev. D (to be published).
[6] A. Lenard, J. Funct. Anal. **10**, 410 (1972).
[7] A. M. Berthier and J. M. Jauch, Lett. Math. Phys. **1**, 93 (1976).
[8] W. O. Amrein and A. M. Berthier, J. Funct. Anal. **24**, 258 (1977).
[9] P. Busch, "Indeterminacy Relations and Simultaneous Measurements in Quantum Theory," Int. J. Theor. Phys. (to be published).

The role of the connection in geometric quantization

I. H. McKenna and K. K. Wan
Department of Theoretical Physics, St. Andrews University, St. Andrews, Scotland

(Received 20 June 1983; accepted for publication 4 November 1983)

The effects on quantization of using different connections are studied. Through explicit examples it is shown that with a proper choice of connection the half-density quantization scheme with the B.W.S. condition can produce physically correct results while the complicated half-form quantization scheme with the corrected B.W.S. condition can produce physically incorrect results.

PACS numbers: 03.65.Ca, 02.40. + m

1. INTRODUCTION

In this paper we discuss an aspect of geometric quantization which seems to have been largely ignored in the past. The nonuniqueness of the choice of the connection on the quantum line bundle is well known but the actual practical effects on quantization do not appear to have been systematically discussed in the existing literature. We are particularly interested here in the effect the choice of connection has on the so-called B.W.S. and corrected B.W.S conditions. To this end we quantize one-dimensional Hamiltonians with potential wells and find that, in the examples we discuss, a judicious (*a posteriori*) choice of connection gives the exact energy eigenvalues while the standard connection gives approximate energy eigenvalues only.

Throughout this paper, unless otherwise stated or explicitly defined, we have adopted the notation and geometric quantization schemes described by Woodhouse.[1]

The basic elements of geometric quantization are a symplectic manifold (M,ω) and a Hermitian line bundle $\mathbf{B}$ over M with (compatible) connection ∇ (a connection means a connection compatible with the Hermitian structure in this paper) with curvature $\omega/\hbar$. We construct the quantum Hilbert space $\mathscr{H}_{P'}$ or $\mathscr{H}^m_{P'}$ associated with a polarization P' of M (we will restrict our attention to real polarizations here) from sections s of $\mathbf{B}$ which are covariantly constant with respect to vector fields in P' and similarly covariantly constant half-densities $\nu \in \Delta_{-1/2}(P')$ or half-forms $\nu \in \delta_{-1/2}(P')$ in some metaplectic structure $\delta_{1/2}$. Explicitly the quantum pre-Hilbert space $W_{P'}$ (or $W^m_{P'}$) consists of smooth sections $\Psi = s\cdot\nu$ of $\mathbf{B}\otimes\Delta_{-1/2}(P')$ or $\mathbf{B}\otimes\delta_{-1/2}(P')$ which obey

$$\nabla_X s = 0; \quad \nabla_X \bar{\nu} = 0 \quad \forall X \in \mathfrak{X}(M,P'),$$

where $\mathfrak{X}(M,P')$ denotes the set of smooth vector fields X such that $X_m \in P'_m$ for each $m \in M$, and which admit a finite inner product given by[1]

$$\langle \Psi | \Psi \rangle = (2\pi\hbar)^{(-1/2)n} \int_{Q'} \langle \Psi,\Psi \rangle,$$

where $Q' = M/P$, $\langle \Psi,\Psi\rangle_{q'} = (s,s)_m |\bar{\nu}|^2_m |\epsilon_\omega|_m$ and where $q'(\epsilon Q') = \pi'(m)$, π' being the projection $M \to Q'$. $(.,.)$ is the Hermitian structure on $\mathbf{B}$ and ϵ_ω is the Liouville form on M. The quantum Hilbert space $\mathscr{H}_{P'}$ is then the completion of the pre-Hilbert space $W_{P'}$.

In quantum mechanics we are normally concerned with a symplectic manifold (M,ω) where M is the cotangent bundle T^*Q of some configuration space Q. In this case, there is a natural polarization, the vertical polarization which we shall denote by P, for which $M/P = Q$ and $\mathscr{H}_P$ or $\mathscr{H}^m_P$ is identifiable with the Hilbert space $L^2(Q)$ of square-integrable functions on Q. We shall restrict ourselves to cotangent bundles from now on.

When M is simply connected, all connections ∇ with curvature $\omega/\hbar$ are equivalent.[1] We are concerned in this paper with what happens when M is not simply connected. There is then an infinite set of inequivalent choices for ∇ and we would expect that the choice of connection has physical consequences.

Let us review some of the relevant properties of a compatible connection on $\mathbf{B}$.

Let s_1 be any nowhere-zero smooth section of $\mathbf{B}$ over some neighborhood $N \subset M$. Given a connection ∇ we can define a connection potential β_1 on N obeying $d\beta_1 = \omega$ by

$$\nabla_X s_1 = -(i/\hbar)(X \lrcorner \beta_1) s_1 \quad \forall X \in V(M),$$

where $V(M)$ denotes the set of smooth vector fields on M. Note that our β differs from Woodhouse's notation by a factor $\hbar$. Given a second nowhere-zero smooth section s_2 on N we can define a second connection potential β_2 on N obeying $d\beta_2 = \omega$ by

$$\nabla_X s_2 = -(i/\hbar)(X \lrcorner \beta_2) s_2 \quad \forall X \in V(M).$$

Since s_1 and s_2 are nowhere-zero there exists a nowhere-zero smooth function u on N such that $s_2 = u s_1$. It then follows that

$$\beta_2 = \beta_1 + i\hbar\, d(\ln u).$$

Now let ∇' be a second connection on $\mathbf{B}$. We can define a corresponding connection potential β'_1 obeying $d\beta'_1 = \omega$ by

$$\nabla'_X s_1 = -(i/\hbar)(X \lrcorner \beta'_2) s_1 \quad \forall X \in V(M).$$

Suppose that s_1, s_2 are global sections on M and that β_1, β'_1 are global connection potentials on M. If there exists on M a smooth global function u of the form $u = \exp(v/i\hbar)$, where v is a real function on M, such that

$$(X \lrcorner \beta'_1) s_1 = (X \lrcorner (\beta_1 - i\hbar d(\ln u))) s_1 \quad \forall X \in V(M),$$

then ∇' is equivalent to ∇ with the equivalence map[1]

$$\rho: \mathbf{B} \to \mathbf{B} \text{ given by } s \to \rho(s) = us.$$

We can readily check that

(1) $(s,s) = (\rho(s), \rho(s))$,

(2) $\rho(\nabla_X s) = \nabla'_X \rho(s)$, and

(3) ∇ being compatible with the Hermitian structure implies ∇' is also compatible with the Hermitian structure.

While such a function exits locally it may well not exist globally. However, if M is contractible, then such a function u exists globally. The reasoning is simple: (1) $d\beta'_1 = \omega = d\beta_1$ implies $\beta'_1 = \beta_1 + \alpha$, where α is a closed 1-form, (2) every closed 1-form on a contractible M is exact by the Poincaré Lemma,[2] i.e., $\alpha = du$ for some global function u.

2. QUANTIZATION WITH EQUIVALENT CONNECTIONS

Let $Q = \mathbb{R}$ the real line, $M = T^*Q = \mathbb{R}^2$, $\mathbf{B} = M \otimes \mathbb{C}$, and let (p,q) be the usual global Cartesian canonical coordinates on T^*Q. The vertical polarization P is spanned by $\partial/\partial p$. Let s_0 be any nowhere-zero smooth global section of $\mathbf{B}$ and let ∇,∇' be connections on $\mathbf{B}$ defined by

$$\left.\begin{aligned}\nabla_X s_0 &= -(i/\hbar)(X \lrcorner\, p dq)s_0\\ \nabla'_X s_0 &= -(i/\hbar)(X \lrcorner\, \beta')s_0\end{aligned}\right\} \quad \forall X \in V(M).$$

Since $M = \mathbb{R}^2$ we can write, by the Poincaré Lemma, $\beta' = pdq - i\hbar d(\ln u)$ for some globally smooth function u.

Now s_0 is covariantly constant along $X \in \mathfrak{X}(M,P)$ with respect to ∇. Hence, with ∇ as connection, we can write any section s polarized with respect to P in the form $s = \psi(q)s_0$.

Now, since ∇ and ∇' are equivalent we have

$$\nabla'_X u s_0 = u\nabla_X s_0 = 0$$

and so, us_0 is also covariantly constant along $X \in \mathfrak{X}(M,P)$ with respect to ∇'.

With ∇' as connection we can write any section s polarized with respect to P as

$$s = \psi(q)us_0 = \psi(q)\exp(-iv/\hbar)s_0.$$

Given a connection ∇ and a polarization P we construct the quantum pre-Hilbert space W_p from sections of $\mathbf{B} \otimes \Delta_{-1/2}(P)$ [or $\mathbf{B} \otimes \delta_{-1/2}(P)$, there being no practical difference between the half-density and half-form quantization in this case] of the form

$$\Phi = \phi(q)s_0|dp|^{-1/2}, \tag{2.1}$$

with scalar product

$$\langle \Phi | \Phi \rangle = \frac{1}{(2\pi\hbar)^{1/2}} \int_{-\infty}^{\infty} |\phi(q)|^2 dq.$$

There is also a unique self-adjoint operator in $\mathscr{H}_P$ corresponding to any classical observable ζ of the form $\zeta = \xi(q)p + \eta(q)$ provided $\xi(q)\partial/\partial q$ is a complete vector field on Q.[3] The operator expression $\hat{\zeta}$ is given, according to the standard half-density quantization scheme, by

$$\hat{\zeta}\Phi = \left[\left(-i\hbar\left(\xi(q)\frac{\partial}{\partial q} + \frac{1}{2}\frac{\partial \xi}{\partial q}\right) + \eta(q)\right)\phi(q)\right]s_0|dp|^{-1/2}. \tag{2.2}$$

Now we can repeat this construction with the connection ∇' to obtain the quantum Hilbert space $\mathscr{H}'_P$ which consists of elements of the form

$$\Phi' = \phi(q)s'_0|dp|^{-1/2}, \quad s'_0 = (\exp v/i\hbar)s_0, \tag{2.3}$$

and the operator expression $\hat{\zeta}'$ in $\mathscr{H}'_P$ is

$$\hat{\zeta}\Phi' = \left[\left(-i\hbar\left(\xi(q)\frac{\partial}{\partial q} + \frac{1}{2}\frac{\partial \xi}{\partial q}\right) + \eta(q)\right)\phi(q)\right] \times s'_0|dp|^{-1/2}. \tag{2.4}$$

Let us introduce a map of W_p onto W'_p, $U_0: W_p \to W'_p$, by $\Phi \to U_0\Phi = \Phi'$, where Φ and Φ' are given by (2.1) and (2.3). The map U_0 is obviously an isometric operator[4,5] with a domain dense in $\mathscr{H}_p$ and a range dense in $\mathscr{H}'_p$. Hence U_0 admits[4] a unique continuous linear extension U to the entire Hilbert space $\mathscr{H}_p$ on which U is isometric. We conclude[5] that U is in fact a unitary map of $\mathscr{H}_p$ onto $\mathscr{H}'_p$. The operators $\hat{\zeta}$ and $\hat{\zeta}'$ are hence unitarily related by $\hat{\zeta}' = U\hat{\zeta}U^\dagger$. In other words, a change to an equivalent connection leads to a unitarily equivalent quantum Hilbert space and quantized observables, a change which leaves the physics content unaltered. We can go a step further. We can identify $\mathscr{H}_p$ with $L^2(Q)$ in two ways: (1) with connection ∇ we can identify $\mathscr{H}_p$ with $L^2(Q)$ by identifying $\Phi \in \mathscr{H}_p$ as given in (2.1) with $\phi(q) \in L^2(Q)$; (2) with connection ∇' we can identify $\mathscr{H}'_p$ with $L^2(Q)$ by $L^2(Q)$ by identifying $\Phi' \in \mathscr{H}'_p$ as given in (2.3) with $\phi(q) \in L^2(Q)$. Then $\hat{\zeta}$ and $\hat{\zeta}'$ both appear as the same operator expression[3]

$$-i\hbar\left(\xi\frac{\partial}{\partial q} + \frac{1}{2}\frac{\partial \xi}{\partial q}\right) + \eta \tag{2.5}$$

in $L^2(Q)$.

3. QUANTIZATION WITH INEQUIVALENT CONNECTIONS

Let Q be the circle $\mathbf{S}^1$, $M = T^*Q = \mathbf{S}^1 \otimes \mathbb{R}$, $\mathbf{B} = M \otimes \mathbb{C}$, and let (p_θ,θ) be the usual coordinates of polar angle θ and its conjugate momentum p_θ on M [strictly θ is not a (global) coordinate but this technicality can be overlooked for our purposes]. The vertical polarization P is spanned by $\partial/\partial p_\theta$.

Let s_0 be any nowhere-zero smooth global section of $\mathbf{B}$ and let ∇ and ∇' be connections on $\mathbf{B}$ defined, respectively, by

$$\nabla_X s_0 = -(i/\hbar)(X \lrcorner\, p_\theta\, d\theta)s_0 \quad \forall X \in V(M)$$

and

$$\nabla'_X s_0 = -(i/\hbar)(X \lrcorner (p_\theta d\theta - \alpha))s_0 \quad \forall X \in V(M),$$

where α is a closed one-form on M. In this case, since M is not simply connected, there exist closed one-forms whch are not exact. In particular let us examine the inexact one-form $\alpha = b\, d\theta$, $b \in \mathbb{R}$.

To simplify things let us restrict ourselves to the half-density quantization scheme here.

With ∇ as connection, the quantum pre-Hilbert space W_p consists of elements of the form

$$\Phi = \phi(\theta)s_0|dp_\theta|^{-1/2}, \tag{3.1}$$

where $\nabla_X s_0 = 0$, $X = \partial/\partial p_\theta$ with finite scalar product

$$\langle \Psi | \Psi \rangle = (2\pi\hbar)^{-1/2} \int_0^{2\pi} |\psi(\theta)|^2\, d\theta. \tag{3.2}$$

The self-adjoint operator expression $\hat{p}_\theta$ in W_p for p_θ is

$$\hat{p}_\theta\Phi = \left[-i\hbar\frac{\partial}{\partial \theta}\phi(\theta)\right]s_0|dp_\theta|^{-1/2}.$$

Let us now try to repeat all this with connection ∇'. We could try to construct the quantum pre-Hilbert space W_p in terms of

$$\Phi' = \phi'(\theta)s_0'|dp_\theta|^{-1/2}, \tag{3.3}$$

where

$$s_0' = (\exp v/i\hbar)s_0, \quad v = b\theta \tag{3.4}$$

as before. Since the function v is not globally smooth we are faced with two possibilities. The first possibility is when $b/\hbar$ happens to be an integer. Then $s_0 = \exp n\theta/i\hbar$ is single valued, i.e., a smooth global section of **B**. Furthermore $\nabla_X' s_0' = 0$, $X = \partial/\partial p_\theta$, on account of (3.2). As a result we can repeat our previous construction to obtain $\mathscr{H}_p'$ and the operator $\hat{p}_\theta'$, and they are unitarily related to $\mathscr{H}_p$ and $\hat{p}_\theta$, respectively. In particular the spectrum of $\hat{p}_\theta$ coincides with that of p_θ'.

The second possibility occurs when $b/\hbar$ is not an integer. Then s_0' in (3.4) is not a smooth global section of **B**. The above construction loses its meaning. To salvage the situation we can introduce

$$\Phi'' = \phi''(\theta)(\exp b\theta/i\hbar)s_0|dp|^{-1/2}, \tag{3.5}$$

which becomes globally smooth if $\phi''(\theta)\exp(b\theta/i\hbar)$ is single values, i.e., $\phi''(\theta)$ must not be single valued in order to maintain the single valueness of $\phi''(\theta)\exp(b\theta/i\hbar)$, hence of Φ''. We can then construct the quantum pre-Hilbert space W_{p_θ}'' and its completion $\mathscr{H}_p''$ in terms of Φ'' as before. The corresponding operator $\hat{p}_\theta''$ in $\mathscr{H}_p''$ possesses a spectrum $n\hbar + b$ which is different from that of $\hat{p}_\theta$. Obviously, $\mathscr{H}_p''$ and $\hat{p}_\theta''$ are not unitarily equivalent to $\mathscr{H}_p$ and $\hat{p}_\theta$ and the two connections ∇, ∇' lead to two physically distinguishable quantum systems.

We see now that when M is not simply connected different choices of connections lead to physically different results. There does not seem to be any *a priori* rules within geometric quantization schemes to single out a "correct" connection. The seemingly natural choice of the connection defined by the connection potential $\beta = p_i\, dq^i$ is not universally applicable since there are known cases where this connection is inappropriate, e.g., the Bohm–Aharonov effect and the examples discussed in later sections. There is one further effect arising from the freedom of choice of connection which seems to have escaped people's attention so far. This is that the choice of connection can blur the distinction between the half-density and the half-form quantization schemes. This will become apparent later.

4. HAMILTONIANS AND B.W.S. CONDITIONS, A GENERAL DISCUSSION

We wish now to study the effect the choice of connection has on the well-known B.W.S. conditions.[1] To this end we propose a quantization scheme to deal with Hamiltonians for one-dimensional systems on the configuration space $Q = \mathbb{R}$ in purely attractive potentials $V(q)$. More precisely $V(q)$ satisfies the following conditions:

$$q\frac{\partial V}{\partial q} > 0 \text{ and } \lim_{q\to+\infty} V(q) = \lim_{q\to-\infty} V(q) = E_0.$$

For convenience we set $q = 0$ at the minimum V_0 of V.

Let X_H be the Hamiltonian vector field on M generated by H and let σ_t be its flow. The flow lines of X_H in M can be found by solving the differential equations

$$\frac{\partial q}{\partial t} = \frac{p}{m}; \quad \frac{\partial p}{\partial t} = -\frac{\partial V}{\partial q}.$$

These flow lines represent the classical orbits of the system. In general an orbit with energy H will be either closed or open and infinite depending on whether $H < E_0$ or $H > E_0$. The flow lines of X_H split M into three distinct regions, labeled by M_c, M_0^+, and M_0^-, where

$$M_c = \{\text{closed flow lines of } X_H\} - \{(0,0)\},$$

$$M_0^+ = \{\text{open flow lines of } X_H; p > 0\},$$

$$M_0^- = \{\text{open flow lines of } X_H; p < 0\}.$$

The origin (0,0) is the critical point of X_H, i.e., $X_H = 0$ at (0,0) and none of the flow lines originated outside meets the origin. We therefore remove the origin from M_c. This is critical later as the vector field X_H spans a polarization in M_c but not M_c plus the origin, and also M_c is not simply connected.

If $V(q)$ is an infinite well, i.e., $E_0 = \infty$, M_0^+ and M_0^- are empty. For a free particle M_c is empty.

The dividing line between M_c and M_0^+ is the line $K_{E_0}^+$ corresponding to $H = E_0; p > 0$ while the dividing line between M_c and M_0^- is the line $K_{E_0}^-$ corresponding to $H = E_0; p < 0$. In each region we can label the flow lines according to their energy by K_H^+ in M_0^+, K_H^- in M_0^-, and K_H^c in M_c. We summarize all this in Fig. 1.

In a neighborhood U of each point $m \in M_c \cup M_0^+ \cup M_0^-$, we can find canonical coordinates $(H(p,q), t(p,q))$, where t is the flow parameter along the flow lines of X_H. If $m \in M_0^+$ or M_0^- these coordinates extend to cover all of M_0^+ or M_0^- (set $t = 0$ when $q = 0$).

If $m \in M_c$ there will be a value of t, T say, dependent on which flow line m lies, such that $\sigma_T(m) = m$. For each flow line K_H^c in M_c we can assign a value of T thus constructing a function $T(H)$ on the flow lines K_H^c. Let us call $T(H)$ the period of K_H^c.

If we fix the origin of t in M_c by setting $t = 0$ when $q = 0; p > 0$, we can construct canonical coordinates (R,θ) by

$$\theta = \frac{2\pi t}{T(H)} \quad R = (2\pi)^{-1}\int_{V_0}^{H} T(H)\, dH.$$

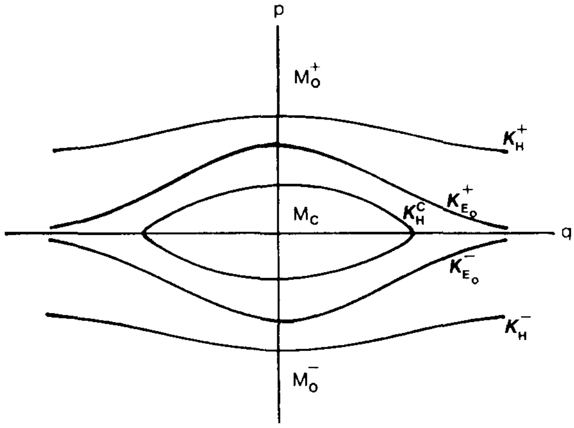


FIG. 1.

These are not global coordinates, so strictly we should define two overlapping charts. This technicality can be overlooked for our purposes.

The advantage of using (R,θ) over (H,t) is that the period of θ, being 2π, is independent of H while the period of t is $T(H)$ and hence, $d\theta$ is a well-defined one-form on M_c while dt is not.

We have three distinct regions of M and the flow σ_t of X_H is complete in each of these three regions. We shall assume that we can quantize the Hamiltonian H separately in these three regons. Such a quantization procedure has been found justifiable for the case of momentum observables by McFarlane and Wan.[6] We are particularly interested in M_c where, as the flow lines of X_H are closed, we can apply B.W.S. conditions[1] to try to derive the discrete part of the spectrum, E_n, of the Hamiltonian.

Let us confine our attention from now on to the region M_c. The prequantum line bundle $\mathbf{B}$ over M can be split into three bundles: $\mathbf{B}_c$ over M_c, $\mathbf{B}_0^+$ over M_0^+, and $\mathbf{B}_0^-$ over M_0^-. We can then choose the connection separately for these three bundles. Since M_0^+, M_0^- are simply connected, the choice of connection in $\mathbf{B}_0^+, \mathbf{B}_0^-$ will have no physical consequences. However, since M_c is not simply connected the choice of connection for $\mathbf{B}_c$ is going to affect the spectrum of the Hamiltonian as we shall see.

Let s_0 by any nonzero section of $\mathbf{B}_c$ and let ∇ be a connection on $\mathbf{B}_c$ defined by

$$\nabla_X s_0 = -(i/\hbar)(X \lrcorner \beta) s_0, \quad X \in V(M_c),$$

where β is a 1-form on M_c such that $d\beta = \omega$ on M_c. Now the vector field $\partial/\partial\theta$ defines a polarization P_c' on M_c. The leaves of P_c' are easily seen to coincide with the flow lines of X_H on M_c. The B.W.S. condition on the leaves K_H^c of P_c' arising from the half-density quantization scheme is

$$\oint_{K_H^c} \beta = 2\pi n\hbar, \quad n = \text{integer}.$$

The corrected B.W.S. condition arising from the half-form quantization scheme[1] is given by

$$\oint_{K_H^c} \beta = 2\pi(n - \tfrac{1}{2})\hbar, \quad n = \text{integer}.$$

We shall now proceed to examine the effect of a change of connection by way of two explicit examples in the next two sections.

5. ONE-DIMENSIONAL SIMPLE HARMONIC OSCILLATOR

The Hamiltonian is given by

$$H = \tfrac{1}{2}(p^2 + q^2),$$

where we have set the mass and the elastic constant to unity.

The vector field X_H is given by

$$X_H = p\frac{\partial}{\partial q} - q\frac{\partial}{\partial p}.$$

The classical orbits satisfying the initial conditions $q(0) = 0$, $p(0) = \sqrt{2H}$ are given by

$$q = \sqrt{2H}\sin t; \quad p = \sqrt{2H}\cos t.$$

Clearly $M_c = M - \{(0,0)\}$ and $T(H) = 2\pi$. The coordinates (R,θ) are then defined by

$$\theta = 2\pi t/T(H) = t,$$

$$R = \frac{1}{2\pi}\int_0^H T(H)\,dH = H.$$

In terms of p and q

$$\theta = \tan^{-1}(q/p); \quad R = \tfrac{1}{2}(p^2 + q^2).$$

Note that R, hence H, is bigger than zero in M_c. Let us choose as connection firstly $\beta = p\,dq$.

The B.W.S. condition

$$\oint_{K_H^c} p\,dq = 2\pi n\hbar, \quad n = \text{integer},$$

is well known to give

$$E_n = n\hbar, \quad n = 1,2,3,\ldots.$$

The allowed values of H are then $E_n = n\hbar$. $n = 1,2,3,\ldots$. The corrected B.W.S. condition meanwhile is known to give

$$E_n = (n - \tfrac{1}{2})\hbar, \quad n = 1,2,3,\ldots.$$

These values form the correct physical spectrum of course.

Thus the corrected B.W.S. condition gives the correct physical results. This is often cited as an example in favor of the half-form scheme over the half-density scheme. However, consider the same example with the connection given by the connection potential

$$\beta' = p(1 + \tfrac{1}{2}\hbar/(p^2 + q^2))dq - \tfrac{1}{2}\hbar(q/(p^2 + q^2))dp,$$

which differs from β by $\tfrac{1}{2}\hbar\,d\theta$, i.e., $\beta' = \beta + \tfrac{1}{2}\hbar\,d\theta$.

The B.W.S. condition now gives

$$\oint_{K_H^c} \beta' = 2\pi n\hbar, \quad n = \text{integer}.$$

The evaluation of the integral serves to illustrate the usefulness of the coordinates (R,θ). Since (as can be verified)

$$\oint_{K_H^c} p\,dq = \int_0^{2\pi} R\,d\theta, \tag{5.1}$$

we have

$$\oint_{K_H^c} \beta' = 2\pi R + \pi\hbar.$$

Hence,

$$E_n = (n - \tfrac{1}{2})\hbar, \quad n = 1,2,3,\ldots.$$

The corrected B.W.S. condition on the other hand gives

$$E_n = n\hbar, \quad n = 1,2,3\ldots.$$

Thus with a different choice of connection, the B.W.S. condition gives the physically correct spectrum while the corrected B.W.S. condition gives a physically incorrect spectrum.

6. THE MODIFIED POSCHL–TELLER POTENTIAL

The modified Pöschl–Teller potential[7] is

$$V(q) = -\tfrac{1}{2}\hbar^2\alpha^2\lambda(\lambda - 1)/\cosh^2(\alpha q),$$

where α, λ are real constants and $\lambda > 1$. This potential is purely attractive and negative with a minimum value V_0 at

$q = 0$ rising to the maximum value zero at infinity. The value V_0 equals $V(0) = -\frac{1}{2}\hbar^2\alpha^2\lambda(\lambda - 1)$. The Hamiltonian is

$$H = \tfrac{1}{2}p^2 - V_0/\cosh^2(\alpha q).$$

We have again set the mass to unity for simplicity. The vector field X_H is

$$X_H = \frac{p\partial}{\partial q} + \frac{2V_0\alpha\sinh(\alpha q)}{\cosh^3(\alpha q)}\frac{\partial}{\partial p}.$$

In this case we have three regions as in Fig. 1. The lines $K^+_{E_0}$ and $K^-_{E_0}$ which serve to divide M into these regions correspond to $H = E_0 = 0$. The regions M_0^+ and M_0^- correspond to the free part of the Hamiltonian. The region M_c contains closed flow lines of X_H and thus we can apply B.W.S. conditions on these flow lines as before.

Explicitly, the flow lines of X_H in M_c are given by

$$\sinh(\alpha q) = ((V_0 - H)/H)^{1/2}\sin((-2H)^{1/2}\alpha t),$$
$$p = (-2H)^{1/2}\cos((-2H)^{1/2}\alpha t)/[H/(V_0 - H) + \sin^2((-2H)^{1/2}\alpha t)]^{1/2}.$$

The function $T(H)$ is given by

$$T(H) = 2\pi/((-2H)^{1/2}\alpha).$$

The coordinates (R,θ) are defined by

$$R = \frac{1}{2\pi}\int_{V_0}^{H} T(H)dH = ((-2V_0)^{1/2} - (-2H)^{1/2})/\alpha, \tag{6.1}$$

$$\theta = 2\pi t/T(H) = (-2H)^{1/2}\alpha t.$$

Let us choose as connection in $\mathbf{B}_c$ the connection with connection potential $\beta = p\,dq$. We again have, as in (5.1),

$$\oint_{K^c_H} p\,dq = \int_0^{2\pi} R\,d\theta = 2\pi R(H).$$

The B.W.S. condition then gives

$$R(E_n) = n\hbar, \quad n = \text{integer},$$

while the corrected B.W.S. condition gives

$$R(E_n) = (n - \tfrac{1}{2})\hbar, \quad n = \text{integer}.$$

The possible values of R in M_c are in the range $(0,(\lambda(\lambda - 1))^{1/2}\hbar)$ on account of (6.1). The allowed values of R by the B.W.S. condition are then

$$R = n\hbar, \quad n = 1,2,\dots \text{ and } n < (\lambda(\lambda - 1))^{1/2},$$

while the allowed values of R by the corrected B.W.S condition are

$$R = (n - \tfrac{1}{2})\hbar, \quad n = 1,2,\dots \text{ and } n < (\lambda(\lambda - 1))^{1/2} + \tfrac{1}{2}.$$

Finally, to calculate the discrete part of the spectrum E_n of the Hamiltonian H we use (6.1) to obtain

$$E_n = -\tfrac{1}{2}\hbar^2\alpha^2((\lambda(\lambda - 1))^{1/2} - n)^2,$$
$$n = 1,2,\dots,n < (\lambda(\lambda - 1))^{1/2}$$

and

$$E_n = -\tfrac{1}{2}\hbar^2\alpha^2((\lambda(\lambda - 1))^{1/2} - n + \tfrac{1}{2})^2,$$
$$n = 1,2,\dots,n < (\lambda(\lambda - 1))^{1/2} + \tfrac{1}{2}$$

for the B.W.S. and the corrected B.W.S. conditions, respectively. The physically correct values[7] are

$$-\tfrac{1}{2}\hbar^2\alpha^2(\lambda - 1 - n)^2, \quad n = 0,1,\dots \text{ and } n \leqslant \lambda - 1.$$

Thus, neither the B.W.S. condition nor the corrected B.W.S. condition gives the correct physical result. In other words the half-form quantization scheme produces incorrect physical results and is thus unsatisfactory in this case when the standard connection is used.

However, let us reconsider the half-density scheme, this time using the connection with potential.

$$\beta' = p\,dq + \hbar\sqrt{\lambda - 1}(\sqrt{\lambda - 1} - \sqrt{\lambda})d\theta.$$

The B.W.S. condition now gives

$$\oint_{K^c_H} (p\,dq + h\sqrt{\lambda - 1}(\sqrt{\lambda - 1} - \sqrt{\lambda})d\theta)$$
$$= 2\pi n\hbar, \quad n = \text{integer},$$
$$\Rightarrow 2\pi R + 2\pi\hbar\sqrt{\lambda - 1}(\sqrt{\lambda - 1} - \sqrt{\lambda}) = 2\pi n\hbar,$$
$$n = 0,1,2\dots,n < \lambda - 1.$$

The allowed values for R are then

$$R = (n - \sqrt{\lambda - 1}(\sqrt{\lambda - 1} - \sqrt{\lambda}))\hbar,$$
$$n = 0,1,2\dots \text{ and } n < \lambda - 1.$$

Consequently we obtain

$$E_n = -\tfrac{1}{2}\hbar^2\alpha^2(\lambda - 1 - n)^2, \quad n = 0,1,2\dots,n < \lambda - 1,$$

which are precisely the desired values.

We could produce the same values using the corrected B.W.S. condition by choosing the connection with potential

$$\beta'' = p\,dq + \hbar(\lambda - \tfrac{1}{2} - \sqrt{\lambda - 1}\sqrt{\lambda})d\theta.$$

7. CONCLUSION

There are two features which emerge from the examples discussed: (1) the B.W.S. condition and the corrected B.W.S. condition with the standard connection give, in general, physically incorrect energy eigenvalues while the judicious choice of connection for either condition can give physically correct eigenvalues; (2) the corrected B.W.S. condition can be replaced by the B.W.S. condition plus an appropriate choice of connection. While the general validity of (1) above has been shown through the examples discussed the general validity of (2) has not been established. But it is not unreasonable to suspect that (2) above may be true for a large class of physically relevant cases. If this is so, then the enormously complex apparatus of half-form quantization schemes appears to be an unnecessary complication, especially in the eyes of physicists.

In view of what has been said, it is clear that the choice of connection in geometric quantization is of physical importance. We have gone some way to give a discussion of this. Perhaps we should mention also that the situation involving the use of a connection different from the standard one to model the Bohm–Aharonov effect[1] is quite a different thing altogether. The Bohm–Aharonov effect arises from the multiply connectedness of the configuration space (and subsequently the phase space), and the connection can be chosen *a priori* on a physical basis by relating it to the magnetic field which gives rise to the effect. In our examples there does not seem to be any *a priori* criteria for selecting the connection. There is obviously room for much further work in this area.

[1]N. Woodhouse, *Geometric Quantization* (Oxford U.P. Oxford, 1980), Chaps. 4, 5, and 6, p. 119, p. 296; J. Sniatycki, *Geometric Quantization and Quantum Mechanics* (Springer, New York, 1980).
[2]C. von Westenholz, *Differential Forms in Mathematical Physics* (North-Holland, Amsterdam, 1981), p. 165, p. 293.
[3]K. McFarlane and K. K. Wan, Int. J. Theor. Phys. **22**, 55 (1983).
[4]E. Prugovecki, *Quantum Mechanics in Hilbert Space* (Academic, New York, 1971).
[5]P. Roman, *Some Modern Mathematics for Physicists and Other Outsiders* (Pergamon, New York, 1975), Vol. 1, pp. 551–553, p. 564.
[6]K. McFarlane and K. K. Wan, J. Phys. A **14**, L1 (1981).
[7]S. Flugge, *Practical Quantum Mechanics* (Springer, New York, 1974), p. 96.

The propagator of the time-dependent forced harmonic oscillator with constant damping

Bin Kang Cheng[a)]
Departamento de Física, Universidade Federal do Paraná, 80.000 Curitiba, Paraná, Brazil

(Received 6 April 1983; accepted for publication 14 October 1983)

The work of Montroll in deriving the propagator of a time-dependent harmonic oscillator is generalized to obtain the propagator of the time-dependent forced harmonic oscillator with constant damping.

PACS numbers: 03.65.Db, 02.90. + p

I. INTRODUCTION

From Feynman's formulation of nonrelativistic quantum mechanics the propagator, probability amplitude for a particle to go from the point (x',t') to the point (x'',t''), can be expressed as[1,2]

$$K(x'',t'';x',t')=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{\frac{i}{\hbar}\int_{t'}^{t''}L(x,\dot{x},t)dt\right\}Dx(t), \tag{1.1}$$

where $L(x,\dot{x},t)$ is the Lagrangian of the dynamical system considered and where $Dx(t)$ is designed to indicate that the integral is over all paths with fixed end points (x',t') and (x'',t''). As is well known that for the quadratic Lagrangian, the propagator has the form[3,4]

$$K(x'',t'';x',t')=\left\{\left|\frac{\partial^2 S_{\rm cl}}{\partial x'\,\partial x''}\right|\Big/2\pi i\hbar\right\}^{1/2}\exp\left\{\frac{iS_{\rm cl}(x'',t'';x',t')}{\hbar}\right\}, \tag{1.2}$$

where $S_{\rm cl}(x'',t'';x',t')$ is the classical action. Here we have assumed that there exists one and only one classical path which passes through (x',t') and (x'',t''). Therefore, the catastrophic phenomenon[5,6] will not be discussed in the present work.

For the time-dependent harmonic oscillator, Montroll[7] firstly transforms the path integral (1.1) into the Gaussian integral

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\{i(Y^TAY+2B^TY)\}\prod_{j=1}^{n}dy_j=(i\pi)^{n/2}(\det A)^{-1/2}\exp(-iB^TA^{-1}B), \tag{1.3}$$

multiplied by a function of x',x'', and τ, where A is an $n\times n$ matrix, Y and B are $n\times 1$ matrices, and Y^T and B^T are, respectively, the transpose matrices of Y and B. He than carries out the calculations as $\tau\to 0$ (or $n\to\infty$). His method has recently been applied for evaluating the propagator of the time-dependent forced harmonic oscillator.[8] In this paper the same method has been generalized to calculate the propagator of the time-dependent forced harmonic oscillator with constant damping term.

In Sec. II we are able to transform our path integral into the Gaussian integral (1.3) multiplied by a function of x', x'', and τ. In Sec. III we show the details of our calculation (also in the Appendix) as $\tau\to 0$ and we also write down the propagator in terms of two functions, which are, respectively, the solutions of the equation of motion of the time-dependent harmonic oscillator with damping and with antidamping. Finally, we discuss the result in Sec. IV and confirm our result for perturbative force being zero by directly calculating (1.2) in Sec. V.

II. FORMULATION

For the time-dependent forced harmonic oscillator with constant damping term, the equation of motion becomes

$$\ddot{x}+\gamma\dot{x}+\omega^2(t)x=q(t)/m, \tag{2.1}$$

where $\omega(t)$ is time-dependent frequency, $q(t)$ time-dependent perturbative force, and γ a constant damping coefficient. Equation (2.1) can easily be obtained by the Lagrangian[9]

$$L(x,\dot{x},t)=e^{\gamma t}\{m[\dot{x}^2-\omega^2(t)x^2]/2+q(t)x\}. \tag{2.2}$$

In spite of its interpretation difficulties in quantum mechanics,[10,11] we are going to use (2.2) as our Lagrangian. Hence the propagator defined by (1.1) can be written as

$$K(x'',t'';x',t')=\lim_{n\to\infty}\left[\prod_{j=1}^{n}\left(\frac{me^{\gamma t_j}}{2\pi i\hbar\tau}\right)^{1/2}\right]$$
$$\times\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{\frac{i\tau}{2\hbar}\left[m\tau^{-2}\sum_{j=1}^{n}e^{\gamma t_j}(x_j-x_{j-1})^2-m\sum_{j=0}^{n-1}e^{\gamma t_j}\omega_j^2x_j^2+2\sum_{j=0}^{n-1}e^{\gamma t_j}q_jx_j\right]\right\}\prod_{j=1}^{n-1}dx_j \tag{2.3}$$

[a)] Work supported in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil.

by Feynman's definition. The extra factor $e^{\gamma t_j}$ is necessary for including dissipative effect.[12] For latter convenience we have set $\tau = (t'' - t')/n$ and $r_j = r(t' + j\tau)$, $r' = r(t')$, and $r'' = r(t'')$ for any function $r(t)$. Now we let $y_j = e^{\gamma t_j/2}(m/2\hbar\tau)^{1/2}x_j$; then Eq. (2.3) can be rewritten as

$$K(x'',t'';x',t') = \lim_{n\to\infty} (i\pi)^{-n/2}\left(\frac{me^{\gamma t''}}{2\hbar\tau}\right)^{1/2}\exp\left\{\frac{i\tau}{2\hbar}[2e^{\gamma t'}q'x' + m\tau^{-2}(e^{\gamma t_1}x'^2 + e^{\gamma t''}x''^2) - me^{\gamma t'}\omega'^2x'^2]\right\}$$

$$\times\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{i\left[\sum_{j=1}^{n-1}(1 + e^{\gamma\tau} - \omega_j^2\tau^2)y_j^2 - 2\sum_{j=0}^{n-1}e^{\gamma\tau/2}y_j\,y_{j+1} + \left(\frac{2\tau^3}{m\hbar}\right)^{1/2}\sum_{j=1}^{n-1}e^{\gamma t_j/2}q_jy_j\right]\right\}\prod_{j=1}^{n-1}dy_j \tag{2.4}$$

since $dx_j = e^{-\gamma t_j/2}(2\hbar\tau/m)^{1/2}dy_j$.

By comparing (1.3) and (2.4) we discover that the matrix A is of the form

$$A = \begin{pmatrix} a_1 & -d & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -d & a_2 & -d & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -d & a_3 & -d & \cdots & 0 & 0 & 0 & 0 \\ \vdots & & & & & & & & \\ 0 & 0 & 0 & 0 & \cdots & -d & a_{n-3} & -d & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -d & a_{n-2} & -d \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -d & a_{n-1} \end{pmatrix}, \tag{2.5}$$

with $a_j = 1 + e^{\gamma\tau} - \omega_j^2\tau^2$ and $d = e^{\gamma\tau/2}$. The column matrix B has the elements

$$b_1 = -y'e^{\gamma\tau/2} + (\tau^3/2m\hbar)^{1/2}Q_1 = -c\tau^{-1/2}e^{\gamma t_1/2}x' + a\tau^{3/2}Q_1, \tag{2.6}$$

$$b_j = a\tau^{3/2}Q_j \quad (j = 2,3,\ldots, n-2), \tag{2.7}$$

and

$$b_{n-1} = -y''e^{\gamma\tau/2} + (\tau^3/2m\hbar)^{1/2}Q_{n-1} = -c\tau^{-1/2}e^{\gamma(t''+\tau)/2}x'' + a\tau^{3/2}Q_{n-1}. \tag{2.8}$$

Here, we have set $Q_j = e^{\gamma t_j/2}q_j$, $c = (m/2\hbar)^{1/2}$ and $a = (2m\hbar)^{-1/2}$. By substituting (1.3) into (2.4) we obtain

$$K(x'',t'';x',t') = \lim_{\tau\to\infty}\left(\frac{me^{\gamma t''}}{2\pi i\hbar\tau\,\det A}\right)^{1/2}\exp\{iB(x'',x',\tau)\} \tag{2.9}$$

and

$$B(x'',x',\tau) = (m/2\hbar\tau)(e^{\gamma t_1}x'^2 + e^{\gamma t''}x''^2) - B^TA^{-1}B. \tag{2.10}$$

We have assumed the factor $\exp\{(i\tau/2\hbar)(-me^{\gamma t'}\omega'^2x'^2 + 2e^{\gamma t'}q'x')\}$ in (2.4) to be one as $\tau\to 0$ (or $n\to\infty$). Now we are only left to calculate the limit values of $\tau\det A$ and $B(x'',x',\tau)$ as $\tau\to 0$. With the help of Eqs. (2.5)–(2.10), these calculations will be carried out in the next section and in the Appendix.

III. CALCULATION

From the matrix A we define A_j and D_j as the following determinants

$$A_1 = a_1,\quad A_2 = \begin{vmatrix} a_1 & -d \\ -d & a_2 \end{vmatrix},\quad A_3 = \begin{vmatrix} a_1 & -d & 0 \\ -d & a_2 & -d \\ 0 & -d & a_3 \end{vmatrix},\ldots,\quad A_{n-1} = \det A,$$

$$D_{n-1} = a_{n-1},\quad D_{n-2} = \begin{vmatrix} a_{n-2} & -d \\ -d & a_{n-1} \end{vmatrix},\quad D_{n-3} = \begin{vmatrix} a_{n-3} & -d & 0 \\ -d & a_{n-2} & -d \\ 0 & -d & a_{n-1} \end{vmatrix},\ldots,\quad D_1 = \det A.$$

It is easy to see that the A_j and the D_j satisfy the recurrence relations

$$A_{j+1} = a_{j+1}A_j - d^2A_{j-1},\quad A_0 = 1\quad (1\leqslant j\leqslant n-2) \tag{3.1}$$

and

$$D_{j-1} = a_{j-1}D_j - d^2d_{j+1},\quad D_n = 1\quad (2\leqslant j\leqslant n-1). \tag{3.2}$$

Furthermore, (3.1) and (3.2) can easily be transformed into the finite-difference equations

$$(D_{j+1} - 2D_j + D_{j-1})/\tau^2 = -\omega_{j-1}^2D_j - \gamma(D_{j+1} - D_j)/\tau \tag{3.3}$$

and

$$(A_{j+1} - 2A_j + A_{j-1})/\tau^2 = -\omega_{j+1}^2 A_j + \gamma(A_j - A_{j-1})/\tau, \tag{3.4}$$

respectively. From (3.1) and (3.2) we see that the end conditions of A_j and D_j are

$$D_{n-1} = a_{n-1} \approx 1 + O(\tau) \approx a_1 = A_1, \tag{3.5}$$

$$(D_{n-1} - D_{n-2})/\tau = \{a_{n-1}(1 - a_{n-2}) + d^2\}/\tau \approx -(1/\tau), \tag{3.6}$$

and

$$(A_2 - A_1)/\tau = \{a_1(a_2 - 1) - d^2\}/\tau \approx 1/\tau \tag{3.7}$$

for τ being very small. In order to overcome the difficulties of divergence in (3.6) and (3.7), we now introduce f_j and g_j by

$$f_j = \tau D_j \quad \text{and} \quad g_j = \tau A_j. \tag{3.8}$$

Equations (3.3) and (3.4) can then be rewritten as the differential equations

$$\ddot{f} + \gamma \dot{f} + \omega^2(t) f = 0, \quad f'' = 0, \ \dot{f}'' = -1 \tag{3.9}$$

and

$$\ddot{g} - \gamma \dot{g} + \omega^2(t) g = 0, \quad g' = 0, \ \dot{g}' = 1 \tag{3.10}$$

with the help of (3.5) – (3.7) in the limit as $\tau \to 0$. Therefore we have

$$\lim_{\tau \to 0}(\tau \det A) = \lim_{\tau \to 0}(\tau A_{n-1}) = \lim_{\tau \to 0} g_{n-1} \approx g(t'') = g'' \tag{3.11}$$

or

$$\lim_{\tau \to 0}(\tau \det A) = \lim_{\tau \to 0}(\tau D_1) = \lim_{\tau \to 0} f_1 \approx f(t') = f'. \tag{3.12}$$

From (3.1), (3.2), and (3.8) we find that the f_j and g_j are related through the formula

$$f_{j+1}\, g_j - d^2 f_{j+2} g_{j-1} = f_j g_{j-1} - d^2 f_{j+1} g_{j-2} = \tau^2 \det A = \tau f_1 = \tau g_{n-1}. \tag{3.13}$$

Hence

$$\begin{aligned} g_j &= \tau f_1 f_{j+2}(f_{j+1} f_{j+2})^{-1} + d^2 f_{j+2} g_{j-1}/f_{j+1} = \tau f_1 f_{j+2}\{(f_{j+1} f_{j+2})^{-1} + d^2 (f_j f_{j+1})^{-1}\} \\ &\quad + d^4 f_{j+1} g_{j-2}/f_j = \cdots = \tau f_1 f_{j+2} \sum_{k=1}^{j+1} (f_k f_{k+1})^{-1} d^{2(j-k+1)}. \end{aligned} \tag{3.14}$$

Specially, we have

$$\tau^2 \sum_{k=1}^{n-1} (f_k f_{k+1})^{-1} d^{2(n-k)} = d^2 g_{n-2}/g_{n-1} \tag{3.15}$$

with the help of (3.14) for $j = n - 2$. By using (2.5) and (3.14) we obtain

$$B^T A^{-1} B = \sum_{j=1}^{n-1} (f_j f_{j+1} d^{2j})^{-1} \left(\sum_{k=j}^{n-1} b_k f_{k+1} d^k \right)^2. \tag{3.16}$$

The above relation has been shown in the Appendix.

By substituting (2.6)–(2.8) into (3.16), we get

$$\begin{aligned} B^T A^{-1} B &= (f_1 f_2 d^2)^{-1} \left\{ a\tau^{3/2} \sum_{k=1}^{n-1} Q_k f_{k+1} d^k - c\tau^{-1/2} \left[e^{\gamma t_1/2} f_2\, dx' + e^{\gamma(t'' + \tau)/2} f_n d^{n-1} x'' \right] \right\}^2 \\ &\quad + \sum_{j=2}^{n-2} (f_j f_{j+1} d^{2j})^{-1} \left\{ a\tau^{3/2} \sum_{k=j}^{n-1} Q_k f_{k+1} d^k - c\tau^{-1/2} e^{\gamma(t'' + \tau)/2} f_n d^{n-1} x'' \right\}^2 \\ &\quad + [f_{n-1} f_n d^{2(n-1)}]^{-1} \{ a\tau^{3/2} Q_{n-1} f_n - c\tau^{-1/2} e^{\gamma(t'' + \tau)/2} f_n x'' \}^2 d^{2(n-1)} \\ &= \frac{c^2 f_2}{\tau f_1} e^{\gamma t_1} x'^2 + \frac{2c^2}{f_1} e^{\gamma t''} x' x'' + c^2 \tau \left\{ \sum_{k=1}^{n-1} (f_k f_{k+1})^{-1} d^{2(n-k)} \right\} e^{\gamma t''} x''^2 - \frac{2ac}{f_1} \left\{ \sum_{k=1}^{n-1} Q_k \tau f_{k+1} d^k \right\} e^{\gamma t'/2} x' \\ &\quad - 2ac\tau \sum_{j=1}^{n-1} (f_j f_{j+1} d^{2j})^{-1} \left\{ \sum_{k=j}^{n-1} Q_k \tau f_{k+1} d^{n+k} \right\} e^{\gamma t''/2} x'' + a^2 \tau \sum_{j=1}^{n-1} (f_j f_{j+1} d^{2j})^{-1} \left\{ \sum_{k=j}^{n-1} Q_k f_{k+1} d^k \right\}^2 \end{aligned} \tag{3.17}$$

after lengthy but straightforward calculations. From (2.10) and (3.17), $B(x'', x', \tau)$ has the form

$$B(x'', x', \tau) = A_\tau x'^2 + B_\tau x' x'' + C_\tau x''^2 + D_\tau x' + E_\tau x'' + F_\tau.$$

As $\tau \to 0$, we obtain

$$\lim_{\tau \to 0} A_\tau = \lim_{\tau \to 0} \left(\frac{m e^{\gamma t_1}}{2\hbar \tau} \right) \left(1 - \frac{f_2}{f_1} \right) = -\frac{m e^{\gamma t'} \dot{f}'}{2\hbar f'}, \tag{3.18}$$

$$\lim_{\tau\to 0} B_\tau = \lim_{\tau\to 0}\left(-\frac{me^{\gamma t''}}{\hbar f_1}\right) = -\frac{me^{\gamma t''}}{\hbar f'}, \tag{3.19}$$

$$\lim_{\tau\to 0} C_\tau = \lim_{\tau\to 0}\left(\frac{me^{\gamma t''}}{2\hbar\tau}\right)\left(1 - e^{\gamma\tau}\frac{g_{n-2}}{g_{n-1}}\right) = \frac{me^{\gamma t''}}{2\hbar}\left(-\gamma + \frac{\dot g''}{g''}\right), \tag{3.20}$$

$$\lim_{\tau\to 0} D_\tau = \frac{1}{\hbar f'}\int_{t'}^{t''} q(t)f(t)e^{\gamma t}\,dt, \tag{3.21}$$

$$\lim_{\tau\to 0} E_\tau = \frac{e^{\gamma t''}}{\hbar f'}\int_{t'}^{t''} q(t)g(t)dt, \tag{3.22}$$

and

$$\lim_{\tau\to 0} F_\tau = -\frac{1}{m\hbar f'}\int_{t'}^{t''} q(t)f(t)e^{\gamma t}\,dt\int_{t'}^{t} q(\theta)g(\theta)d\theta. \tag{3.23}$$

Equation (3.15) has been used in deriving Eq. (3.20). Equations (3.21)–(3.23) have been shown in the Appendix.

By substituting (3.12) and (3.18)–(3.23) into (2.9), we finally obtain our principal result:

$$K(x'',t'';x',t') = \left(\frac{me^{\gamma t''}}{2\pi i\hbar f'}\right)^{1/2}\times\exp\left\{\frac{m}{2i\hbar f'}\left[\dot f' e^{\gamma t'}x'^2 + 2e^{\gamma t''}x'x'' + (\gamma f' - \dot g'')e^{\gamma t''}x''^2\right]\right\}$$
$$\times\exp\left\{\frac{i}{\hbar f'}\left[x'\int_{t'}^{t''} q(t)f(t)e^{\gamma t}\,dt + e^{\gamma t''}x''\int_{t'}^{t''} q(t)g(t)dt - \frac{1}{m}\int_{t'}^{t''} q(t)f(t)e^{\gamma t}\,dt\int_{t'}^{t} q(\theta)g(\theta)d\theta\right]\right\}. \tag{3.24}$$

Here we assume that $f' = g'' \neq 0$ for excluding catastrophic phenomenon. The propagator has been written in terms of $f(t)$ and $g(t)$ which are, respectively, the solutions of the equation of motion of the time-dependent harmonic oscillator with damping and with antidamping. For $\gamma = 0$, (3.24) is exactly equivalent to (3.13) in Ref. 8 as we expect.

IV. RESULT

It can easily be shown that the solutions of (3.9) and (3.10) are

$$f(t) = s(t)e^{-\gamma(t-t'')/2}\sin[\nu'' - \nu(t)] \tag{4.1}$$

and

$$g(t) = s(t)e^{-\gamma(t'-t)/2}\sin[\nu(t) - \nu'], \tag{4.2}$$

respectively, where $s(t)$ and $\nu(t)$ are the amplitude and the phase of the time-dependent harmonic oscillator with constant damping (or antidamping). In order to satisfy their boundary conditions, we must have

$$\ddot s(t) - s'^2 s^{-3}(t) + \Omega^2(t)s(t) = 0 \tag{4.3}$$

and

$$s^2(t)\dot\nu(t) = s', \tag{4.4}$$

where $\Omega^2(t) = \omega^2(t) - \gamma^2/4$. We also have $s' = s''$, $\dot\nu' = \dot\nu''$ and $s'\dot\nu' = s''\dot\nu'' = 1$ since $f' = g''$. With the help of (4.1)–(4.4), (3.20) can be rewritten as

$$K(x'',t'';x',t') = \left[\frac{me^{\gamma(t'+t'')/2}\dot\nu'}{2\pi i\hbar\sin\Phi(t'',t')}\right]^{1/2}\exp\left\{\frac{m\dot\nu'}{4i\hbar}\left[(2\dot s' - \gamma s')e^{\gamma t'}x'^2 - (2\dot s'' - \gamma s'')e^{\gamma t''}x''^2\right]\right\}$$
$$\times\exp\left\{\frac{im\dot\nu'}{2\hbar}\left[(e^{\gamma t'}x'^2 + e^{\gamma t''}x''^2)\cot\Phi(t'',t') - 2e^{\gamma(t'+t'')/2}x'x''\csc\Phi(t'',t')\right]\right\}$$
$$\times\exp\left\{\frac{i\dot\nu'}{\hbar\sin\Phi(t'',t')}\left[x'e^{\gamma t'/2}\int_{t'}^{t''} q(t)s(t)e^{\gamma t/2}\sin\Phi(t'',t)dt + e^{\gamma t''/2}x''\int_{t'}^{t''} q(t)s(t)e^{\gamma t/2}\sin\Phi(t,t')dt\right.\right.$$
$$\left.\left.-\frac{1}{m}\int_{t'}^{t''} q(t)s(t)e^{\gamma t/2}\sin\Phi(t'',t)dt\int_{t'}^{t} q(\theta)s(\theta)e^{\gamma\theta/2}\sin\Phi(\theta,t')d\theta\right]\right\}, \tag{4.5}$$

with $\Phi(\alpha,\beta) = \nu(\alpha) - \nu(\beta)$ for any two arbitrary times α and β. When the perturbative force is absent, $q(t) = 0$ and $\omega(t)$ is a constant, (4.5) reduced to the propagator evaluated by Papadopoulos.[13] For $\gamma = 0$, (4.6) in Ref. 8 is also reproduced from the above equation. However, we should mention that our result is in agreement with (80) of Khandekar and Lawande.[14] Here we need both $f(t)$ and $g(t)$ to evaluate the propagator. It seems to agree with the idea of Feshbach and Tikochinsky.[15]

V. CLASSICAL PATH

Let us define a new function $F(t)$ as

$$F(t) = s(t)e^{-\gamma(t-t')/2}\sin\Phi(t,t') = g(t)e^{-\gamma(t-t')} \tag{5.1}$$

which satisfies the differential equation

$$\ddot{F} + \gamma\dot{F} + \omega^2(t)F = 0, \quad F' = 0 \text{ and } \dot{F}' = 1. \tag{5.2}$$

Since the Wronskian of $f(t)$ and $F(t)$ at t' is different from zero, $W[f',F'] = f' \neq 0$, they are two linearly independent solutions of (2.1) with $q(t) = 0$. Therefore, the classical path of the time-dependent forced harmonic oscillator with constant damping term has the form[15]

$$\bar{x}(t) = c_1 f(t) + c_2 F(t) + \int_{t'}^{t} \frac{q(\theta)[f(\theta)F(t) - f(t)F(\theta)]d\theta}{f'}, \tag{5.3}$$

where c_1 and c_2 are two arbitrary constants to be determined by the end conditions $\bar{x}' = x'$ and $\bar{x}'' = x''$. By substituting (5.1) into (5.3), we then obtain

$$\bar{x}(t) = \frac{1}{f'}\left\{f(t)\left[x' - e^{\gamma t'}\int_{t'}^{t} q(t)g(t)e^{\gamma t}\,dt\right] + g(t)e^{-\gamma t}\left[e^{\gamma t''}x'' - e^{\gamma t'}\int_{t}^{t''} q(t)f(t)dt\right]\right\} \tag{5.4}$$

after simplifications.

From now on we only consider the case $q(t) = 0$. Hence we have from (5.4)

$$\bar{x}(t) = (1/f')\{f(t)x' - e^{-\gamma(t-t'')}g(t)x''\} \tag{5.5}$$

and

$$\dot{\bar{x}}(t) = (1/f')\{\dot{f}(t)x' - e^{-\gamma(t-t'')}[\dot{g}(t) - \gamma g(t)]x''\}. \tag{5.6}$$

With the help of (2.2), (3.9), (3.10), (5.5), and (5.6), we finally get

$$S_{\rm cl}(x'',t'';x',t') = -(m/2f')[\dot{f}'e^{\gamma t'}x'^2 + 2e^{\gamma t''}x'x'' + (\gamma f' - \dot{g}'')e^{\gamma t''}x''^2] \tag{5.7}$$

and

$$\left|\frac{\partial^2 S_{\rm cl}(x'',t'';x',t')}{\partial x'\,\partial x''}\right| = \frac{me^{\gamma t''}}{f'} \tag{5.8}$$

after lengthy but straightforward calculations. Now by substituting (5.7) and (5.8) into (1.2) we obtain (3.24) with $q(t) = 0$ as we expect.

APPENDIX

The elements of A^{-1}, represented by a_{jk}^{-1}, are determined by finding the cofactor of A. Therefore we have from (2.5) that

$$a_{jk}^{-1} = d^{j-k}A_{k-1}D_{j+1}/D_1 = D^{j-k}g_{k-1}f_{j+1}/\tau f_1, \quad j>k \tag{A1}$$

and

$$a_{jk}^{-1} = d^{k-j}A_{j-1}D_{k+1}/D_1 = d^{k-j}g_{j-1}f_{k+1}/\tau f_1, \quad j\leqslant k. \tag{A2}$$

With the help of (3.14), (A1), and (A2), (3.16) can be obtained,

$$\begin{aligned}
B^TA^{-1}B &= \sum_{j,k=1}^{n-1} b_j a_{jk}^{-1} b_k \\
&= \frac{1}{\tau f_1}\left\{\sum_{k=1}^{n-2} b_k\, g_{k-1} d^{-k} \sum_{j=k+1}^{n-1} b_j f_{k+1} d^j + \sum_{j=1}^{n-1} b_j\, g_{j-1} d^{-j} \sum_{k=1}^{n-1} b_k f_{k+1} d^k\right\} \\
&= \sum_{k=j}^{n-2} b_k f_{k+1} d^{-k}\left\{\sum_{m=1}^{k} (f_m f_{m+1})^{-1} d^{2(k-m)}\right\} \sum_{j=k+1}^{n-1} b_j f_{j+1} d^j \\
&\quad + \sum_{j=1}^{n-1} b_j f_{j+1} d^{-j}\left\{\sum_{m=1}^{j} (f_m f_{m+1})^{-1} d^{2(j-m)}\right\} \sum_{k=j}^{n-1} b_k f_{k+1} d^k \\
&= \sum_{j=1}^{n-1} (f_j f_{j+1} d^{2j})^{-1} \sum_{m=j}^{n-2} \sum_{k=m+1}^{n-1} (b_m f_{m+1} d^m)(b_k f_{k+1} d^k) \\
&\quad + \sum_{j=1}^{n-1} (f_j f_{j+1} d^{2j})^{-1} \sum_{m=j}^{n-1} \sum_{k=m}^{n-1} (b_m f_{m+1} d^m)(b_k f_{k+1} d^k) = \sum_{j=1}^{n-1} (f_j f_{j+1} d^{2j})^{-1}\left\{\sum_{k=j}^{n-1} b_k f_{k+1} d^k\right\}^2,
\end{aligned}$$

after lengthy but straightforward calculations.

By using (3.14), we get

$$\lim_{\tau\to0} D_\tau = \lim_{\tau\to0}\left(\frac{2ac}{f_1}\right)\left\{\sum_{k=1}^{n-1} Q_k \tau f_{k+1} d^k\right\}e^{\gamma t'/2} = \lim_{\tau\to0}\left(\frac{e^{\gamma t'/2}}{\hbar f_1}\right)\times\left\{\sum_{k=1}^{n-1} q_k \tau f_{k+1} e^{\gamma(t'+k\tau)/2} e^{\gamma k\tau/2}\right\} = \lim_{\tau\to0}\left(\frac{1}{\hbar f_1}\right)$$

$$\times\left\{\sum_{k=1}^{n-1} q_k \tau f_k e^{\gamma(t'+k\tau)}\right\} = \frac{1}{\hbar f'}\int_{t'}^{t''} q(t) f(t) e^{\gamma t}\, dt$$

since $t_k = t' + k$,

$$\lim_{\tau\to 0} E_\tau = \lim_{\tau\to 0} (2ac\tau e^{\gamma t''/2}) \sum_{j=1}^{n-1} (f_j f_{j+1} d^{2j})^{-1} \left\{\sum_{k=j}^{n-1} Q_k \tau f_{k+1} d^{n+k}\right\}$$

$$= \lim_{\tau\to 0}\left(\frac{\tau e^{\gamma t''/2}}{\hbar}\right)\left\{Q_{n-1} f_n \tau \sum_{j=1}^{n-1} (f_j f_{j+1})^{-1} d^{2(n-j-1)} + Q_{n-2} f_{n-1}\tau \sum_{j=1}^{n-2} (f_j f_{j+1})^{-1} d^{2(n-j-2)+1} + \cdots + Q_1 f_2 \tau (f_1 f_2)^{-1} d^{n-2}\right\}$$

$$= \lim_{\tau\to 0}\left(\frac{\tau e^{\gamma t''/2}}{\hbar}\right)\left\{\frac{Q_{n-1} f_n g_{n-2}}{f_1 f_n} + \frac{Q_{n-2} f_{n-1} g_{n-3} d}{f_1 f_{n-1}} + \frac{Q_{n-3} f_{n-2} g_{n-4} d^2}{f_1 f_{n-2}} + \cdots + \frac{Q_1 f_2 g_0 d^{n-2}}{f_1 f_2}\right\}$$

$$= \lim_{\tau\to 0}\left(\frac{e^{\gamma t''/2}}{\hbar f_1}\right)\{q_{n-1}\tau g_{n-2} + g_{n-2}\tau g_{n-3} + \cdots + q_1 \tau g_0\} e^{\gamma(n-1)\tau/2} = \frac{e^{\gamma t''}}{\hbar f'}\int_{t'}^{t''} q(t)\, g(t)\, dt,$$

and

$$\lim_{\tau\to 0} F_\tau = -\lim_{\tau\to 0}\left\{a^2\tau \sum_{j=1}^{n-1} (f_j f_{j+1} d^{2j})^{-1}\left[\sum_{k=j}^{n-1} Q_k \tau f_{k+1} d^k\right]^2\right\} = -\lim_{\tau\to 0}\left(\frac{\tau^2}{2m\hbar}\right)\left\{Q_{n-1}^2 f_n^2\, \tau \sum_{j=1}^{n-1} (f_j f_{j+1})^{-1} d^{2(n-j-1)}\right.$$

$$\left. + \left[2Q_{n-1} f_n d^{n-1} + Q_{n-2} f_{n-1} d^{n-2}\right]\tau \sum_{j=1}^{n-2} (f_j f_{j+1})^{-1} d^{n-2} + \cdots + 2\tau \sum_{j=1}^{n-2} Q_{n-j} f_{n-j+1} d^{n-j} + Q_1 f_2 \frac{f_2}{f_1}\right\}$$

$$= -\lim_{\tau\to 0}\left(\frac{\tau^2}{2m\hbar f_1}\right)\left\{\sum_{j=1}^{n-1} Q_j^2 f_{j+1} g_{j-1} + 2Q_{n-1} f_n \sum_{j=2}^{n-1} Q_{n-j} g_{n-j-1} d^{n-j} + \cdots + 2Q_2 f_3 Q_1\right\}$$

$$= -\lim_{\tau\to 0}\left(\frac{1}{m\hbar f_1}\right)\sum_{j=2}^{n-1} q_j \tau f_{j+1} e^{\gamma(t'+j\tau)/2} \sum_{k=1}^{j-1} q_k g_{k-1} \tau e^{\gamma[t'+(j-1)\tau]/2} = -\frac{1}{m\hbar f'}\int_{t'}^{t''} q(t) f(t) e^{\gamma t}\, dt \int_{t'}^{t} q(\theta) g(\theta)\, d\theta.$$

[1] R. P. Feynman, Rev. Mod. Phys. **20**, 367 (1948).

[2] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).

[3] J. M. Van Vleck, Proc. Nat. Acad. Sci. USA **14**, 178 (1928).

[4] C. Morette, Phys. Rev. **81**, 848 (1951); Ann. Phys. (N. Y.) **97**, 367 (1976).

[5] L. S. Schulman, *Techniques and Applications of Path Integration* (Wiley, New York, 1981).

[6] S. Levit, "Semiclassical Approximation to Path Integrals—Phases and Catastrophes," in *Path Integrals and Their Applications in Quantum Statistical and Solid State Physics*, edited by G. J. Papadoplous and J. T. Devreese (Plenum, New York, 1978).

[7] E. W. Montroll, Commn. Pure Appl. Math. **5**, 415 (1952).

[8] B. K. Cheng, Rev. Bras. Fis. **13**, 220 (1983).

[9] P. Havas, Nuovo Cimento Suppl. **5**, 363 (1957).

[10] R. W. Hasse, J. Math. Phys. **16**, 2005 (1975).

[11] D. M. Greenberger, J. Math. Phys. **20**, 762 (1979).

[12] By using Kanai's Hamiltonian $H = (p^2/2m)e^{-\gamma t} + V(x)e^{\gamma t}$, this factor can be obtained directly by evaluating the phase-space path integral, see C. Garrod, Rev. Mod. Phys. **38**, 483 (1966).

[13] G. J. Papadopoulos, J. Phys. A **7**, 209 (1974).

[14] D. C. Khandekar and S. V. Lawande, J. Math. Phys. **20**, 1870 (1979).

[15] H. Feshbach and Y. Tikochinsky, *Contribution to A Festschrift for I. I. Rabi*, edited by L. Motz (New York Academy of Sciences, New York, 1977).

[16] F. Brauer and J. A. Nohel, *Ordinary Differential Equations* (Benjamin, New York, 1973).

Propagation of a Dirac particle. A path integral approach

Takashi Ichinose
Department of Mathematics, Hokkaido University, 060-Sapporo, Japan
Hiroshi Tamura
Department of Physics, Hokkaido University, 060-Sapporo, Japan

(Received 6 December 1983; accepted for publication 13 January 1984)

Rigorous path integral formulas are given which represent, in two space-time dimensions, the fundamental solution of the Cauchy problem for the Dirac equation as well as the retarded and advanced propagators for the Dirac particle. It is also shown that the theory can be applied to a free particle and a particle in a central electric field in four space-time dimensions and reveals some aspects of the path integral. Heuristically discussed is the connection with the phase space path integral or Hamiltonian path integral.

PACS numbers: 03.65.Db, 11.10.Qr

I. INTRODUCTION

Suggested by Dirac's remarks[1] concerning the relation of classical actions to quantum mechanics, Feynman[2,3] made two physical postulates to initiate the idea of path integral in quantum mechanics. In pure-imaginary-time quantum mechanics, Kac[4] has given a mathematical meaning to the path integral in terms of the Wiener measure.[5] Namely he established a path integral formula representing the solution of the heat equation for the Schrödinger operator with a scalar potential, which is called the Feynman–Kac formula.[5] Its further extension to the Schrödinger operator with both scalar and vector potentials is the Feynman–Kac–Itô formula[5] which involves the Itô integral. The Laplace transform of this formula for the Schrödinger operator in four-dimensional space-time with respect to a fifth variable as the proper time gives rise to a path integral representation of the Euclidean propagator for a Klein–Gordon particle in an external electromagnetic field.

The aim of this paper is to make a path integral approach to the Dirac equation[6]

$$\partial_t \phi(t,\mathbf{x}) = [-\boldsymbol{\alpha}(\boldsymbol{\partial} - ie\mathbf{A}(t,\mathbf{x})) - im\beta - ie\Phi(t,\mathbf{x})]\phi(t,\mathbf{x}) \tag{1.1}$$

for a particle of rest mass m and charge e in an external electromagnetic field. Here $\boldsymbol{\alpha} = (\alpha_1,\alpha_2,\alpha_3)$, and the α_j and β are the Dirac matrices. $\Phi(t,\mathbf{x})$ and $\mathbf{A}(t,\mathbf{x})$ are, respectively, the scalar and vector potentials of the field. The natural units are used in which the light velocity c and the constant $\hbar = h/2\pi$ with Planck's constant h equal 1.

In this paper we present a construction of path space measures and establish path integral formulas which represent the fundamental solution of the Cauchy problem for (1.1) as well as the retarded and advanced propagators for a Dirac particle, both in two space-time dimensions. The path space measures constructed are countably additive and differ from the Wiener measure. They have support, when $m > 0$, on the set of the Lipschitz continuous paths with slopes smaller than or equal to the light velocity 1, while, when $m = 0$, on the set of the paths with slopes exactly equal to the light velocity 1. This support property reminds us of the "Zitterbewegung"[7] of the Dirac particle. The theory can also be applied to two cases in four space-time dimensions, the path integral for the free Dirac equation and that for the Dirac equation for a central electric field, for they are reduced to the equations with two independent variables by use of the Radon transform and the spherical coordinates, respectively. The path integral formulas established show a close analogy with the Feynman–Kac formula and the Feynman–Kac–Itô formula for the heat equation. It is further seen by a heuristic argument that they coincide with what is obtained from the phase space path integral[8] or Hamiltonian path integral.[9]

Our construction of each countably additive path space measure follows Nelson's method[10] of construction of the Wiener measure. Crucial is the proof of continuity of a certain linear functional on the Banach space of the continuous functions on the path space defined by use of the fundamental solution for the free Dirac equation, which assures application of the Riesz-type representation theorem. The problem is connected with the L^∞ well-posedness of the Cauchy problem for a hyperbolic system of the first order with two independent variables.

The previous work[11] dealt with related and general problems. The present paper focuses on the Dirac equation in relativistic quantum mechanics. The new characteristic feature is the introduction of *conditional* path space measures so as to make a direct treatment of the fundamental solutions of the Cauchy problems and the propagators. The path integral formulas for the propagators (Theorem 2.3) were quoted without detailed proof in Ref. 12.

Section II is devoted to path integral representations for the fundamental solutions of Cauchy problems with the Dirac equation and the retarded and advanced propagators, in two space-time dimensions. The characteristic functionals of the path space measures are also given. The proof is given in Sec. III. Section IV is concerned with the path integral for the Dirac equation in four space-time dimensions. Section V refers to a heuristic derivation of the path integral for the Dirac equation. All the treatments but in this section are mathematically rigorous.

$\mathbb{C}^d$ is the vector space of complex d-column-vectors and $(\mathbb{C}^d)'$ that of complex d-row-vectors. $M_d(\mathbb{C})$ is a vector space

 0022-2488/84/061810-10$02.50

complex $d\times d$ matrices. The norm of a $d\times d$ matrix $N=(N_{jk})$ is defined by $|N|=\max_{1<j<d}\sum_{k=1}^{d}|N_{jk}|$. $\langle\cdot,\cdot\rangle$ is the bilinear inner product and $(\cdot,\cdot)$ the physicist's inner product.

II. PATH INTEGRAL REPRESENTATIONS

In two space-time dimensions, we give path integral formulas for the fundamental solutions of two Cauchy problems with the Dirac equation.[6] One of them is further used to give path integral formulas for the retarded and advanced propagators. By $|r,s|$ is meant the interval $\{t;\ r\leqslant t\leqslant s\}$ or $\{t;\ r\geqslant t\geqslant s\}$ according to $r<s$ or $r>s$. We use the convention of summation over repeated Greek indices.

A. The fundamental solutions of the Cauchy problems

We consider the Dirac equation (1.1) in two-dimensional space-time. We write $A^0(t,x)=\Phi(t,x)$ and $A^1(t,x)=\mathbf{A}(t,x)$. Set $A_\rho(t,x)=g_{\rho\sigma}A^\sigma(t,x)$, $\rho=0,1$, with the metric tensor

$$(g_{\rho\sigma})=\begin{pmatrix}1&0\\0&-1\end{pmatrix}.$$

Then Eq. (1.1) becomes

$$\partial_t\phi(t,x)=[-\alpha(\partial_x+ieA_1(t,x))-im\beta-ieA_0(t,x)\phi(t,x)],\quad t\in\mathbb{R},\ x\in\mathbb{R}. \tag{2.1}$$

Here α and β are 2×2 Hermitian matrices with $\alpha^2=\beta^2=1$ and $\alpha\beta+\beta\alpha=0$. We assume for simplicity that the $A_\rho(t,x)$, $\rho=0,1$, are real-valued continuous functions in space-time $\mathbb{R}\times\mathbb{R}=\mathbb{R}^2$. One Cauchy problem we consider is that for (2.1) with data $\phi(r,x)=g(x)$.

By putting $x^0=t$ and $x^1=x$, (2.1) is rewritten as

$$iH\phi(x)\equiv[(\partial_0+ieA_0(x))+\alpha(\partial_1+ieA_1(x))+im\beta]\phi(x)=0, \tag{2.2}$$

where $x=(x^0,x^1)\in\mathbb{R}^2$ and $\partial_\rho=\partial/\partial x^\rho$, $\rho=0,1$. H is essentially self-adjoint on $C_0^\infty(\mathbb{R}^2;\mathbb{C}^2)$, and defines[13] a self-adjoint operator in $L^2(\mathbb{R}^2;\mathbb{C}^2)$.

We introduce the proper time[14] τ to consider the other Cauchy problem for

$$\partial_\tau\psi(\tau,x)=-iH\psi(\tau,x),\quad \tau\in\mathbb{R},\ x\in\mathbb{R}^2, \tag{2.3}$$

with data $\psi(r,x)=g(x)$. Let $K^{\mathrm{I}}(s,x;r,y)$ and $K^{\mathrm{II}}(s,x;r,y)$ be the fundamental solutions of the Cauchy problems for (2.1) and (2.2), respectively:

$$\phi(s,x)=\int_{\mathbb{R}}K^{\mathrm{I}}(s,x;r,y)g(y)dy, \tag{2.4}$$

$$\psi(s,x)=(e^{-i(s-r)H}g)(x)=\int_{\mathbb{R}^2}K^{\mathrm{II}}(s,x;r,y)g(y)dy. \tag{2.5}$$

Then they admit the following path integral representations. Set $A(t,x)=(A_0(t,x),A_1(t,x))$. $M_2(\mathbb{C})$ is the space of complex 2×2 matrices.

Theorem 2.1: There exists a unique $\mathscr{S}'(\mathbb{R}\times\mathbb{R};M_2(\mathbb{C}))$-valued countably additive measure $\nu_{s;r}^{\mathrm{I}}$ on the Banach space $C(|r,s|;\mathbb{R})$ of the one-dimensional continuous path $X:|r,s|\to\mathbb{R}$ such that for every $A(t,x)$,

$$\iint_{\mathbb{R}\times\mathbb{R}}\overline{f(x)}K^{\mathrm{I}}(s,x;r,y)g(y)dx\,dy=\int(f,d\nu_{s;r}^{\mathrm{I}}(X)g)\exp\Big[-\mathrm{i}\Big(\int_r^s eA_0(t,X(t))dt+\int_r^s eA_1(t,X(t))dX(t)\Big)\Big], \tag{2.6}$$

with (f,g) in $\mathscr{S}(\mathbb{R};(\mathbb{C}^2)')\times\mathscr{S}(\mathbb{R};\mathbb{C}^2)$. The support of $\nu_{s;r}^{\mathrm{I}}$ is on the set of the Lipschitz continuous paths $X:|r,s|\to\mathbb{R}$ satisfying

for each a, b with $r\leqslant a<b\leqslant s$ when $r<s$
or $r\geqslant a>b\geqslant s$ when $r>s$,
$$|X(b)-X(a)|\leqslant|b-a|\quad [|X(b)-X(a)|=|b-a| \text{ in case } m=0]. \tag{2.7}$$

The set function $\nu_{s,f;r,g}^{\mathrm{I}}$ defined by

$$\nu_{s,f;r,g}^{\mathrm{I}}(\cdot)=\langle\bar f\otimes g,\ \nu_{s;r}^{\mathrm{I}}(\cdot)\rangle=(f,\nu_{s;r}^{\mathrm{I}}(\cdot)g) \tag{2.8}$$

is a complex-valued countably additive measure on the Banach space $C(|r,s|;\mathbb{R})$ with support on the set of the Lipschitz continuous paths X satisfying (2.7) and $X(r)\in\operatorname{supp}g$, $X(s)\in\operatorname{supp}f$.

Theorem 2.2: There exists a unique $\mathscr{S}'(\mathbb{R}^2\times\mathbb{R}^2;M_2(\mathbb{C}))$-valued countably additive measure $\nu_{s;r}^{\mathrm{II}}$ on the Banach space $C(|r,s|;\mathbb{R}^2)$ of the two-dimensional continuous paths $X:|r,s|\to\mathbb{R}^2$, $X(\tau)=(X^0(\tau),X^1(\tau))$, such that for every $A(x)$,

$$(f,e^{-i(s-r)H}g)=\iint_{\mathbb{R}^2\times\mathbb{R}^2}\overline{f(x)}K^{\mathrm{II}}(s,x;r,y)g(y)dx\,dy=\int(f,d\nu_{s;r}^{\mathrm{II}}(X)g)\exp\Big[-i\int_r^s eA_\rho(X(\tau))dX^\rho(\tau)\Big], \tag{2.9}$$

with (f,g) in $\mathscr{S}(\mathbb{R}^2;(\mathbb{C}^2)')\times\mathscr{S}(\mathbb{R}^2;\mathbb{C}^2)$. The support of $\nu_{s;r}^{\mathrm{II}}$ is on the set of the Lipschitz continuous paths $X:|r,s|\to\mathbb{R}^2$ satisfying

for each a,b with $r\leqslant a<b\leqslant s$ when $r<s$
or $r\geqslant a>b\geqslant s$ when $r>s$,
$$X^0(b)-X^0(a)=b-a,\ |X^1(b)-X^1(a)|\leqslant|b-a|\quad [|X^1(b)-X^1(a)|=|b-a| \text{ in case } m=0]. \tag{2.10}$$

The set function $\nu_{s,f;r,g}^{\mathrm{II}}$ defined by

$$\nu_{s,f;r,g}^{\mathrm{II}}(\cdot)=\langle\bar f\otimes g,\ \nu_{s;r}^{\mathrm{II}}(\cdot)\rangle=(f,\nu_{s;r}^{\mathrm{II}}(\cdot)g) \tag{2.11}$$

is a complex-valued countably additive measure on the Banach space $C(|r,s|;\mathbb{R}^2)$ with support on the set of the Lipschitz continuous paths X satisfying (2.10) and $X(r)\in\operatorname{supp}g$, $X(s)\in\operatorname{supp}f$.

Remark 1: $\nu_{s;r}^{\mathrm{I}}$ and $\nu_{s;r}^{\mathrm{II}}$ may be regarded as *conditional* path space measures (cf. conditional Wiener measure[5]). In fact, introduce formal *conditional* path space measures $\nu_{s,x;r,y}^{\mathrm{I}}$ and $\nu_{s,x;r,y}^{\mathrm{II}}$ as

$$\nu_{s,f;r,g}^{\mathrm{I}}(\cdot)=\iint_{\mathbb{R}\times\mathbb{R}}\overline{f(x)}\nu_{s,x;r,y}^{\mathrm{I}}(\cdot)g(y)dx\,dy$$

and

$$\nu_{s,f;r,g}^{\mathrm{II}}(\cdot)=\iint_{\mathbb{R}^2\times\mathbb{R}^2}\overline{f(x)}\nu_{s,x;r,y}^{\mathrm{II}}(\cdot)g(y)dx\,dy.$$

Then Theorems 2.1 and 2.2 look like

$$K^{\mathrm{I}}(s,x;r,y)=\int d\nu_{s,x;r,y}^{\mathrm{I}}(X)\exp\Big[-i\Big(\int_r^s eA_0(t,X(t))dt+\int_r^s eA_1(t,X(t))dX(t)\Big)\Big] \tag{2.6$'$}$$

and

$$K^{\rm II}(s,x;r,y)=\int d\nu^{\rm II}_{s,x;r,y}(X)\exp\left[-i\int_r^s eA_\rho(X(\tau))dX^\rho(\tau)\right]. \tag{2.9$'$}$$

The "support" of $\nu^{\rm I}_{s,x;r,y}$ [resp., $\nu^{\rm II}_{s,x;r,y}$] is on the set of the Lipschitz continuous paths X satisfying (2.7) [resp., (2.10)] and $X(r)=y$, $X(s)=x$.

Remark 2: The characteristic functionals of the path space measures $\nu^{\rm I}_{s,f;r,g}$ and $\nu^{\rm II}_{s,f;r,g}$ are given as follows. T is the time-ordering symbol.

(i)

$$C^{\rm I}_{s,f;r,g}(Y)\equiv\int d\nu^{\rm I}_{s,f;r,g}(X)\exp\left[i\int_r^s X(t)Y(t)dt\right]$$

$$=\int_{\mathbb{R}}\overline{\tilde f\left(p+\int_r^s Y(t)dt\right)}\,T\exp\left\{i\int_r^s\left[\alpha\left(p+\int_r^u Y(t)dt\right)+m\beta\right]du\right\}\tilde g(p)dp,$$

where

$$\tilde f(p)=(2\pi)^{-1/2}\int_{\mathbb{R}}e^{-ixp}f(x)dx.$$

(ii)

$$C^{\rm II}_{s,f;r,g}(Y)\equiv\int d\nu\nu^{\rm II}_{s,f;r,g}(X)\exp\left[i\int_r^s X(\tau)Y(\tau)d\tau\right]$$

$$=\int_{\mathbb{R}^2}\overline{\tilde f\left(p+\int_r^s Y(\tau)d\tau\right)}\,T\exp\left\{i\int_r^s\left[\left(p_0+\int_r^u Y_0(\tau)d\tau\right)+\alpha\left(p_1+\int_r^u Y_1(\tau)d\tau\right)+m\beta\right]du\right\}\tilde g(p)dp,$$

where

$$\tilde f(p)=(2\pi)^{-1}\int_{\mathbb{R}^2}e^{-ixp}f(x)dx$$

and $xp=x^\rho p_\rho$ and $X(\tau)Y(\tau)=X^\rho(\tau)Y_\rho(\tau)$.

Remark 3: Typical paths which have the property (2.7) or (2.10) are the zigzag paths, which reminds us of the "Zitterbewegung"[7] of the Dirac particle.

Remark 4: Feynman and Hibbs[3] give briefly a cryptic description of the fundamental solution $K^{\rm I}_0(s,x;r,y)$ of the Cauchy problem for the *free* Dirac equation in two space-time dimensions (see also Ref. 15 and Ref. 16). It may be interesting to study a relation.

Remark 5: Daletskii[17] treated some related problem but *did* construct no countably additive path space measure.

B. The retarded and advanced propagators

The propagator[18] for a two-space-time-dimensional Dirac particle is a 2×2 matrix-valued function (distribution) which is a solution of the Green's function equation

$$[\gamma^\rho(-i\partial_\rho+eA_\rho(x))+m]S(x,y;m)=\delta(x-y),\quad x,y\in\mathbb{R}^2. \tag{2.12}$$

Here $\gamma^0=\beta$ and $\gamma^1=\beta\alpha$. γ^0 is Hermitian and γ^1 anti-Hermitian with $\gamma^\rho\gamma^\sigma+\gamma^\sigma\gamma^\rho=2g^{\rho\sigma}1$, where $\rho,\sigma=0,1$. The convention of summation over repeated Greek indices is used. The left-hand side of (2.12) is nothing but $\gamma^0 HS(x,y;m)$.

Then the retarded and advanced propagators $S_{\rm ret}(x,y;m)$ and $S_{\rm adv}(x,y;m)$ have the following path integral representations.

Theorem 2.3: For (f,g) in $\mathscr{D}(\mathbb{R}^2;(\mathbb{C}^2)')\times\mathscr{D}(\mathbb{R}^2;\mathbb{C}^2)$,

$$\iint_{\mathbb{R}^2\times\mathbb{R}^2}\overline{f(x)}S_{\rm ret}(x,y;m)g(y)dx\,dy$$

$$=i\int_0^\infty d\tau\int(f,d\nu^{\rm II}_{\tau,0}(X)\gamma^0 g)\exp\left[-i\int_0^\tau eA_\rho(X(s))dX^\rho(s)\right] \tag{2.13}$$

and

$$\iint_{\mathbb{R}^2\times\mathbb{R}^2}\overline{f(x)}S_{\rm adv}(x,y;m)g(y)dx\,dy$$

$$=-i\int_{-\infty}^0 d\tau\int(f,d\nu^{\rm II}_{\tau,0}(X)\gamma^0 g)\exp\left[-i\int_0^\tau eA_\rho(X(s))dX^\rho(s)\right]. \tag{2.14}$$

Remark 1: With the formal conditional path space measure $\nu^{\rm II}_{\tau,x;0,y}$ introduced in Remark 1 to Theorems 2.1 and 2.2, (2.13) and (2.14) become

$$S_{\rm ret}(x,y;m)$$

$$=i\int_0^\infty d\tau\int d\nu^{\rm II}_{\tau,x;0,y}(X)\gamma^0\exp\left[-i\int_0^\tau eA_\rho(X(s))dX^\rho(s)\right] \tag{2.13$'$}$$

and

$$S_{\rm adv}(x,y;m)$$

$$=-i\int_{-\infty}^0 d\tau\int d\nu^{\rm II}_{\tau,x;0,y}(X)\gamma^0\exp\left[-i\int_0^\tau eA_\rho(X(s))dX^\rho(s)\right]. \tag{2.14$'$}$$

Remark 2: For the Feynman propagator $S_F(x,y;m)$ we have not such a neat formula.

Proof of Theorem 2.3: We prove only (2.13) for the retarded propagator. The proof for the advanced propagator is similar. Let (f,g) be in $\mathscr{D}(\mathbb{R}^2;(\mathbb{C}^2)')\times\mathscr{D}(\mathbb{R}^2;\mathbb{C}^2)$. Note that H is a self-adjoint operator in $L^2(\mathbb{R}^2;\mathbb{C}^2)$ and

$$(\epsilon + iH)^{-1} = \int_0^\infty d\tau\, e^{-\epsilon\tau} e^{-i\tau H}.$$

Then by Theorem 2.2 we have

$$(f, i(\epsilon + iH)^{-1}\gamma^0 g) = i\int_0^\infty d\tau\, e^{-\epsilon\tau}\int (f, d\nu^{\mathrm{II}}_{\tau;0}(X)\gamma^0 g) \times \exp\left[-i\int_0^\tau eA_\rho(X(s))dX^\rho(s)\right]. \quad (2.15)$$

Since f and g have compact support, the τ-integral on the right-hand side of (2.15) is reduced to that over a finite interval in view of the support property of $\nu^{\mathrm{II}}_{f,\tau;0,g}$. Therefore, it converges to the right-hand side of (2.13) as $\epsilon\downarrow 0$, by the Lebesgue bounded convergence theorem. On the other hand, by definition of the retarded propagator the left-hand side of (2.15) converges to that of (2.13) as $\epsilon\downarrow 0$. This proves Theorem 2.3.

III. PROOFS OF THEOREMS 2.1 AND 2.2

We shall prove only Theorem 2.2. Theorem 2.1 will be shown similarly. Without loss of generality we may assume $r = 0$ to construct $\nu^{\mathrm{II}}_{s;0}$. The proof consists of three parts. First we construct the path space measure $\nu^{\mathrm{II}}_{s;0}$ and then study its support property. Finally we establish the formula (2.9).

A. Construction of the path space measure $\nu^{II}_{s;0}$

We construct $\nu^{\mathrm{II}}_{s;0}$, following Nelson's method[10,5,19] of construction of the Wiener measure.

First consider the Cauchy problem for the free equation to (2.3),

$$\begin{aligned}\partial_\tau\psi(\tau,x) &= -iH_0\psi(\tau,x)\\ &\equiv[-\partial_0 - \alpha\partial_1 - im\beta]\psi(\tau,x),\\ &\tau\in\mathbb{R},\ x = (x^0,x^1)\in\mathbb{R}^2,\end{aligned} \quad (3.1)$$

with initial data $\psi(0,x) = g(x)$. Let $K^{\mathrm{II}}_0(s,x)$ be the fundamental solution

$$\psi(s,x) = (e^{-isH_0}g)(x) = \int_{\mathbb{R}^2} K^{\mathrm{II}}_0(s,x-y)g(y)dy.$$

It is given by

$$K^{\mathrm{II}}_0(s,x) = 2^{-1}\delta(x^0 - s)[\partial_s - \alpha\partial_1 - im\beta] \times (J_0(m(s^2 - |x^1|^2)^{1/2})\theta(s - |x^1|)), \quad (3.2)$$

where $J_0(t)$ is the Bessel function of order zero, and $\theta(t)$ the Heaviside function $\theta(t) = 1$ for $t>0$, $= 0$ for $t<0$. When $m>0$, for each fixed $(s,x)\in\mathbb{R}\times\mathbb{R}^2$, $K^{\mathrm{II}}_0(s-r,x-y)$ has support on the union $\Gamma^{\mathrm{II}}(s,x)$ of the backward and forward conoids of dependence

$$\Gamma^{\mathrm{II}}(s,x) = \{(r,y)\in\mathbb{R}\times\mathbb{R}^2; x^0 - y^0 = s - r, |x^1 - y^1|\leqslant|s-r|\}.$$

When $m = 0$, its support is exactly on the two characteristics. We set

$$R^{\mathrm{II}}(s,x;r) = \Gamma^{\mathrm{II}}(s,x)\cap(\{r\}\times\mathbb{R}^2). \quad (3.3)$$

$C_\infty(\mathbb{R}^2;\mathbb{C}^2)$ denotes the Banach space of the $\mathbb{C}^2$-valued continuous functions in $\mathbb{R}^2$ which vanish at infinity. Its dual space, denoted by $M(\mathbb{R}^2;(\mathbb{C}^2)')$, is the Banach space of the $(\mathbb{C}^2)'$-valued measures on $\mathbb{R}^2$ with bounded variation.[20]

Lemma 3.1: (i) $e^{-i\tau H_0}$ is a continuous linear operator of $C_\infty(\mathbb{R}^2;\mathbb{C}^2)$ into itself, and satisfies

$$\|Ne^{-i\tau H_0}g\|\leqslant e^{m|\tau|}\|Ng\|, \quad (3.4)$$

for g in $C_\infty(\mathbb{R}^2;\mathbb{C}^2)$. Here N is a unitary matrix satisfying

$$N\alpha N^{-1} = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}.$$

(ii) $e^{-i\tau H_0}$ is a continuous linear operator of $M(\mathbb{R}^2;\mathbb{C}^2)$ into itself, and satisfies

$$\|Ne^{-i\tau H_0}\mu\|\leqslant e^{m|\tau|}\|N\mu\|, \quad (3.5)$$

for μ in $M(\mathbb{R}^2;\mathbb{C}^2)$.

(iii) $e^{-i\tau H_0}$ is a continuous linear operator $\mathscr{S}(\mathbb{R}^2;\mathbb{C}^2)$ into itself.

Proof: (i) implies (iii), and also (ii) by duality. Thus we have only to prove (i). Essential is the L^∞ well-posedness of the Cauchy problem for the first-order hyperbolic system with two independent variables,[21] which can be solved along the characteristics. The solution $\psi(s,x)$ of the Cauchy problem for (3.1) satisfies

$$|N\psi(\tau,x)|\leqslant e^{m|\tau|}\max\{|N\psi(0,y)|; y\in R^{\mathrm{II}}(\tau,x;0)\}, \quad (3.6)$$

which implies (3.4).

Now we are ready to construct the path space measure $\nu^{\mathrm{II}}_{s;0}$. Let $\dot{\mathbb{R}}^2 = \mathbb{R}^2\cup\{\infty\}$ be the one-point compactification of $\mathbb{R}^2$, and for each fixed $s\in\mathbb{R}$ let $\mathscr{X}_{s,0} = \Pi_{|0,s|}\dot{\mathbb{R}}^2 = (\dot{\mathbb{R}}^2)^{|0,s|}$ be the product of the uncountably many copies of $\dot{\mathbb{R}}^2$. By the Tychonoff theorem[22] $\mathscr{X}_{s,0}$ is a compact Hausdorff space in the product topology. It may be regarded as the space of all paths $X:|0,s|\to\dot{\mathbb{R}}^2$, possibly discontinuous and possibly passing through the infinity ∞. Let $C(\mathscr{X}_{s,0})$ be the Banach space of the complex-valued continuous functions on $\mathscr{X}_{s,0}$ and $C_{\mathrm{fin}}(\mathscr{X}_{s,0})$ the subspace of those Φ in $C(\mathscr{X}_{s,0})$ for which there exist a finite partition,

$$\begin{aligned}&0 = s_0 < s_1 < \cdots < s_n = s \text{ when } s>0 \text{ or}\\ &0 = s_0 > s_1 > \cdots > s_n = s \text{ when } s<0\end{aligned} \quad (3.7)$$

of the interval $|0,s|$ and a complex-valued bounded continuous function $F(x_{(0)},x_{(1)},\ldots,x_{(n)})$ on $(\mathbb{R}^2)^{n+1}$ such that

$$\Phi(X) = F(X(s_0),X(s_1),\ldots,X(s_n)). \quad (3.8)$$

Define, for each fixed $s\in\mathbb{R}$, a functional $L_{s,0}(\Phi;f,\mu)$ which is linear in $\Phi\in C_{\mathrm{fin}}(\mathscr{X}_{s,0})$ and sesquilinear in $(f,\mu)\in C_\infty(\mathbb{R}^2;(\mathbb{C}^2)')\times M(\mathbb{R}^2;\mathbb{C}^2)$ by

$$\begin{aligned}L_{s,0}(\Phi;f,\mu) = &\overbrace{\int_{\mathbb{R}^2}dx_{(n)}\cdots\int_{\mathbb{R}^2}}^{n}dx_{(1)}\\ &\times\int_{\mathbb{R}^2}\overline{f(x_{(n)})}K^{\mathrm{II}}_0(s_n - s_{n-1},x_{(n)} - x_{(n-1)})\\ &\times K^{\mathrm{II}}_0(s_{n-1} - s_{n-2},x_{(n-1)} - x_{(n-2)})\cdots\\ &\times K^{\mathrm{II}}_0(s_1 - s_0,x_{(1)} - x_{(0)})\\ &\times F(x_{(0)},x_{(1)},\ldots,x_{(n)})d\mu(x_{(0)}).\end{aligned} \quad (3.9)$$

When stressing the sesquilinearity of $L_{s,0}(\Phi;f,\mu)$ we shall write it also $(L_{s,0}\Phi)(f,\mu)$.

Then we can make multiple use of Lemma 3.1 to prove the following lemma, which is of crucial importance in this paper.

Lemma 3.2: (i) For each fixed (f,μ) in $C_\infty(\mathbb{R}^2;(\mathbb{C}^2)')\times M(\mathbb{R}^2;\mathbb{C}^2)$, $L_{s,0}(\Phi;f,\mu)$ is well defined on $C_{\rm fin}(\mathscr{X}_{s,0})$; it is independent of the choice of F corresponding to Φ.

(ii) The following inequality holds:

$$|L_{s,0}(\Phi;f,\mu)|\leqslant Ce^{m|s|}\|\Phi\|\,\|f\|\,\|\mu\| \tag{3.10}$$

for every Φ in $C_{\rm fin}(\mathscr{X}_{s,0})$ and every pair (f,μ) in $C_\infty(\mathbb{R}^2;(\mathbb{C}^2)')\times M(\mathbb{R}^2;\mathbb{C}^2)$, with $C=|N|\,|N^{-1}|\leqslant 2$.

The consequence of Lemma 3.2 is the following. Since $C_{\rm fin}(\mathscr{X}_{s,0})$ is dense in $C(\mathscr{X}_{s,0})$ by the Stone–Weierstrass theorem,[22] the inequality (3.10) holds also for $\Phi\in C(\mathscr{X}_{s,0})$. By Lemma 3.2 $L_{s,0}\Phi$ is a continuous sesquilinear form on $C_\infty(\mathbb{R}^2;(\mathbb{C}^2)')\times M(\mathbb{R}^2;\mathbb{C}^2)$, and so on $\mathscr{S}(\mathbb{R}^2;(\mathbb{C}^2)')\times\mathscr{S}(\mathbb{R}^2;\mathbb{C}^2)$, because both linear embeddings of $\mathscr{S}(\mathbb{R}^2;(\mathbb{C}^2)')$ into $C_\infty(\mathbb{R}^2;(\mathbb{C}^2)')$ and of $\mathscr{S}(\mathbb{R}^2;\mathbb{C}^2)$ into $M(\mathbb{R}^2;\mathbb{C}^2)$ are continuous. Then the kernel theorem[22] enables us to regard $L_{s,0}\Phi$ as an element in the space

$$\mathscr{S}'(\mathbb{R}^2\times\mathbb{R}^2;M_2(\mathbb{C}))=(\mathscr{S}(\mathbb{R}^2;(\mathbb{C}^2)')\otimes\mathscr{S}(\mathbb{R}^2;\mathbb{C}^2))'. \tag{3.11}$$

Here note that $M_2(\mathbb{C})=\mathbb{C}^2\otimes(\mathbb{C}^2)'$. Hence $L_{s,0}$ is a continuous linear mapping of $C(\mathscr{X}_{s,0})$ into the space (3.11). Further $L_{s,0}$ is weakly compact[23] because the space (3.11) is reflexive, and even compact[24] because the space (3.11) is a Montel space. Then the Riesz-type representation theorem[25] yields the following representation of $L_{s,0}$ in terms of a countably additive measure on $\mathscr{X}_{s,0}$.

Theorem 3.3: There exists a unique countably additive measure $\nu^{\rm II}_{s;0}$ defined on the Borel sets in $\mathscr{X}_{s,0}$ and having values in $\mathscr{S}'(\mathbb{R}^2\times\mathbb{R}^2;M_2(\mathbb{C}))$ such that

(a) $\nu^{\rm II}_{s;0}$ is of bounded q-variation for each continuous seminorm q on $\mathscr{S}'(\mathbb{R}^2\times\mathbb{R}^2;M_2(\mathbb{C}))$, i.e.,

$$q\text{-Var }\nu^{\rm II}_{s;0}\equiv\sup q(\Sigma_j\nu^{\rm II}_{s;0}(E_j)c_j)<\infty,$$

where the supremum is taken over all finite partitions $\{E_j\}$ of $\mathscr{X}_{s,0}$ into disjoint Borel sets and all collections $\{c_j\}$ of complex numbers with $|c_j|\leqslant 1$;

(b) for each (f,g) in $\mathscr{S}(\mathbb{R}^2;(\mathbb{C}^2)')\times\mathscr{S}(\mathbb{R}^2;\mathbb{C}^2)$ the set function $\nu^{\rm II}_{s,f;0,g}$ defined by

$$\nu^{\rm II}_{s,f;0,g}(E)=\langle\bar f\otimes g,\nu^{\rm II}_{s;0}(E)\rangle=(f,\nu^{\rm II}_{s;0}(E)g),$$

is a complex-valued countably additive regular measure on the Borel sets E in $\mathscr{X}_{s,0}$;

(c) $\|L_{s,0}\|_q\equiv\sup\{q(L_{s,0}\Phi);\ \|\Phi\|\leqslant 1,\ \Phi\in C(\mathscr{X}_{s,0})\}$

$$=q\text{-Var }\nu^{\rm II}_{s;0}\leqslant C'e^{m|s|}$$

with a constant C' depending only on the seminorm q;

(d) for each Φ in $C(\mathscr{X}_{s,0})$,

$$L_{s,0}\Phi=\int_{\mathscr{X}_{s,0}}d\nu^{\rm II}_{s;0}(X)\Phi(X). \tag{3.12}$$

Remark 1: In terms of the formal conditional path space measure $\nu^{\rm II}_{s,x;r,y}$ in Remark 1 to Theorems 2.1 and 2.2, the expression (3.12) looks like

$$(L_{s,0}\Phi)(x,y)=\int_{\mathscr{X}_{s,0}}d\nu^{\rm II}_{s,x;0,y}(X)\Phi(X). \tag{3.12$'$}$$

Remark 2: There is another way to construct a path space measure. First note that we can define $L_{s,0}(\Phi;f,\mu)$ to establish Lemma 3.2 for the pair $C(\dot{\mathbb{R}}^2;(\mathbb{C}^2)'),M(\dot{\mathbb{R}}^2;\mathbb{C}^2)$ in place of the pair $C_\infty(\mathbb{R}^2;(\mathbb{C}^2)')$, $M(\mathbb{R}^2;\mathbb{C}^2)$. Here $C(\dot{\mathbb{R}}^2;(\mathbb{C}^2)')$ is the Banach space of the $(\mathbb{C}^2)'$-valued continuous functions on $\dot{\mathbb{R}}^2$. Its dual space $M(\dot{\mathbb{R}}^2;\mathbb{C}^2)$ is the Banach space of the $\mathbb{C}^2$-valued measures on $\dot{\mathbb{R}}^2$ with bounded variation.[22] Define $L_{s;0,\mu}$ by

$$(L_{s;0,\mu}\Phi)(f)=L_{s,0}(\Phi;f,\mu),$$

with a fixed $\mu\in M(\dot{\mathbb{R}}^2;\mathbb{C}^2)$, for $\Phi\in C(\mathscr{X}_{s,0})$ and $f\in C(\dot{\mathbb{R}}^2;(\mathbb{C}^2)')$. By the Riesz representation theorem, $L_{s;0,\mu}\Phi$ is regarded as an element in $M(\dot{\mathbb{R}}^2;\mathbb{C}^2)$, so that $L_{s;0,\mu}$ is a continuous linear operator of $C(\mathscr{X}_{s,0})$ into $M(\dot{\mathbb{R}}^2;\mathbb{C}^2)$. Further it is seen[26] that $L_{s;0,\mu}$ is weakly compact. Then by the Riesz-type representation theorem,[27] there exists a unique countably additive measure $\nu^{\rm II}_{s;0,\mu}$ defined on the Borel sets in $\mathscr{X}_{s,0}$ and having values in $M(\dot{\mathbb{R}}^2;\mathbb{C}^2)$ such that

(a) $\nu^{\rm II}_{s;0,\mu}$ is of bounded variation;

(b) for each f in $C(\dot{\mathbb{R}}^2;(\mathbb{C}^2)')$ the set function $\nu^{\rm II}_{s,f;0,\mu}$ defined by

$$\nu^{\rm II}_{s,f;0,\mu}(E)=\langle\bar f,\nu^{\rm II}_{s;0,\mu}(E)\rangle$$

is a complex-valued countably additive regular measure on the Borel sets E in $\mathscr{X}_{s,0}$;

(c) $\|L_{s;0,\mu}\|=\text{Var }\nu^{\rm II}_{s;0,\mu}\leqslant Ce^{m|s|}\|\mu\|$ with $C=|N|\,|N^{-1}|$;

(d) for each Φ in $C(\mathscr{X}_{s,0})$,

$$L_{s;0,\mu}\Phi=\int_{\mathscr{X}_{s,0}}d\nu^{\rm II}_{s;0,\mu}(X)\Phi(X).$$

Notice we can choose, for μ, $\delta^1_y\equiv\begin{pmatrix}\delta_y\\0\end{pmatrix}$ or $\delta^2_y\equiv\begin{pmatrix}0\\\delta_y\end{pmatrix}$, where $\delta_y=\delta(\cdot-y)$ is the Dirac measure at y, an element of $M(\dot{\mathbb{R}}^2)$. Further details are omitted.

B. Support of the path space measure $\nu^{\rm II}_{s;0}$

We shall now see the measure $\nu^{\rm II}_{s;0}$ has the support property as in Theorm 2.2. Let us introduce the *infinity path* X_∞ by $X_\infty(\tau)=\infty$ for every $\tau\in|r,s|$. It suffices then to show the support of $\nu^{\rm II}_{s;0}$ is concentrated on the set

$S(0,s)=\{X\in\mathscr{X}_{s,0};X:|0,s|\to\mathbb{R}^2$ and (2.10) with $r=0$ holds$\}\cup\{X_\infty\}$.

For (2.10) implies the Lipschitz continuity of X, and if f and g are rapidly decreasing we have $(f,\nu_{s,0}(\{X_\infty\})g)=0$. For $0\leqslant a<b\leqslant s$ when $s>0$ or $0\geqslant a>b\geqslant s$ when $s<0$ let $G(a,b)$ be the set of those X in $\mathscr{X}_{s,0}$ for which one of the following two holds:

(i) both $X(a)$ and $X(b)$ are in $\mathbb{R}^2$ and either

$$X^0(b)-X^0(a)\neq b-a \text{ or } |X^1(b)-X^1(a)|>|b-a|;$$

(ii) one of $X(a)$ and $X(b)$ is in $\mathbb{R}^2$ but the other is ∞. Then the $G(a,b)$ are all open in $\mathscr{X}_{s,0}$, and

$$S(0,s)^C=\cup G(a,b),$$

where the union is taken over all a, b with $0\leqslant a<b\leqslant s$ when $s>0$ or $0\geqslant a>b\geqslant s$ when $s<0$.

To show $\nu^{\rm II}_{s;0}$ vanishes on $S(0,s)^C$ we have only to show, by localization principle of a measure,[28] that $\nu^{\rm II}_{s;0}$ vanishes on each $G(a,b)$. Since $C_{\rm fin}(\mathscr{X}_{s,0})$ is dense in $C(\mathscr{X}_{s,0})$ we have only to prove that $L_{s,0}(\Phi;f,g)=0$ for every $\Phi\in C_{\rm fin}(\mathscr{X}_{s,0})$ with supp $\Phi\subseteq G(a,b)$ and for every pair (f,g) in $\mathscr{S}(\mathbb{R}^2;(\mathbb{C}^2)')\times\mathscr{S}(\mathbb{R}^2;\mathbb{C}^2)$. If Φ is in $C_{\rm fin}(\mathscr{X}_{s,0})$ there are a finite partition (3.7) of the interval $|0,s|$ and a bounded continuous function $F(x_{(0)},x_{(1)},...,x_{(n)})$ on $(\dot{\mathbb{R}}^2)^{n+1}$ such that (3.8) holds. We may assume that $s_j = a$, $s_k = b$ for some $j<k$, so that supp F is included in the set of those points $(x_{(0)},x_{(1)},...,x_{(n)})$ in $(\dot{\mathbb{R}}^2)^{n+1}$ for which one of the following two holds:

(i)′ both $x_{(j)} = (x^0_{(j)}, x^1_{(j)})$ and $x_{(k)} = (x^0_{(k)},x^1_{(k)})$ are in $\mathbb{R}^2$ and either $x^0_{(k)} - x^0_{(j)} \neq s_k - s_j$ or $|x^1_{(k)} - x^1_{(j)}| > |s_k - s_j|$;

(ii)′ one of $x_{(j)}$ and $x_{(k)}$ is in the $\mathbb{R}^2$ but the other is ∞. Therefore, especially, if $(x_{(0)}, x_{(1)},...,x_{(n)})$ is a point in $(\mathbb{R}^2)^{n+1}$, there exists an integer l with $j<l \leqslant k$ such that either $x^0_{(l)} - x^0_{(l-1)} \neq s_l - s_{l-1}$ or $|x^1_{(l)} - x^1_{(l-1)}| > |s_l - s_{l-1}|$. By the support property of the fundamental solution $K^{\rm II}_0$ $(s-r,x-y)$, we have $K^{\rm II}_0(s_l - s_{l-1}, x_{(l)} - x_{(l-1)}) = 0$. It follows from (3.9) with $d\mu(x_{(0)}) = g(x_{(0)})dx_{(0)}$ that $L_{s,0}$ $(\Phi;f,g) = 0$.

The assertion for $m=0$ will also be seen from the support property of $K^{\rm II}_0(s-r,x-y)$. This proves the support property of $\nu^{\rm II}_{s;0}$.

C. Proof of the path integral formula (2.9)

We prove the formula (2.9) with $r=0$. We note the proof of the path integral formula (2.6) is analogous, though more complicated because of the t-dependence of $A(t,x)$. The proof of the general case is referred to Ref. 11.

To prove (2.9) with $r=0$, define the operator

$$(T(\tau)g)(x) = \int_{\mathbb{R}^2} K^{\rm II}_0(\tau,x-y)\exp[-ieA_\rho(y)(x^\rho - y^\rho)]g(y)dy \tag{3.13}$$

for g in $C^1_\infty(\mathbb{R}^2;\mathbb{C}^2)$, the Banach space of the $\mathbb{C}^2$-valued continuously differentiable functions in $\mathbb{R}^2$ which together with their first derivatives vanish at infinity.

We need the following lemma.

Lemma 3.4: Assume $A(x)$ is in $C^\infty_0(\mathbb{R}^2;\mathbb{R}^2)$. Then $T(\tau)$ defines a bounded linear operator of $C_\infty(\mathbb{R}^2;\mathbb{C}^2)$ into itself and $\|T(\tau)\| \leqslant Ce^{m|\tau|}$ with a constant C. Further $\partial_\tau(T(\tau)g)$ converges to $-iHg$ in the norm of L^∞ as $\tau \to 0$.

Proof: The first half follows from the L^∞ well-posedness of the Cauchy problem for (3.1). To show the second note that, for fixed $(\tau,x)\in\mathbb{R}\times\mathbb{R}^2$, the support $R^{\rm II}(\tau,x;0)$ in y of $K^{\rm II}_0(\tau,x-y)$ is bounded, and

$$\partial_\tau K^{\rm II}_0(\tau,x-y) = [\partial_{y^0} + \alpha\partial_{y^1} - im\beta]K^{\rm II}_0(\tau,x-y).$$

Thus we have

$$\begin{aligned}\partial_\tau(T(\tau)g)(x) = -\int_{\mathbb{R}^2}\{&K^{\rm II}_0(\tau,x-y)\partial_{y^0}(\exp[-ieA_\rho(y)(x^\rho - y^\rho)]g(y))\\ &+\alpha K^{\rm II}_0(\tau,x-y)\partial_{y^1}(\exp[-ieA_\rho(y)(x^\rho - y^\rho)]g(y))\\ &+im\beta K^{\rm II}_0(\tau,x-y)\exp[-ieA_\rho(y)(x^\rho - y^\rho)]g(y)\}dy.\end{aligned}$$

Then

$$\begin{aligned}&\partial_\tau(T(\tau)g)(x) + i(Hg)(x)\\ &= -\Big\{\int_{\mathbb{R}^2} K^{\rm II}_0(\tau,x-y)\exp[-ieA_\rho(y)(x^\rho - y^\rho)][\partial_{y^0} + ieA_0(y) - i(\partial_{y^0}eA_\rho(y))(x^\rho - y^\rho)]g(y)dy\\ &\quad - (\partial_{x^0} + ieA_0(x))g(x)\Big\}\\ &\quad -\alpha\Big\{\int_{\mathbb{R}^2} K^{\rm II}_0(\tau,x-y)\exp[-ieA_\rho(y)(x^\rho - y^\rho)][\partial_{y^1} + ieA_1(y) - i(\partial_{y^1}eA_\rho(y))(x^\rho - y^\rho)]g(y)dy\\ &\quad - (\partial_{x^1} + ieA_1(x))g(x)\Big\}\\ &\quad -im\beta\Big\{\int_{\mathbb{R}^2} K^{\rm II}_0(\tau,x-y)\exp[-ieA_\rho(y)(x^\rho - y^\rho)]g(y)dy - g(x)\Big\}.\end{aligned}$$

The right-hand side above converges to zero in the norm of L_∞ as $\tau \to 0$. Lemma 3.4 is thus proved.

Now we are in a position to prove the path integral formula (2.9) with $r=0$. By (3.9) in Theorem 3.3 we have for (f,g) in $\mathscr{S}(\mathbb{R}^2;(\mathbb{C}^2)')\times\mathscr{S}(\mathbb{R}^2;\mathbb{C}^2)$ with $s_j = js/n$

$$(f,T(s/n)^n g) = \int (f,d\nu^{\rm II}_{s;0}(X)g)\exp\Big[-i\sum_{j=1}^n A_\rho(X(s_{j-1}))(X^\rho(s_j) - X^\rho(s_{j-1}))\Big]. \tag{3.14}$$

The integrand on the right hand of (3.14) is uniformly bounded, and convergent to $\exp[-i\int_0^s A_\rho(X(\tau))dX^\rho(\tau)]$ as $n\to\infty$ for every Lipschitz continuous path $X:[0,s]\to\mathbb{R}^2$, i.e, for almost every path X, because $\nu^{\rm II}_{s,f;0,g}$ has, as seen in Sec. III B, support on the set of the Lipschitz continuous paths. Thus by the Lebesgue bounded convergence theorem, the right-hand side of (3.14) converges to that of (2.9) as $n\to\infty$.

To show the convergence of the left-hand side of (3.14) we assume first that $A(x)$ is in $C^\infty_0(\mathbb{R}^2;\mathbb{R}^2)$. In fact, we show that $T(s/n)^n g$ converges to $e^{-isH}g$ in the norm of L^∞ as $n\to\infty$.

By Lemma 3.4 we can see that for every g in $C^1_\infty(\mathbb{R}^2;\mathbb{C}^2)$

$$\|\partial_\tau(T(\tau) - e^{-i\tau H})g\| \to 0$$

as $\tau\to 0$. Since

$$\begin{aligned}&\|n(T(s/n) - e^{-i(s/n)H})g\|\\ &= n\Big\|\int_0^{s/n}\partial_\tau(T(\tau) - e^{-i\tau H})g\,d\tau\Big\|\\ &\leqslant s\,\sup_{\tau\in[0,s/n]}\|\partial_\tau(T(\tau) - e^{-i\tau H})g\|,\end{aligned}$$

we have

$$\|n(T(s/n) - e^{-i(s/n)H})g\| \to 0 \text{ as } n\to\infty. \tag{3.15}$$

Further $\{n(T(s/n) - e^{-i(s/n)H})\}_{n=1}^{\infty}$ is a family of bounded operators of $C_{\infty}^{1}(\mathbb{R}^2;\mathbb{C}^2)$ into $C_{\infty}(\mathbb{R}^2;\mathbb{C}^2)$ and for each fixed g in $C_{\infty}^{1}(\mathbb{R}^2;\mathbb{C}^2)$,

$$\|n(T(s/n) - e^{-i(s/n)H})g\| \tag{3.16}$$

is uniformly bounded for n. Therefore, by the uniform boundedness principle,[22] (3.16) is uniformly bounded for both n and g in the unit ball of $C_{\infty}^{1}(\mathbb{R}^2;\mathbb{C}^2)$. It follows that the convergence in (3.15) is uniform on compact subsets of $C_{\infty}^{1}(\mathbb{R}^2;\mathbb{C}^2)$. Since

$$\begin{aligned}&\|(T(s/n)^n - e^{-isH})g\| \\ &= \Big\|\sum_{j=1}^{n} T(s/n)^{j-1}(T(s/n) - e^{-i(s/n)H})e^{-i((n-j)s/n)H}g\Big\| \\ &\leqslant Cne^{m|s|}\sup_{\tau\in|0,s|}\|(T(s/n) - e^{-i(s/n)H})e^{-i\tau H}g\|\end{aligned}$$

with $C = |N|\,|N^{-1}|$ by Lemma 3.1, and the set $\{e^{-i\tau H}g;\ \tau\in|0,s|\}$ is a compact subset of $C_{\infty}^{1}(\mathbb{R}^2;\mathbb{C}^2)$, the theorem is proved for A in $C_0^{\infty}(\mathbb{R}^2;\mathbb{R}^2)$.

Next we assume that A is in $C(\mathbb{R}^2;\mathbb{R}^2)$. Choose a sequence $\{A^{(n)}(x)\}_{n=1}^{\infty}$ in $C_0^{\infty}(\mathbb{R}^2;\mathbb{R}^2)$ such that the $A^{(n)}(x)$ are uniformly bounded on each compact set in $\mathbb{R}^2$ and $A^{(n)}(x)\to A(x)$ on each compact set of $\mathbb{R}^2$ as $n\to\infty$. For each n define a self-adjoint operator $H^{(n)}$ by (2.2) with $A^{(n)}(x)$ in place of $A(x)$. Then for (f,g) in $\mathscr{S}(\mathbb{R}^2;(\mathbb{C}^2)')\times\mathscr{S}(\mathbb{R}^2;\mathbb{C}^2)$

$$\begin{aligned}&(f,\exp[-isH^{(n)}]g) \\ &= \int d\nu_{s,f;0,g}^{\mathrm{II}}(X)\exp\Big[-i\int_0^s eA_p^{(n)}(X(\tau))dX^p(\tau)\Big].\end{aligned} \tag{3.17}$$

Since $A^{(n)}(X(\tau))$ converges to $A(X(\tau))$ for every Lipschitz continuous path $X:|0,s|\to\mathbb{R}^2$ and so for almost every path X because of the support property of $\nu_{s,f;0,g}^{\mathrm{II}}$, the right-hand side of (3.17) converges to that of (2.9) as $n\to\infty$ by the Lebesgue bounded convergence theorem.

On the other hand, $H^{(n)}$ and H are essentially self-adjoint[13] on $C_0^{\infty}(\mathbb{R}^2;\mathbb{C}^2)$. For g in $C_0^{\infty}(\mathbb{R}^2;\mathbb{C}^2)$, $H^{(n)}g$ converges to Hg in $L^2(\mathbb{R}^2;\mathbb{C}^2)$ as $n\to\infty$. It follows[29] that for g in $L^2(\mathbb{R}^2;\mathbb{C}^2)$, $\exp[-isH^{(n)}]g$ converges to e^{-isH} in the norm of L^2 as $n\to\infty$. This proves (2.9), completing the proof of Theorem 2.2.

IV. THE DIRAC EQUATION IN FOUR SPACE-TIME DIMENSIONS

We consider the problem of the path integral for the four-space-time dimensional Dirac equation (1.1). The Dirac matrices α_j and β are 4×4 Hermitian matrices satisfying $\alpha_j^2 = \beta^2 = 1$, $\alpha_j\beta + \beta\alpha_j = 0$, $1\leqslant j\leqslant 3$, and $\alpha_j\alpha_k + \alpha_k\alpha_j = 0$, $j\neq k$.

Our method does not seem to establish the L^{∞} well-posedness to get Lemma 3.2 for the Cauchy problem for the free equation to (1.1),

$$\partial_t\phi(t,\mathbf{x}) = \Big[-\sum_{j=1}^{3}\alpha_j\partial_j - im\beta\Big]\phi(t,\mathbf{x}). \tag{4.1}$$

However, we can deal with two special cases, the path integral for the free Dirac equation and that for the Dirac equation for a central electric field, which are reduced to the equations with two independent variables as considered in Sec. II.

A. The free Dirac equation

We use the Radon transform[30] to reduce the problem with four independent variables to that with two independent variables.[31]

The Radon transform $\hat{g}$ of a function g defined in $\mathbb{R}^3$ is by definition

$$\hat{g}(\xi,\boldsymbol{\omega}) = \int_{\mathbb{R}^3} g(\mathbf{x})\delta(\xi - \mathbf{x}\boldsymbol{\omega})d\mathbf{x},$$

where $\xi\in\mathbb{R}$ and $\boldsymbol{\omega} = (\omega_1,\omega_2,\omega_3)$ is a unit vector in $\mathbb{R}^3$. The following Plancherel theorem holds:

$$\int_{\mathbb{R}^3}\overline{f(\mathbf{x})}g(\mathbf{x})d\mathbf{x} = 2^{-1}(2\pi)^{-2}\int_{|\omega|=1}d\boldsymbol{\omega}\int_{\mathbb{R}}d\xi\overline{\hat{f}_{\xi}(\xi,\boldsymbol{\omega})}\hat{g}_{\xi}(\xi,\boldsymbol{\omega}), \tag{4.2}$$

where $\hat{f}_{\xi} = \partial_{\xi}\hat{f}$ and $\hat{g}_{\xi} = \partial_{\xi}\hat{g}$.

Then the fundamental solution $K_0(t,\mathbf{x}-\mathbf{y})$ of the Cauchy problem for the free Dirac equation (4.1) admits the following path integral representation. Note that there is a unitary matrix $N(\boldsymbol{\omega})$ such that $N(\boldsymbol{\omega})(\Sigma_{j=1}^{3}\alpha_j\omega_j)N(\boldsymbol{\omega})^{-1} = \alpha_1$ and $N(\boldsymbol{\omega})\beta N(\boldsymbol{\omega})^{-1} = \beta$.

Theorem 4.1: There exists a unique $\mathscr{S}'(\mathbb{R}\times\mathbb{R};M_4(\mathbb{C}))$-valued countably additive measure $\nu_{t;0}^{\mathrm{I}}$ on the Banach space $C(|0,t|;\mathbb{R})$ of the one-dimensional continuous paths $\Xi:|0,t|\to\mathbb{R}$ such that for (f,g) in $\mathscr{S}\mathbb{R}^3;(\mathbb{C}^4)')\times\mathscr{S}(\mathbb{R}^3;\mathbb{C}^4)$

$$\begin{aligned}(f(\cdot),\phi(t,\cdot)) &= \iint_{\mathbb{R}^3\times\mathbb{R}^3}\overline{f(\mathbf{x})}K_0(t,\mathbf{x}-\mathbf{y})g(\mathbf{y})d\mathbf{x}\,d\mathbf{y} \\ &= 2^{-1}(2\pi)^{-2}\int_{|\omega|=1}d\boldsymbol{\omega} \\ &\quad\times\int(\hat{f}_{\xi}(\cdot,\boldsymbol{\omega}),N(\boldsymbol{\omega})^{-1}d\nu_{t;0}^{\mathrm{I}}(\Xi)N(\boldsymbol{\omega})\hat{g}_{\xi}(\cdot,\boldsymbol{\omega})).\end{aligned} \tag{4.3}$$

The support of $\nu_{t;0}^{\mathrm{I}}$ is on the set of the Lipschitz continuous paths $\Xi:|0,t|\to\mathbb{R}$ satisfying

$$\begin{aligned}&\text{for each } a,b \text{ with } 0\leqslant a<b\leqslant t \text{ when } t>0 \\ &\text{or } 0\geqslant a>b\geqslant t \text{ when } t<0, \\ &|\Xi(b) - \Xi(a)|\leqslant|b-a| \\ &[|\Xi(b) - \Xi(a)| = |b-a| \text{ in case } m=0].\end{aligned} \tag{4.4}$$

The set function

$$\nu_{t;\hat{f}_{\xi}(\cdot,\omega);0,\hat{g}_{\xi}(\cdot,\omega)}^{\mathrm{I}}(\cdot) = (\hat{f}_{\xi}(\cdot,\boldsymbol{\omega}),N(\boldsymbol{\omega})^{-1}\nu_{t;0}^{\mathrm{I}}(\cdot)N(\boldsymbol{\omega})\hat{g}_{\xi}(\cdot,\boldsymbol{\omega}))$$

is a complex-valued countably additive measure on the Banach space $C(|0,t|;\mathbb{R})$ with support on the set of the Lipschitz continuous paths Ξ satisfying (4.4) and $\Xi(0)\in\operatorname{supp}\hat{g}_{\xi}(\cdot,\boldsymbol{\omega})$, $\Xi(t)\in\operatorname{supp}\hat{f}_{\xi}(\cdot,\boldsymbol{\omega})$.

Proof: The Radon transform of (4.1) yields

$$\partial_t\hat{\phi}(t,\xi,\boldsymbol{\omega}) = \Big[-\Big(\sum_{j=1}^{3}\alpha_j\omega_j\Big)\partial_{\xi} - im\beta\Big]\hat{\phi}(t,\xi,\omega). \tag{4.5}$$

Multiply (4.5) by $N(\boldsymbol{\omega})$ from the left. Then we have

$$\partial_t\eta(t,\xi,\boldsymbol{\omega}) = [-\alpha_1\partial_{\xi} - im\beta]\eta(t,\xi,\boldsymbol{\omega}), \tag{4.5$'$}$$

with $\eta(t,\xi,\boldsymbol{\omega}) = N(\boldsymbol{\omega})\hat{\phi}(t,\xi,\boldsymbol{\omega})$. For $\boldsymbol{\omega}$ fixed, (4.5)$'$ is a first-order hyperbolic system with two independent variables t and ξ. In the same way as in the proof of Theorems 2.1 and 2.2 we can construct the path space measure $\nu_{t;0}^{\mathrm{I}}$ with the property mentioned in Theorem 4.1.

To get (4.3), differentiate by ξ both sides of (4.5). Then if $\hat{\phi}_\xi(t,\xi,\omega) \equiv \partial_\xi \hat{\phi}(t,\xi,\omega)$ is the solution of the Cauchy problem for (4.5) with initial data $\hat{\phi}_\xi(0,\xi,\omega) = \hat{g}_\xi(\xi,\omega) \equiv \partial_\xi \hat{g}(\xi,\omega)$, it has the following path integral representation:

$$(\hat{f}_\xi(\cdot,\omega),\hat{\phi}_\xi(t,\cdot,\omega))$$
$$= \int (\hat{f}_\xi(\cdot,\omega), N(\omega)^{-1}\, d\nu^{\mathrm{I}}_{t;0}(\Xi) N(\omega)\hat{g}_\xi(\cdot,\omega)).$$

The formula (4.3) follows from this with the aid of the Plancherel formula (4.2). This proves Theorem 4.1.

Remark: Formal substitution of $\delta_{\mathbf{x}} = \delta(\cdot - \mathbf{x})$ and $\delta_{\mathbf{y}} = \delta(\cdot - \mathbf{y})$ into f and g yields the following intuitive expression of (4.3):

$$K_0(t,\mathbf{x}-\mathbf{y}) = 2^{-1}(2\pi)^{-2}\int_{|\omega|=1} d\omega \int \delta'(\Xi(t) - \mathbf{x}\omega)$$
$$\times N(\omega)^{-1}\, d\nu^{\mathrm{I}}_{t;0}(\Xi) N(\omega)\delta'(\Xi(0) - \omega\mathbf{y}). \quad (4.3)'$$

B. The Dirac equation for a central electric field

The Dirac equation for a central electric field can be separated in spherical coordinates.[6] The radial Dirac equation is

$$\partial_t \chi(t,r) = -iH^\kappa \chi(t,r), \qquad t \in \mathbb{R},\ r \in (0,\infty),$$
$$H^\kappa = H_0^\kappa + V(r), \quad H_0^\kappa = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\partial_r + \begin{pmatrix} m & \dfrac{\kappa}{r} \\ \dfrac{\kappa}{r} & -m \end{pmatrix}, \tag{4.6}$$

with $V = e\Phi$, where κ is a positive and negative integer. We assume that $V(r)$ is a real-valued continuous function in $(0,\infty)$, and H^κ is self-adjoint[32] in $L^2((0,\infty), dr;\mathbb{C}^2)$. The following theorem is, though of a rather restrictive character, concerned with a path integral representation for the solution $\chi(t,r)$ of the Cauchy problem for (4.6) with initial data $\chi(0,r) = g(r)$.

Theorem 4.2: Let f and g be in $\mathscr{S}(\mathbb{R}^+;(\mathbb{C}^2)')$ and $\mathscr{S}(\mathbb{R}^+;\mathbb{C}^2)$, restrictions of $\mathscr{S}(\mathbb{R};(\mathbb{C}^2)')$ and $\mathscr{S}(\mathbb{R};\mathbb{C}^2)$ to $\mathbb{R}^+ = [0,\infty)$, respectively. If for each s with $0 \leqslant s \leqslant t$ when $t > 0$ or with $0 \geqslant s \geqslant t$ when $t < 0$, the intersection

$$\{r \in \mathbb{R}^+; |x - r| \leqslant |t - s|,\ x \in \operatorname{supp} f\}$$
$$\cap \{r \in \mathbb{R}^+; \quad |r - y| \leqslant |s|,\ y \in \operatorname{supp} g\} \tag{4.7}$$

does not contain 0, there exists a unique complex-valued countably additive measure $\nu^+_{\kappa;t,f;0,g}$ on the set of the one-dimensional continuous paths $R:|0,t| \to \mathbb{R}^+$ such that for every $V(r)$,

$$(f, e^{-itH^\kappa} g)_{L^2((0,\infty),dr)} = \int_{\mathbb{R}^+} \overline{f(r)}\chi(t,r)dr$$
$$= \int d\nu^+_{\kappa;t,f;0,g}(R)\exp\left[-i\int_0^t V(R(s))ds\right]. \tag{4.8}$$

The support of $\nu^+_{\kappa;t,f;0,g}$ is on the set of the Lipschitz continuous paths $R:|0,t| \to \mathbb{R}^+$ satisfying

$$\begin{aligned} &\text{for each } a, b \text{ with } 0 \leqslant a < b \leqslant t \text{ when } t > 0 \\ &\text{or } 0 \geqslant a > b \geqslant t \text{ when } t < 0, \\ &|R(b) - R(a)| \leqslant |b - a| \end{aligned} \tag{4.9}$$

and $R(0) \in \operatorname{supp} g$, $R(t) \in \operatorname{supp} f$.

Proof: We give only an outline of the proof, for it proceeds with a similar argument used in the proof of Theorem 2.2.

Let t, f, and g be as in Theorem 4.2. We consider only the case $t > 0$. The free equation to (4.6) is

$$\partial_s \chi(s,r) = -iH_0^\kappa \chi(s,r), \quad 0 < s \leqslant t, \quad r \in (0,\infty). \tag{4.10}$$

Then for the Cauchy problem for (4.10) we have the following lemma [cf. Lemma 3.1 and (3.6)]. Let $r_0 > 0$ be the minimum of the set (4.7), so that $|\kappa|/r_0$ is an upper bound of $|\kappa|/r$ with r in the set (4.7).

Lemma 4.3: If r is in supp f with $r \geqslant \min(\operatorname{supp} g) - s$, then

$$|N\chi(s,r)| \leqslant e^{Ms} \max\{|N\chi(0,u)|;\quad u \in \operatorname{supp} g, r - s \leqslant u \leqslant r + s\},$$

where $M = m + |\kappa|/r_0$, and $N = 2^{-1/2}\begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}$.

Proof: We only note the following. Multiplying Eq. (4.10) by N from the left, we have with $\eta(t,r) = N\chi(t,r)$,

$$\partial_t \eta(t,r)$$
$$= \left[\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\partial_r + \begin{pmatrix} 0 & -m - \dfrac{i\kappa}{r} \\ m - \dfrac{i\kappa}{r} & 0 \end{pmatrix}\right]\eta(t,r). \tag{4.10$'$}$$

Notice that Lemma 4.3 yields the support property of the fundamental solution $K_0^+(s,r)$ of the Cauchy problem for (4.10).

To construct the path space measure let $\mathscr{R}_{t,0} = \Pi_{[0,t]}\dot{\mathbb{R}}^+$ be the product of the infinitely many copies of $\dot{\mathbb{R}}^+ = \mathbb{R}^+ \cup \{\infty\}$, the one-point compactification of $\mathbb{R}^+$. Let $C(\mathscr{R}_{t,0})$ be the Banach space of the continuous functions on the compact Hausdorff space $\mathscr{R}_{t,0}$, and $C_{\mathrm{fin}}(\mathscr{R}_{t,0})$ its subspace of those Φ for which there exist a finite partition $0 = t_0 < t_1 < \cdots < t_n = t$ of the interval $[0,t]$ and a complex-valued bounded continuous function $F(r_{(0)}, r_{(1)},\ldots,r_{(n)})$ on $(\dot{\mathbb{R}}^+)^{n+1}$ such that $\Phi(R) = F(R(t_{(0)}), R(t_{(1)}),\ldots,R(t_{(n)}))$.

Define a linear functional $L_{\kappa;t,f;0,g}$ on $C_{\mathrm{fin}}(\mathscr{R}_{t,0})$ by

$$L_{\kappa;t,f;0,g}\Phi = \overbrace{\int_{\mathbb{R}^+}\cdots\int_{\mathbb{R}^+}}^{n+1} \overline{f(r_{(n)})}K_0^+(t_n - t_{n-1}, r_{(n)} - r_{(n-1)})\cdots$$
$$\times K_0^+(t_1 - t_0, r_{(1)} - r_{(0)})F(r_{(0)}, r_{(1)},\ldots,r_{(n)})g(r_{(0)})dr_{(0)}\,dr_{(1)}\cdots dr_{(n)}.$$

Then the following lemma will be shown with use of Lemma 4.3 (cf. Lemma 3.2).

Lemma 4.4: $L_{\kappa;t,f;0,g}\Phi$ is well defined on $C_{\mathrm{fin}}(\mathscr{R}_{t,0})$ and

$$|L_{\kappa;t,f;0,g}\Phi| \leqslant 2e^{Mt}\|\Phi\|\,\|f\|_{L^1((0,\infty),dr)}\|g\|_{L^\infty((0,\infty),dr)}$$

for every Φ in $C_{\mathrm{fin}}(\mathscr{R}_{t,0})$, where $M = m + |\kappa|/r_0$.

Since $C_{\text{fin}}(\mathscr{R}_{t,0})$ is dense in $C(\mathscr{R}_{t,0})$, it follows from Lemma 4.4 with the Riesz-type representation theorem[27] that there exists a unique complex-valued countably additive measure $\nu^{+}_{\kappa;t,f;0,g}$ defined on the Borel sets in $\mathscr{R}_{t,0}$ such that

$$L_{\kappa;t,f;0,g}\Phi = \int_{\mathscr{R}_{t,0}} d\nu^{+}_{\kappa;t,f;0,g}(R)\Phi(R)$$

for $\Phi \in C(\mathscr{R}_{t,0})$. The support property of the measure will be seen from that of $K_0^{+}(s,r)$.

Once the path space measure $\nu^{+}_{\kappa;t,f;0,g}$ is constructed the proof of the formula (4.8) will be accomplished as in Sec. IIIC.

Remark 1: The restriction for t and the supports of f and g in Theorem 4.2 mean that the information which starts from g at time 0 to reach f at time t has never passed through the center of the potential. We need it, for the free Dirac equation (4.10) contains the $1/r$ singularity in H_0^{κ}, which invalidates Lemma 4.3.

Remark 2: Even when $m = 0$, it cannot be asserted that the support of $\nu^{+}_{\kappa;t,f;0,g}$ is on the set of those paths $R:[0,t] \to \mathbb{R}^{+}$ satisfying $|R(b) - R(a)| = |b - a|$ instead of (4.9) for the same a,b. The presence of the $i\kappa/r$ term in H_0^{κ} tends to warp the paths.

These facts may suggest that the problem is after all four-space-time-dimensional.

V. HEURISTIC DERIVATION OF PATH INTEGRAL FOR THE DIRAC EQUATION

In this section we shall heuristically see what should be the path integral for the fundamental solution of the Cauchy problem for the Dirac equation (1.1). Our strategy is to exploit the method of phase space path integral[8] or Hamiltonian path integral.[9]

We begin with the action[33] for a relativistic positive-energy (resp., negative-energy) particle of mass m and charge e in an electromagnetic field,

$$S(s,r;\mathbf{P},\mathbf{X}) = \int_r^s [\mathbf{P}(t)\dot{\mathbf{X}}(t) \mp \sqrt{(\mathbf{P}(t) - e\mathbf{A}(t,\mathbf{X}(t)))^2 + m^2} - e\Phi(t,\mathbf{X}(t))]dt. \tag{5.1}$$

Here $\mathbf{X}(t)$ and $\mathbf{P}(t)$ are the position and momentum. We assume for definiteness that $s > r$. In the formulas with double signs, the upper one corresponds to the positive-energy particle and the lower one to the negative-energy particle. The method of phase space path integral or Hamiltonian path integral assumes that the fundamental solution $K(s,\mathbf{x};r,\mathbf{y})$ of the Cauchy problem for the Dirac equation (1.1), which is the probability amplitude that a quantized charged particle with positive energy (resp., negative energy) at position $\mathbf{y}$ at time r will be at position $\mathbf{x}$ at time s, is given by a formal "integral"

$$\int_{\mathscr{P}_{s,\mathbf{x};r,\mathbf{y}}} e^{iS(s,r;\mathbf{P},\mathbf{X})}\mathscr{D}(\mathbf{P})\mathscr{D}(\mathbf{X}). \tag{5.2}$$

Here $\mathscr{D}(\mathbf{P})\mathscr{D}(\mathbf{X})$ is a formal "measure" $\Pi_{t\in[r,s]}(2\pi)^{-3}\,d\,\mathbf{P}(t)d\,\mathbf{X}(t)$ on the space $\mathscr{P}_{s,\mathbf{x};r,\mathbf{y}}$ of the phase space paths $(\mathbf{P}(t),\mathbf{X}(t))$ satisfying $\mathbf{X}(r) = \mathbf{y}$ and $\mathbf{X}(s) = \mathbf{x}$ with $\mathbf{P}(t)$ unrestricted. In this formal phase space path "integral" we first make the change of variables: $\mathbf{X}'(t) = \mathbf{X}(t)$, $\mathbf{P}'(t) = \mathbf{P}(t) - e\mathbf{A}(t,\mathbf{X}(t))$ and next use Dirac's prescription[6]

$$\pm(\mathbf{P}'(t)^2 + m^2)^{1/2} = \boldsymbol{\alpha}\mathbf{P}'(t) + m\beta. \tag{5.3}$$

Then we have [writing $(\mathbf{P}(t),\mathbf{X}(t))$ again instead of $(\mathbf{P}'(t),\mathbf{X}'(t))$]

$$K(s,\mathbf{x};r,\mathbf{y}) = \int_{\mathscr{P}_{s,\mathbf{x};r,\mathbf{y}}} T\exp\left\{i\int_r^s [(\mathbf{P}(t) + e\mathbf{A}(t,\mathbf{X}(t)))\dot{\mathbf{X}}(t) - (\boldsymbol{\alpha}\mathbf{P}(t) + m\beta) - e\Phi(t,\mathbf{X}(t))]dt\right\}\Pi_{t\in[r,s]}(2\pi)^{-3}\,d\mathbf{P}(t)d\mathbf{X}(t), \tag{5.4}$$

where T stands for the time-ordering symbol. We understand (5.4) to be defined with a time division procedure, i.e.,

$$K(s,\mathbf{x};r,\mathbf{y}) = \lim \int_{\mathbf{R}^3}\overbrace{\int_{\mathbf{R}^6}\cdots\int_{\mathbf{R}^6}}^{n-1}\prod_{j=1}^{n}\exp\left\{i\left[(\mathbf{p}_{(j-1)} + e\mathbf{A}(t_{j-1},\mathbf{x}_{(j-1)}))\frac{\mathbf{x}_{(j)} - \mathbf{x}_{(j-1)}}{t_j - t_{j-1}} - (\boldsymbol{\alpha}\mathbf{p}_{(j-1)} + m\beta) - e\Phi(t_{j-1},\mathbf{x}_{(j-1)})\right](t_j - t_{j-1})\right\}$$
$$\times(2\pi)^{-3}\,d\mathbf{p}_{(0)}(2\pi)^{-3}(d\mathbf{p}_{(1)}\,d\mathbf{x}_{(1)})\cdots(2\pi)^{-3}(d\mathbf{p}_{(n-1)}\,d\mathbf{x}_{(n-1)}). \tag{5.5}$$

Note that the integrand of the integral on the right of (5.5) is rewritten as

$$\prod_{j=1}^{n}\{\exp[i\mathbf{x}_{(j)}\mathbf{p}_{(j-1)}]\exp[-i(\boldsymbol{\alpha}\mathbf{p}_{(j-1)} + m\beta)(t_j - t_{j-1})]\exp[-i\mathbf{x}_{(j-1)}\mathbf{p}_{(j-1)}]\}$$
$$\times\exp\left\{i\sum_{k=1}^{n}[e\mathbf{A}(t_{k-1},\mathbf{x}_{(k-1)})(\mathbf{x}_{(k)} - \mathbf{x}_{(k-1)}) - e\Phi(t_{k-1},\mathbf{x}_{(k-1)})(t_k - t_{k-1})]\right\}.$$

Here $r = t_0 < t_1 < \cdots < t_n = s$ and $\mathbf{x}_{(0)} = \mathbf{y}$, $\mathbf{x}_{(j)} = \mathbf{X}(t_j)$, $\mathbf{x}_{(n)} = \mathbf{x}$, and the product $\Pi_{j=1}^{n}$ is time-ordered with time increasing from the right to the left. The limit is taken for $n \to \infty$ and $\max_{1\le j\le n}(t_j - t_{j-1}) \to 0$. If $K_0(t,\mathbf{x})$ is the fundamental solution of the Cauchy problem for the free Dirac equation (4.1), then the $\mathbf{p}_{(j)}$ integrations on the right-hand side of (5.5) yield

$$K(s,\mathbf{x};r,\mathbf{y}) = \lim\overbrace{\int_{\mathbf{R}^3}\cdots\int_{\mathbf{R}^3}}^{n-1}K_0(t_n - t_{n-1},\mathbf{x}_{(n)} - \mathbf{x}_{(n-1)})\cdots K_0(t_1 - t_0,\mathbf{x}_{(1)} - \mathbf{x}_{(0)})$$
$$\times\exp\left\{i\sum_{j=1}^{n}[e\mathbf{A}(t_{j-1},\mathbf{x}_{(j-1)})(\mathbf{x}_{(j)} - \mathbf{x}_{(j-1)}) - e\Phi(t_{j-1},\mathbf{x}_{(j-1)})(t_j - t_{j-1})]\right\}d\mathbf{x}_{(1)}\,d\mathbf{x}_{(2)}\cdots d\mathbf{x}_{(n-1)}.$$

If a path space measure $\nu_{s,\mathbf{x};r,\mathbf{y}}$ should be constructed from the product of K_0's and the Lebesgue measures $d\mathbf{x}_{(j)}$, we should get

$$K(s,\mathbf{x};r,\mathbf{y}) = \int_{\mathscr{X}_{s,\mathbf{x};r,\mathbf{y}}} d\nu_{s,\mathbf{x};r,\mathbf{y}}(\mathbf{X})\exp\Big[-i\int_r^s e\Phi(t,\mathbf{X}(t))dt + i\int_r^s e\mathbf{A}(t,\mathbf{X}(t))d\,\mathbf{X}(t)\Big], \tag{5.6}$$

where $\mathscr{X}_{s,\mathbf{x};r,\mathbf{y}}$ is the space of the configuration paths $\mathbf{X}(t)$ satisfying $\mathbf{X}(r) = \mathbf{y}$ and $\mathbf{X}(s) = \mathbf{x}$. Thus the formula (5.6), in the covariant notation as mentioned before (2.1), is what we have obtained for the Dirac equation (2.1) in two space-time dimensions, that is, (2.6) or (2.6)'.

ACKNOWLEDGMENTS

One of us (T. I.) would like to thank Professor H. Araki, Professor H. Ezawa, and Professor J. R. Klauder for helpful discussions on some related physical problems. The other (H.T.) would like to thank Professor A. Kanazawa for his warm encouragements. We are also grateful to Professor Klauder for his encouraging and critical interest in the present work.

[1]P. A. M. Dirac, *The Principles of Quantum Mechanics* (Clarendon, Oxford, 1935), 2nd ed., Sec. 33; also Phys. Z. Sowjetunion **3**, 64 (1933); Rev. Mod. Phys. **17**, 195 (1945).

[2]R. P. Feynman, "Space-time approach to non-relativistic quantum mechanics," Rev. Mod. Phys. **20**, 367 (1948).

[3]R. P. Feynman and A. P. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).

[4]M. Kac, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (University of California, Berkeley, 1951), p. 189.

[5]See, e.g., B. Simon, *Functional Integration and Quantum Physics* (Academic, New York, 1979).

[6]See, e.g., J. D. Bjorken and S. D. Drell, *Relativistic Quantum Mechanics* (McGraw-Hill, New York, 1964).

[7]E. Schrödinger, Sitzungsber. Preuss. Akad. Wiss. Phys.-Math. Kl. **24**, 418 (1930); see also Ref. 6, p.38.

[8]Confer M. M. Mizrahi, J. Math. Phys. **19**, 298 (1978); Erratum, *ibid*. **21**, 1965 (1980).

[9]Confer C. Garrod, Rev. Mod. Phys. **38**, 483 (1966).

[10]E. Nelson, J. Math. Phys. **5**, 332 (1964).

[11]T. Ichinose, "Path integral for a hyperbolic system of the first order," Duke Math. J. **51** (to appear); also Proc. Jpn. Acad. Ser. A **58**, 290 (1982).

[12]T. Ichinose, "Path integral formulation of the propagator for a two-dimensional Dirac particle," in *Proceedings of the VIIth International Congress on Mathematical Physics* ($M\cap\Phi$), Boulder, Colorado, 1–10 Aug. 1983 (to appear).

[13]See, e.g, P. R. Chernoff, Pacific, J. Math. **72**, 361 (1977), Theorem 2.1.

[14]Confer Y. Nambu, Prog. Theor. Phys. **5**, 82 (1950).

[15]G. Rosen, *Formulations of Classical and Quantum Dynamical Theory* (Academic, New York, 1969), Appendix *E*, p.118.

[16]G. V. Riazanov, Zh. Eksp. Teor. Fiz. **33**, 1437 (1958) [Soviet Phys. JETP **6**, 1107 (1958)].

[17]Yu. L. Daletskii, Dokl. Akad. Nauk SSSR [Soviet Math. Dokl. **2**, 1634 (1961)]; Uspehi Mat. Nauk. **17 (5)**, 3 (1962) [Russian Math. Surveys **17**, 1 (1962)]. See also Yu. L. Daletskii and S. V. Fomin, Theory Probab. Its Appl. **10**, 304 (1965).

[18]See, e.g., C. Itzykson and J.-B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980), Chap. 2, 2–5.

[19]See also M. Reed and B. Simon, *Methods of Modern Mathematical Physics, II: Fourier Analysis, Self-Adjointness* (Academic, New York, 1975), Chap. X, Sec. 11, p. 276.

[20]See, e.g., E. Hewitt and K. A. Ross, *Abstract Harmonic Analysis, I* (Springer, Berlin, 1963), Chap. 3, Sec. 14 (14.10), p. 170.

[21]See I. G. Petrovsky, *Lectures on Partial Differential Equations* (Interscience, New York, 1964); R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Wiley, New York, 1962), Vol. II.

[22]See, e.g., M. Reed and B. Simon, *Methods of Modern Mathematical Physics, I: Functional Analysis*, revised and enlarged ed. (Academic, New York, 1980).

[23]See, e.g., G. Köthe, *Topological Vector Spaces* (Springer, Berlin, 1969, 1979), Vol. II, Chap. 8, Sec. 42, 2 (1), p.204.

[24]Note that in the space (3.11) the following classes of sets coincide: the bounded sets, the weakly relatively compact sets, and the relatively compact sets. See, e.g., Ref. 23, Vol. I, Chap. 4, Sec. 21, 5 (4), p. 263, Sec. 27, 2, p. 369; and L. Schwartz, *Thèorie des Distributions*, Tome II (Hermann, Paris, 1959), Chap, 7, Sec. 4. See also I. M. Gelfand and G. E. Shilov, *Generalized Functions*, in *Spaces of Fundamental and Generalized Functions* (Academic, New York, 1968), Vol. 2.

[25]K. Swong, Math. Ann. **155**, 270 (1964), Corollary to Theorem 6.2, p. 286. Strong countable additivity follows from the fact that weak and strong sequential convergence coincide in the space (3.11). See, e.g., Ref. 23, Vol. I.Sec. 27, 2 (7), p. 371.

[26]See, e.g., A Grothendieck, *Topological Vector Spaces* (Gordon and Breach, New York, 1973), Chap. 5, Part 4, 1, Exercise 6, p. 226.

[27]N. Dunford and J. T. Schwartz, *Linear Operators*, Part I: *General Theory* (Interscience, New York, 1958), Theorem VI. 7.3, p. 493.

[28]N. Bourbaki, *Intégration* (Hermann, Paris, 1965), 2nd ed., Chaps. 1–4, Chap. 3, Sec. 2, no. ', Corr. to Prop. 1; *ibid*. (Hermann, Paris, 1959), Chap. 6, Sec. 2, no. 1, Remarque 4.

[29]See, e.g., T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed. (Springer, Berlin, 1966, 1976), Chap. VIII, Sec. 1, Cor. 1.6 and Chap. IV, Sec. 3, Theorem 3.6 (Trotter–Kato), or Ref. 22, Theorem VIII.26 and Theorem VIII.21 (Trotter–Kato) with Chap. VIII, Problem 21.

[30]See, e.g., I. M. Gelfand, M. I. Graev, and N. Ya. Vilenkin, *Generalized Functions*, in *Integral Geometry and Representation Theory* (Academic, New York, 1966), Vol. 5, Chap. 1, Sec. 1.

[31]D. Ludwig, Commun. Pure Appl. Math. **19**, 49 (1966), and P. D. Lax and R. S. Phillips, Commun. Pure Appl. Math. **23**, 409 (1970); Errata, *ibid*. **24**, 279 (1971).

[32]See, e.g., W. D. Evans, Proc. London Math. Soc. **(3) 20**, 537 (1970), Theorem 2; J. Weidmann, Math. Z. **119**, 349 (1971), Sec. 9.

[33]See, e.g., L. D. Landau and E. M. Lifschitz, *Course of Theoretical in Physics*, in *The Classical Theory of Fields*, 4th revised English ed. (Pergamon, Oxford, 1975), Vol. 2.

On the path integral quantization of the damped harmonic oscillator

Christopher C. Gerry
Department of Physics, St. Bonaventure University, St. Bonaventure, New York 14778

(Received 28 October 1983; accepted for publication 13 January 1984)

A time-dependent canonical transformation is performed in the path integral expression of the propagator for a damped harmonic oscillator to reduce it to a harmonic oscillator with modified frequency. The transformation is carried out in a properly time-symmetrized expression of the lattice space path integral by making expansions about the midpoint of each time interval.

PACS numbers: 03.65.Fd

I. INTRODUCTION

Recently there have been a number of papers dealing with the quantization of nonconservative systems.[1-3] In particular, Gzyl[4] has recently considered the quantization of a damped harmonic oscillator by means of a canonical transformation which reduces the problem to standard harmonic oscillator with a modified frequency. In the quantum case this amounts to a unitary change of representation where the solution for the representation functions results in the quantization of the system.[5]

In this paper we wish to consider a path integral quantization for this system where we shall also employ a canonical transformation to reduce the problem to a harmonic oscillator. We note that the equation of motion of the damped oscillator (for unit mass) is given by

$$\ddot{x} + \lambda \dot{x} + \omega^2 x = 0 \tag{1.1}$$

which can be obtained from the Lagrangian

$$L = \tfrac{1}{2} e^{\lambda t}(\dot{x}^2 - \omega^2 x^2) \tag{1.2}$$

by the Euler–Lagrange equations or from the Hamiltonian

$$H = \tfrac{1}{2} e^{-\lambda t} p^2 + (\omega^2 x^2/2) e^{\lambda t} \tag{1.3}$$

via the Hamilton equations $\dot{x} = \partial H/\partial p, \dot{p} = -\partial H/\partial x$. If we now make the contact transformation $x = \exp(-\lambda t/2) y$, in Eq. (1.2) we have

$$L = \tfrac{1}{2}(\dot{y}^2 - \lambda \dot{y} y - \omega_\lambda^2 y^2), \tag{1.4}$$

where $\omega_\lambda^2 = \omega^2 - \lambda^2/4$. The term $-\lambda y \dot{y}/2$ does not contribute to the equations of motion (it is a total derivative in time) so it is easily seen that Eq. (1.4) is essentially the Lagrangian for a harmonic oscillator of modified frequency in the y coordinate.

Now previously, Khandekar and Lawande[6] have considered a path integral expression for the propagator, based on the Lagrangian of Eq. (1.2). They write

$$K(x'',t'';x't') = \int \exp\left(\frac{i}{\hbar}\int_{t'}^{t''} L\, dt\right) \mathscr{D}x(t), \tag{1.5}$$

which in the lattice space reads

$$K(x'',t'';x',t') = \lim_{N\to\infty} A_N \int \cdots \int \exp\left\{\frac{i}{\hbar} \times \sum_k S_k(x_k,x_{k-1},\epsilon)\right\} \prod_{k=1}^{N-1} dx_k, \tag{1.6}$$

where

$$S_k = \int_{t_{k-1}}^{t_k} L\, dt \simeq \epsilon L(x_k,x_{k-1},\epsilon) = e^{\lambda t_k}[(1/2\epsilon)(x_k - x_{k-1})^2 - (\epsilon/2)\omega^2 x_k^2], \tag{1.7}$$

and where

$$A_N = \prod_{k=1}^{N} \left(\frac{e^{\lambda t_k}}{2\pi i \hbar \epsilon}\right)^{1/2}. \tag{1.8}$$

The path integration of Eq. (1.6) can then be carried out straightforwardly owing to the quadratic form of the Lagrangian.

However, if one could make the transformation $x = \exp(-\lambda t/2) y$ in the path integral, one would presumably need only the propagator for the harmonic oscillator of frequency ω_λ (in the y representation) to obtain the propagator of the damped oscillator. However, a moment's reflection will show that in making a transformation such as $x_k = \exp(-\lambda t_k/2) y_k$ in Eq. (1.6) or Eq. (1.7), one encounters many ambiguities in regard to the kinetic energy term $\exp(\lambda t_k)(x_k - x_{k-1})^2$. For instance, the term $x_{k-1}^2 \exp(\lambda t_k)$ gives rise to $y_{k-1}^2 \exp[\lambda(t_k - t_{k-1})]$ and terms of order $\epsilon = t_k - t_{k-1}$ must be retained. It is also not obvious as to how the modified frequency ω_λ will appear from this transformation.

In the next section of this paper we illustrate how such a transformation may in fact be carried out after we first find an alternate expression for the Lagrangian path integral in which we symmetrize the time variable. It should be pointed out that the transformation we are using, in contrast to Gzyl's,[4] is only a time-dependent contact transformation whereas his also involves the momentum. Such transformations that mix coordinates and momenta have no place in path integration in the Schrödinger representation.

II. PATH INTEGRAL AND CONTACT TRANSFORMATIONS

Since the Hamiltonian of Eq. (1.3) is time dependent, we write the equation of the evolution of the state $|\psi(t)\rangle$ as

$$|\psi(t'')\rangle = U(t'',t')|\psi(t')\rangle, \tag{2.1}$$

where

$$U(t'',t') = T \exp \int_{t'}^{t''} \left(-\frac{i}{\hbar} H(t)\right) dt \tag{2.2}$$

is the time-ordered product. We write this as

 0022-2488/84/061820-03$02.50

$$U(t'',t') = \lim_{N\to\infty} \exp\left(-\frac{i}{\hbar}\epsilon H(\bar{t}_N)\right) \times \exp\left(-\frac{i}{\hbar}\epsilon H(\bar{t}_{N-1})\right) \times\cdots\times \exp\left(-\frac{i}{\hbar}\epsilon H(\bar{t}_2)\right) \times \exp\left(-\frac{i}{\hbar}\epsilon H(\bar{t}_1)\right), \tag{2.3}$$

where $\epsilon = (t'' - t')/N, t_k = t' + k\epsilon$, and $\bar{t}_k = (t_k + t_{k-1})/2$, for $k = 0,1,\ldots,N$. Here we have made the ansatz of employing a symmetrized time $\bar{t}_k$ in Eq. (2.3) rather than the customary t_k. The reason for this will become clear shortly. We now follow the standard procedure[7] and introduce the propagator as

$$K(x'',t'';x',t') = \langle x''|U(t'',t')|x'\rangle, \tag{2.4}$$

and with the completeness relations

$$\int_{-\infty}^{\infty} dx|x\rangle\langle x| = 1, \tag{2.5a}$$

$$\int_{-\infty}^{\infty} dp|p\rangle\langle p| = 1, \tag{2.5b}$$

inserted between the appropriate partitions of Eq. (2.3) we obtain the phase-space path integral

$$K(x'',t'';x't') = \int \mathscr{D}p(t)\mathscr{D}x(t)\exp\left\{\frac{i}{\hbar}\int_{t'}^{t''}[p\dot{x} - H]dt\right\} \tag{2.6}$$

$$= \lim_{N\to\infty}\int_{-\infty}^{\infty}\prod_{k=1}^{N}\frac{dp_k}{2\pi\hbar}\prod_{k=1}^{N-1}dx_k \times \prod_{k=1}^{N}\exp\left\{\frac{i}{\hbar}\left[p_k(x_k - x_{k-1}) - \epsilon\left(\frac{e^{-\lambda\bar{t}_k/2}}{2}p_k^2 + \frac{\omega^2x_k^2}{2}e^{\lambda t_k}\right)\right]\right\}, \tag{2.7}$$

where $x'' = x_N$ and $x' = x_0$. Note that the time-dependent factor of p_k^2 contains $\bar{t}_k$. This is appropriate because the paths in p-space are distributional,[8] i.e., taken as constant over the interval $(k,k-1)$ but discontinuous on the endpoints. (A better notation might be $p_{k-1/2}$, see Ref. 8.) We thus have used the average time over the interval. Whether one uses $\bar{t}_k$ or t_k in the factor of x_k^2 is immaterial as the difference is of order ϵ^2, which can be ignored in the limit $N\to\infty$.

We now proceed by evaluating the p-integrals of Eq. (2.7). As these are Gaussian one easily obtains

$$K(x'',t'';x't') = \lim_{N\to\infty} A'_N\int\cdots\int\prod_{k=1}^{N}[e^{\lambda\bar{t}_k}]^{1/2} \times \prod_{k=1}^{N-1}dx_k\prod_{k=1}^{N}\exp\left\{\frac{i}{\hbar}\left[\frac{e^{\lambda\bar{t}_k}}{2\epsilon}(x_k - x_{k-1})^2 - \frac{\epsilon}{2}\omega^2x_k^2e^{\lambda t_k}\right]\right\}, \tag{2.8}$$

where $A'_N = (2\pi i\hbar\epsilon)^{-N/2}$. Note that the continuum limit of Eqs. (1.6) and (2.8) is the same although they differ in their lattice space formulations.

We now perform a time-dependent contact transformation such that the explicit time dependence of Eq. (2.8) is removed. As we indicated, one is tempted to make a straightforward substitution of $x_k = \exp(-\lambda t_k/2)y_k$ into Eq. (2.8) or (1.6). As has been demonstrated elsewhere,[9,10] extreme caution must be used in implementing canonical transformations in path integrals. It is particularly important to retain up to order ϵ.

To this end we consider a general case of the form $x = f(t)y$, where $f(t) = \exp(-\lambda t/2)$ for our case of interest. In analogy to the midpoint method of contact canonical transformation discussed in Refs. 9 and 10, we here make expansions about the midpoint in the time interval, i.e., $\bar{t}_k$. That is, to order ϵ, we write

$$x_k = y_k[f(\bar{t}_k) + \tfrac{1}{2}\dot{f}(\bar{t}_k)\epsilon], \tag{2.9a}$$

$$x_{k-1} = y_{k-1}[f(\bar{t}_k) - \tfrac{1}{2}\dot{f}(\bar{t}_k)\epsilon], \tag{2.9b}$$

where the dot denotes time differentiation. We then have

$$\Delta x_k = x_k - x_{k-1} = f(\bar{t}_k)\Delta y_k + \dot{f}(\bar{t}_k)\bar{y}_k\epsilon, \tag{2.10}$$

where $\Delta y_k = y_k - y_{k-1}$ and $\bar{y}_k = (y_k + y_{k-1})/2$. Also we have

$$(\Delta x_k)^2 = [f(\bar{t}_k)]^2(\Delta y_k)^2 + 2\epsilon f(\bar{t}_k)\dot{f}(\bar{t}_k)\bar{y}_k\Delta y_k + [\dot{f}(\bar{t}_k)]^2\epsilon^2\bar{y}_k^2. \tag{2.11}$$

With $f(t) = \exp(-\lambda t/2)$ we obtain

$$(\Delta x_k)^2 = e^{\lambda\bar{t}_k}(\Delta y_k)^2 - \epsilon\lambda e^{-\lambda\bar{t}_k}\bar{y}_k\Delta y_k + (\lambda^2/4)\epsilon^2\bar{y}_k^2e^{\lambda\bar{t}_k}, \tag{2.12}$$

thus the short time action of Eq. (2.8)

$$S_k = (1/2\epsilon)e^{\lambda\bar{t}_k}(\Delta x_k)^2 - \epsilon(\omega^2x^2/2)e^{\lambda t_k} \tag{2.13}$$

becomes upon substitution of Eqs. (2.9) and (2.12)

$$S_k = (1/2\epsilon)(\Delta y_k)^2 - \tfrac{1}{2}\lambda\bar{y}_k\Delta y_k - (\epsilon/2)\omega_\lambda^2y_k^2, \tag{2.14}$$

where $\omega_\lambda^2 = \omega^2 - \lambda^2/4$. In the continuous limit one easily extracts the Lagrangian of Eq. (1.4).

We must also transform the "measure" part of the path integral. Using the symmetrizing procedure discussed in Refs. 9 and 10, we have for the general case, to order ϵ

$$\prod_{k=1}^{N-1}dx_k = [f(t'')f(t')]^{-1/2} \times \prod_{k=1}^{N}f(\bar{t}_k)\prod_{k=1}^{N-1}dy_k. \tag{2.15}$$

In our case this becomes

$$\prod_{k=1}^{N-1}dx_k = e^{\lambda(t''+t')/4} \times \prod_{k=1}^{N}[e^{-\lambda\bar{t}_k}]^{1/2}\prod_{k=1}^{N-1}dy_k. \tag{2.16}$$

Now from Eqs. (2.14) and (2.16) we obtain the transformed path integral as

$$K(x'',t'';x',t') = e^{\lambda(t''+t')/4}\tilde{K}(y'',t'';y',t'), \tag{2.17}$$

where

$$\tilde{K}(y'',t'',y',t') = \lim_{N\to\infty}A'_N\int\cdots\int\prod_{k=1}^{N-1}dy_k \times \prod_{k=1}^{N}\exp\left\{\frac{i}{2\hbar}\left[\frac{(\Delta y_k)^2}{\epsilon} - \lambda\bar{y}_k\Delta y_k - \epsilon\omega_\lambda^2y_k^2\right]\right\}. \tag{2.18}$$

We can remove all the explicit time dependence of the y-representation by using the integral equation

$$\psi(x'',t'') = \int K(x'',t'';x't')\psi(x',t')dx', \quad (2.19)$$

Eq. (2.18), and the change of variable $x = \exp(-\lambda t/2)y$ on the endpoints such that we may redefine the wave functions as $\tilde{\psi}(y,t) = e^{-\lambda t/4}\psi(e^{-\lambda t/2}y,t)$ to obtain

$$\tilde{\psi}(y'',t'') = \int \tilde{K}(y'',t'';y',t')\tilde{\psi}(y',t')dy'. \quad (2.20)$$

We now have only to evaluate the path integral in Eq. (2.18). This may be written symbolically as

$$\tilde{K}(y'',t'';y',t') = \int \mathscr{D}y(t)\exp\left\{\frac{i}{\hbar}\int_{t'}^{t''} L(y,\dot{y})dt\right\}, \quad (2.21)$$

where $L(y,\dot{y})$ is given by Eq. (1.4). As noted earlier, the action calculated with this Lagrangian contains a path independent term, i.e.,

$$\begin{aligned} S &= \int_{t'}^{t''} \frac{1}{2}(\dot{y}^2 - \lambda y\dot{y} - \omega_\lambda^2 y^2)dt \\ &= -\frac{\lambda}{4}(y'' - y''^2) + \int_{t'}^{t''} \frac{1}{2}(\dot{y}^2 - \omega_\lambda^2 y^2)dt. \end{aligned} \quad (2.22)$$

Thus we are left with a path integral for a harmonic oscillator of frequency ω_λ. The result is of course very well known[11] yielding

$$\begin{aligned} &\tilde{K}(y'',t'';y',t') \\ &= \left(\frac{\omega_\lambda}{2\pi i\hbar \sin\omega_\lambda(t''-t')}\right)^{1/2} \\ &\quad\times\exp\left\{\frac{i\omega_\lambda}{2\hbar\sin\omega_\lambda(t''-t')}\right. \\ &\quad\times[(y''^2 + y'^2)\cos\omega_\lambda(t''-t') \\ &\quad\left. - 2y'y''] - (i\lambda/4\hbar)(y''^2 - y'^2)\right\}. \end{aligned} \quad (2.23)$$

Thus we obtain from Eqs. (2.18) and (2.23) the final result

$$\begin{aligned} K(x'',t'';x't') = &\frac{e^{\lambda(t''+t')/4}}{[(2\pi i\hbar/\omega_\lambda)\sin\omega_\lambda(t''-t')]^{1/2}} \\ &\times\exp\left\{\frac{i\omega_\lambda}{2\hbar\sin\omega_\lambda(t''-t')}\right. \\ &\times[(e^{\lambda t''}x''^2 + e^{\lambda t'}x'^2)\cos\omega_\lambda(t''-t') \\ &- 2x''x'e^{\lambda(t''+t')/2}] \\ &\left. - (i\lambda/4)(e^{\lambda t''}x''^2 - e^{\lambda t'}x'^2)\right\}. \end{aligned}$$

III. CONCLUSIONS

This result may be compared with the calculations of Jannussis *et al.*[12] who use the solutions of the Schrödinger equation[13] to obtain the propagator by the definition

$$K(x''t'';x't') = \sum_n \psi_n^*(x't')\psi_n(x'',t'').$$

Their result differs from ours in the sign of the last term. We believe ours to be correct as the sign can be directly traced to the term $-\lambda y\dot{y}/2$ in Eq. (1.4).

Now in arriving at the above result we have started out with a phase-space path integral in which the time has been averaged over each interval of the time lattice. The Lagrangian path integral, Eq. (2.8) thus obtained differs from the usual form of Eqs. (1.6)–(1.8). Our expression has allowed us to follow a procedure analogous to the midpoint method of contact canonical transformations of Refs. 9 and 10 to transform the problem into the simpler problem of the harmonic oscillator. Apparently there is great ambiguity in the correct definition of the propagator when time explicitly appears. It is not clear how such a transformation could be carried out using Eqs. (1.7) and (1.8). Finally it must also be stated that there must be ambiguity in the definition of the phase-space path integral as it is not clear what form this should have for Eqs. (1.6)–(1.8), where the time parameter appears unsymmetrized over the time intervals. We thus conclude that our form Eq. (2.7) is the more appropriate.

[1]N. A. Lemos, Phys. Rev. D **24**, 2338 (1981).
[2]N. A. Lemos, Phys. Rev. D **24**, 1036 (1981).
[3]R. M. Santilli, Hadronic J. **5**, 264 (1982).
[4]H. Gzyl, Phys. Rev. A **27**, 2297 (1983).
[5]See M. Moshinsky and T. H. Seligman, Ann. Phys. **114**, 243 (1978) and references therein.
[6]D. C. Khandekar and S. V. Lawande, J. Math. Phys. **20**, 1870 (1979).
[7]See L. S. Schulman *Techniques and Applications of Path Integration* (Wiley, New York, 1981).
[8]C. Garrod, Rev. Mod. Phys. **38**, 483 (1966).
[9]J. L. Gervais and H. Jevicki, Nucl. Phys. B **110**, 53 (1976).
[10]C. C. Gerry, J. Math. Phys. **24**, 874 (1983).
[11]R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965), p. 63.
[12]A. D. Jannussis, G. N. Brodimas, and A. Streclas, Phys. Lett. A **74**, 6 (1979).
[13]V. V. Dodonov and V. I. Man'co, Nuovo Cimento B **44**, 265 (1978).

The Dirac equation and Hestenes' geometric algebra

J. Dwayne Hamilton
Selkirk College, Castlegar, British Columbia, Canada V1N 3J1

(Received 30 August 1983; accepted for publication 2 December 1983)

Hestenes' geometric algebra and Dirac spinors are reviewed and united into a common mathematical formalism, a unification that establishes the Dirac equation as being manifestly covariant under the Lorentz group, and one that needs no matrix representation of the Dirac algebra. New and simple methods of amplitude or "trace" calculations are then described. A number of problems are then considered within the context of the new approach, such as relativistic spin projections, new and covariant C and T-transformations and spinors for massless and Majorana fields.

PACS numbers: 03.65.Fd, 11.10.Qr, 11.30.Cp, 02.10. + w

1. INTRODUCTION

The Dirac equation, for over half a century the main theoretical tool among the few at the disposal of the physicist wanting to understand the enigmatic electron and other presumably fundamental spin-$\frac{1}{2}$ particles, stands virtually alone among the great equations of physics in that it is not generally regarded as being independent of the reference frame in which it is to be employed, that is to say, as being *manifestly* covariant under transformations of the Lorentz group. The reason for this is the apparent advantage to be gained by regarding the Dirac matrices (γ_μ) in the Dirac equation (units: $\hbar = c = 1$)

$$(i\gamma^\mu\partial_\mu - m - e\gamma^\mu A_\mu(x))\psi(x) = 0 \tag{1.1}$$

(with all of the symbols having their usual meaning) as fixed matrices,[1] with a few especially simple representations[2] of them being universally adopted and employed. These matrices then operate on the four-component column matrix $\psi(x)$, frequently referred to as a "Dirac spinor," scrambling its components which, in turn, represent spin-$\frac{1}{2}$ particles and antiparticles and their spin states. A considerable literature exists concerning these matrices,[2–4] their representations and transformations, their efforts on spinors,[5–7] and how they are used to calculate results of physical interest,[8,9] but only rarely (I have found but two examples[10,11] and none at all in the most recent texts) is it even mentioned that if the (γ_μ) transform as the basis 4-vectors of a reference frame (the viewpoint of this paper), abandoning fixed matrix representations, then does the Dirac equation have the appearance and structure of a manifestly covariant equation as far as the Lorentz group is concerned.

Not unrelated to the Dirac equation, but for other reasons as well, there is considerable current interest in the structure of the algebras and groups[12] which have proven to be so successful in their applications to modern particle physics. In particular, the unitary, orthogonal, and other groups as well as the Clifford algebras (among which are to be included the Pauli and Dirac algebras), which are not unrelated to them, have been for some time, and remain, active research topics. Can certain of these algebras be directed to the covariance and interpretation of the Dirac equation, and, indeed, eliminate *all* reference to matrices and their representations and provide a suitable geometric setting for the important Dirac spinors? I believe so.

The purpose of this paper is to show that the Dirac algebra described by Hestenes[13–17] and others,[18–21] a Clifford algebra based on the usual four-dimensional Minkowskian space-time, is the natural formalism in which to embed Dirac's four-component spinors. Hestenes' version of this algebra is particularly singled out because of its informality, its ease of application and manipulation, and its power. Transformations within the four-dimensional vector space of spinors are accomplished by multiplication by elements of the Dirac algebra $\mathscr{D}$, an algebra isomorphic to the well-known algebra of the Dirac matrices. The whole formalism will be seen to be covariant under transformations of the Lorentz group, and no need whatever will exist for any matrix representation of the algebra or the vector space of spinors; most of what follows will even be expressed in a component-free way. Section 2 will be needed to review the Dirac algebra and its subalgebra, the Pauli algebra, and their application to rotations and Lorentz transformations; Sec. 3 will informally define the vector space of Dirac spinors and its dual space of "adjoint" spinors, which will then be united, in Sec. 4, through an important and fundamental new "theorem" that unites the two, and which describes a new, simple and useful way of calculating the ubiquitous amplitudes and "traces" of relativistic quantum mechanics. Section 5, finally, will apply these techniques to a number of simple problems in relativistic quantum theory, namely: (a) calculating spin projection probabilities; (b) the description by the covariant Dirac equation of an electron in a magnetic field; (c) a new spinor basis needed for massless spin-$\frac{1}{2}$ particles; (d) expressing the charge conjugation transformation (C) and time-reversal transformation (T) in a new and covariant way that avoids the specific representations of the Dirac matrices hitherto employed [the parity (P) transformation is trivially reinterpreted but remains unchanged from what is conventional]; and (e) a look at how Majorana fields have to be expressed in the new notation, since, among other reasons, the charge conjugation transformation described in Sec. 5D forbids the existence of self-conjugate fields.

The literature recognizes two distinct (and, unfortunately, confusing) concepts for the appellation "spinor": (a) it is currently fashionable,[6,7,16,22] for example, to call the Lorentz transformation L in a change of reference frame such as

$$\gamma'_\mu = L\gamma_\mu L^{-1} \tag{1.2}$$

 0022-2488/84/061823-10$02.50

a spinor, or a rotation R in a change of three-dimensional basis

$$\sigma_i' = R\sigma_i R^{-1}, \tag{1.3}$$

such objects, in their respective algebras, being represented, if necessary, by 4×4 or 2×2 matrices, or even by quaternionlike objects; (b) the two-component "Pauli spinor" or the four-component "Dirac spinor," such objects being represented by *column* matrices, and traditionally called "spinors." It is in this latter sense only that the word "spinor" will be used in this paper, with the unfortunately vaguer terms "operator" or "multivector" being used for the first sense.

The mathematics that follows is quite informal and makes no claim to mathematical rigor; my aim is rather to develop and display a practical and simple, but powerful, spinor algebra, one with which facility can be readily acquired. The formal and technical aspects of the algebras used here have been widely developed and discussed in the literature; it is there that the interested reader will be referred in the likely event he finds the presentation here too informal or incomplete, although many references given here only sample this vast literature.

2. REVIEW OF THE DIRAC AND PAULI ALGEBRAS, ROTATIONS, AND LORENTZ TRANSFORMATIONS

The notation and formalism used in this paper would perhaps be best illustrated by a brief review of Hestenes' geometric algebra as it applies to Minkowskian space-time and transformations within it. The reader is referred elsewhere[13–17] for a complete review and explanation of the algebra and its wealth of identities and relationships.

If A and B are 4-vectors in Minkowskian space-time then a general "geometric product" is denoted by simple juxtaposition:

$$\begin{aligned} AB &= \tfrac{1}{2}(AB+BA) + \tfrac{1}{2}(AB-BA) \\ &\equiv A\cdot B + A\wedge B \\ &= B\cdot A - B\wedge A, \end{aligned} \tag{2.1}$$

where $A\cdot B$ is the usual scalar product and the *bivector* $A\wedge B$ can be regarded as the antisymmetric tensor product. This "exterior" or "outer" product, denoted by "$\wedge$," essentially the same one used with differential forms,[23] is defined to be associative, a property which, with its antisymmetric nature, implies that

$$A\wedge B\wedge C\wedge D\wedge E = 0 \tag{2.2}$$

for any five or more 4-vectors. The 16-dimensional vector space (or 32-dimensional, if multiplication by complex numbers is admitted) $\mathscr{C}_4 \equiv \mathscr{D}$, the Clifford algebra based on a four-dimensional vector space, has as an arbitrary element the *multivector*

$$\begin{aligned} M = {}& \text{scalar}(S) + \text{vector}(V) + \text{bivector}(B) + \text{trivector}(T) \text{ or} \\ & \text{pseudovector} + \text{pseudoscalar } (P). \end{aligned} \tag{2.3}$$

If (γ_μ) $(\mu = 0,1,2,3)$ is an orthonormal set of basis 4-vectors (which constitutes a reference frame, with γ_0 the reference frame's 4-velocity), or *tetrad*,

$$\gamma_\mu\cdot\gamma_\nu = \tfrac{1}{2}(\gamma_\mu\gamma_\nu + \gamma_\nu\gamma_\mu) \equiv \eta_{\mu\nu} \tag{2.4}$$

(where $\eta_{00}\equiv 1$, $\eta_{ij}\equiv -\delta_{ij}$), then M has the form

$$\begin{aligned} M = {}& \alpha + \alpha^\mu\gamma_\mu + \tfrac{1}{2}\alpha^{\mu\nu}\gamma_\mu\wedge\gamma_\nu \\ & + (1/3!)\beta^{\lambda\mu\nu}\gamma_\lambda\wedge\gamma_\mu\wedge\gamma_\nu \\ & (\text{or } \beta^\mu\gamma_5\gamma_\mu) + \beta\gamma_5, \end{aligned} \tag{2.5}$$

where $\alpha, \alpha^\mu, \alpha^{\mu\nu} (= -\alpha^{\nu\mu}), \beta^{\lambda\mu\nu} (= -\beta^{\mu\lambda\nu}, \text{etc.}) (\text{or } \beta^\mu), \beta$ is a set of 16 (possibly complex) scalars, and where[24]

$$\gamma_5 \equiv \gamma_0\wedge\gamma_1\wedge\gamma_2\wedge\gamma_3 = \gamma_0\gamma_1\gamma_2\gamma_3 \tag{2.6}$$

is the unit pseudoscalar, which satisfies $(\gamma_5)^2 = -1$.

The subalgebra consisting of even multivectors of the Dirac algebra $\mathscr{D}$ is the algebra $\mathscr{C}_3 \equiv \mathscr{P}$,[13,19] the Pauli algebra, because a basis *triad* (σ_i) $(i = 1,2,3)$ defined by

$$\sigma_i \equiv \gamma_i\wedge\gamma_0 = \gamma_i\gamma_0 \tag{2.7}$$

satisfies

$$\sigma_i\cdot\sigma_j = \tfrac{1}{2}(\sigma_i\sigma_j + \sigma_j\sigma_i) = \delta_{ij}, \tag{2.8}$$

the requirement of a basis in ordinary three-dimensional space. Elements of $\mathscr{P}$ have the form[25]

$$m = \alpha + a_i\sigma_i + \tfrac{1}{2}B_{ij}\sigma_i\wedge\sigma_j + i\beta \tag{2.9a}$$

$$\begin{aligned} &= \alpha + \mathbf{a} + B\ (\text{or } i\mathbf{b}) + i\beta \\ &= (\alpha + i\beta) + (\mathbf{a} + i\mathbf{b}), \end{aligned} \tag{2.9b}$$

where α and β are real scalars and $\mathbf{a}$ and $\mathbf{b}$ are (real) 3-vectors, and where the unit pseudoscalar i $(i^2 = -1)$ is defined by

$$\begin{aligned} i &\equiv \sigma_1\wedge\sigma_2\wedge\sigma_3 = \sigma_1\sigma_2\sigma_3 \\ &= \gamma_0\gamma_1\gamma_2\gamma_3 = \gamma_5, \end{aligned} \tag{2.10}$$

a most important, and perhaps surprising, result, one with geometric significance, as Hestenes[16] has emphasized. I use the notation "i" in the Pauli algebra because $i\sigma_i = \sigma_i\, i$, so that in practice, in $\mathscr{P}$, it behaves no differently than the unit imaginary; thus $\mathscr{P}$ appears to consist of complex scalars and complex 3-vectors, as (2.9b) illustrates. But, since $\gamma_5\gamma_\mu = -\gamma_\mu\gamma_5$, its behavior in $\mathscr{D}$ is quite different. To minimize the conflict with traditional notation, the symbol "i" may be considered to be $\sqrt{-1}$ wherever it appears; however, "γ_5" must be kept $\mathscr{D}$ and will be replaced by "i" whenever 4-vector equations are reduced to 3-vector equations, which will be so indicated whenever it is being done.

The algebras $\mathscr{P}$ and $\mathscr{D}$ provide for quite an elegant solution to the rotation problem.[13,16] If the triad (σ_i) is rotated in the right-hand sense by an angle θ about the direction $\hat{n}$ (where the caret denotes a unit 3-vector other than the basis vectors) to become the new triad (σ_i'), then

$$\sigma_i' = R\sigma_i R^{-1}, \tag{2.11}$$

where

$$\begin{aligned} R &= \exp(-\tfrac{1}{2}i\hat{n}\theta) \\ &= \cos(\tfrac{1}{2}\theta) - i\hat{n}\sin(\tfrac{1}{2}\theta). \end{aligned} \tag{2.12}$$

There is also a Lorentz covariant way of expressing this spatial rotation, a way that, to my knowledge, has not appeared before, although it is a trivial generalization of (2.11) and (2.12). If the spatial 4-vectors of the tetrad (γ_μ) are rotated as above into the spatial 4-vectors of the tetrad (γ_μ'), then $\gamma_0' = \gamma_0$ and

$$\gamma_i' = R\gamma_i R^{-1}, \tag{2.13}$$

where

$$R = \exp(-\tfrac{1}{2}\gamma_5 N\gamma_0\theta) = \cos(\tfrac{1}{2}\theta) - \gamma_5 N\gamma_0 \sin(\tfrac{1}{2}\theta), \tag{2.14}$$

where N is the unit spacelike 4-vector ($N^2 = -1$) that satisfies $N\cdot\gamma_0 = 0$ and

$$\hat{n} = N\wedge\gamma_0 = N\gamma_0. \tag{2.15}$$

Since γ_0 anticommutes with each of γ_5 and N, we have $R\gamma_0 = \gamma_0 R$ and so (2.13) could be written

$$\gamma'_\mu = R\gamma_\mu R^{-1}. \tag{2.16}$$

Note that (2.11) and (2.12) follow from (2.14)–(2.16) by the replacement $\gamma_5 \to i$ after the equation has been written in terms of 3-vectors. This illustrates the remarks made in the previous paragraph.

Following Hestenes,[13] we can define "conjugate" multivectors of $\mathscr{D}$ such as

$$\widetilde{M} \equiv M, \quad \text{with all 4-vector products reversed,} \tag{2.17}$$

$$= S + V - B - T + P,$$

using the notation (2.3), and a kind of Hermitian conjugate[13,26]

$$M^\dagger \equiv V\widetilde{M}V, \tag{2.18}$$

where V is a unit timelike 4-vector ($V^2 = 1$) [$V = \gamma_0$, for some reference frame (γ_μ)]. Note that this Hermitian conjugate is not a Lorentz-covariant object, a fact that considerably reduces its utility, except that $\gamma_5^\dagger = -\gamma_5$ (so $i^\dagger = -i$) in all frames of reference. The 3-vectors (σ_i) defined by (2.7) satisfy $\sigma_i^\dagger = \sigma_i$ in the reference frame (γ_μ) where $V \equiv \gamma_0$, and (2.9b) becomes

$$m^\dagger = (\alpha - i\beta) + (\mathbf{a} - i\mathbf{b}) \tag{2.19}$$

$$= m, \quad \text{with all 3-vector products reversed,}$$

so this operation of Hermitian conjugation is compatible with that generally employed, but is here endowed with geometric significance. Note, now, that the rotation multivector R, in either the form (2.12) or (2.14), satisfies $R^\dagger = R^{-1}$: R is *unitary* [but only, I emphasize, in the reference frame (γ_μ) where γ_0 has been used for V in (2.18)]. If a complex scalar (α) multiplies a multivector M of $\mathscr{D}$, I will define

$$(\alpha M)^\dagger \equiv \alpha^* M^\dagger, \tag{2.20}$$

where the asterisk denotes the complex conjugate and where I will always take the basis vectors to be *real*: $\gamma_\mu^* \equiv \gamma_\mu$ (whether or not a real matrix representation might exist—matrices are irrelevant here).

A proper Lorentz transformation from the tetrad (γ_μ) to the tetrad (γ'_μ) is expressed by

$$\gamma'_\mu = L\gamma_\mu L^{-1}, \tag{2.21}$$

with[13,26] $L^{-1} = \widetilde{L}$ [note that, from (2.14), $\widetilde{R} = R^{-1}$, because $\gamma_5 N\gamma_0$ is a bivector]. A (proper) infinitesimal Lorentz transformation

$$\gamma'_\mu = \gamma_\mu + \epsilon_\mu{}^\nu\gamma_\nu \tag{2.22a}$$

is solved by

$$L = 1 - \tfrac{1}{4}\epsilon^{\mu\nu}\gamma_\mu\wedge\gamma_\nu, \tag{2.22b}$$

which satisfies $L^{-1} = \widetilde{L}$. Since γ_5 commutes with this L, it is an invariant under *proper* Lorentz transformations. A pure boost (that is, a Lorentz transformation with no rotation of the spatial axes) from (γ_μ) to (γ'_μ) is expressed by[26,27]

$$L \equiv H = (1 + \gamma'_0\gamma_0)[2(1 + \gamma'_0\cdot\gamma_0)]^{-1/2}, \tag{2.23a}$$

with

$$H^{-1} = \widetilde{H} = (1 + \gamma_0\gamma'_0)[2(1 + \gamma'_0\cdot\gamma_0)]^{-1/2}. \tag{2.23b}$$

Since $H^\dagger = V\widetilde{H}V = H$ for $V = \gamma_0$ or γ'_0 we say that such Lorentz transformations are Hermitian, but, again, this is not a generally covariant concept. The Lorentz transformation (2.23a) can be expressed in 3-vector or exponential notation,[13,28] but not covariantly, for which reason it will be avoided.

If (γ_μ) is an inertial reference frame, with proper time τ, then

$$\dot{\gamma}_\mu \equiv \frac{d\gamma_\mu}{d\tau} = 0, \tag{2.24}$$

and the triad $(\sigma_i) = (\gamma_i\gamma_0)$ is not rotating,

$$\dot{\sigma}_i = 0, \tag{2.25}$$

but if the triad $(\sigma'_i) = (\gamma'_i\gamma'_0) = (\gamma'_i\gamma_0)$ is rotating, then from (2.11) follows[24]

$$\dot{\sigma}'_i = \dot{R}R^{-1}\sigma'_i - \sigma'_i\dot{R}R^{-1} = \boldsymbol{\omega}\times\sigma'_i, \tag{2.26}$$

where the angular velocity $\boldsymbol{\omega}$ is obtained from the definition

$$\dot{R}R^{-1} \equiv -\tfrac{1}{2}i\boldsymbol{\omega}, \tag{2.27}$$

since $\dot{R}R^{-1}$ is a bivector (in both $\mathscr{P}$ and $\mathscr{D}$). In covariant 4-vector notation we have

$$\dot{\gamma}'_i = \dot{R}R^{-1}\gamma'_i - \gamma'_i\dot{R}R^{-1} = \gamma_5\gamma_0(\Omega\wedge\gamma'_i), \tag{2.28}$$

where

$$\dot{R}R^{-1} \equiv -\tfrac{1}{2}\gamma_5\Omega\gamma_0, \tag{2.29}$$

and Ω is the spacelike ($\Omega\cdot\gamma_0 = 0$) angular velocity 4-vector:

$$\boldsymbol{\omega} = \Omega\wedge\gamma_0 = \Omega\gamma_0, \tag{2.30}$$

with $\Omega^2 = -|\boldsymbol{\omega}|^2$. The reader should again note that the transition from covariant 4-vector notation (2.29) to 3-vector notation within a reference frame requires $\gamma_5 \to i$, which can then be regarded as the unit imaginary.

3. DIRAC SPINORS

Dirac spinors are elements of a vector space $\mathscr{S}$, and transform among themselves, we assume, when multiplied on the left by multivectors of the Dirac algebra $\mathscr{D}$. That is, if, $\psi\in\mathscr{S}$, then $M\psi\in\mathscr{S}$, for all $M\in\mathscr{D}$. The dual spinor space is denoted $\bar{\mathscr{S}}$: if $\bar{\psi}\in\bar{\mathscr{S}}$, then multiplication is on the right and $\bar{\psi}M\in\bar{\mathscr{S}}$. In the conventional terminology $\bar{\psi}$ is called the *adjoint* to the spinor ψ. Products of the form $\bar{\psi}\phi$ will be defined to be (complex) scalars, and we will see that products of the form $\phi\bar{\psi}$ are elements of $\mathscr{D}$. All products displayed here are defined to be associative, provided the result is meaningful. For example, $\bar{\psi}\phi\chi$ becomes $(\bar{\psi}\phi)\chi$, which is meaningful, but $\bar{\psi}(\phi\chi)$ is not.

The spinor equation $\phi = M\psi$ will be defined to have as its analog in $\bar{\mathscr{S}}$, $\bar{\phi} = \bar{\psi}\widetilde{M}$, where $\widetilde{M}$ is defined in (2.17) and is assumed here to have real components. However, if α is a complex scalar

$$\phi = \alpha M\psi \leftrightarrow \bar{\phi} = \alpha^* \bar{\psi}\widetilde{M}. \tag{3.1}$$

(For this reason dual spinors should have been denoted $\tilde{\psi}$[29]; however, the notation $\bar{\psi}$ is virtually universal.)

The Dirac algebra $\mathscr{D}$ permits only two different sets of commuting projection operators. If V is a timelike 4-vector ($V^2 = 1$), then an obvious first choice is the set

$$\Lambda_\pm \equiv \tfrac{1}{2}(1 \pm V), \tag{3.2}$$

since $\Lambda_\pm \Lambda_\pm = \Lambda_\pm, \Lambda_\pm \Lambda_\mp = 0, \Lambda_+ + \Lambda_- = 1$ are satisfied. If, temporarily, we use γ_0 as a specific V, a second choice could be the set $\frac{1}{2}(1 \pm i\gamma_1\gamma_2)$, since $(i\gamma_1\gamma_2)^2 = 1$ and $(i\gamma_1\gamma_2)$ commutes with γ_0. No other independent projection operators exist that commute with these two sets. Now, since $i\gamma_1\gamma_2 = -i\gamma_5\gamma_3\gamma_0$, the second set of projection operators will be taken to be

$$\Sigma_\pm = \tfrac{1}{2}[1 \pm (-i)\gamma_5\gamma_3\gamma_0], \tag{3.3}$$

where the signs will be justified later by the physical interpretation.[8] To generalize, the two sets of projection operators (3.2) and

$$\Sigma_\pm \equiv \tfrac{1}{2}[1 \pm (-i)\gamma_5 SV], \tag{3.4}$$

where $S^2 = -1, S\cdot V = 0$, imply that the vector space of spinors is four-dimensional. Thus, in general, Dirac spinors are four-component objects.

The projection operators $\Sigma_\pm$ are the spin projection operators because in the transition to the Pauli algebra $\gamma_5 \to i$ and since $SV = S \wedge V = \hat{s}$ [in the frame of reference (γ_μ) with 4-velocity $V = \gamma_0$], we have

$$\Sigma_\pm = \tfrac{1}{2}(1 \pm \hat{s}), \tag{3.5}$$

the usual and only projection operators in the Pauli algebra.[30] Note that although the V in (3.4) is not conventional[31] (it appears here, I believe, for the first time) it is required for the proper physical interpretation as (3.4) cannot become (3.5) without it—note its necessary appearance in (2.14).

What follows now is standard fare in the textbooks[1,2,8,24]: If $\phi_\pm$ and $\chi_\pm$ constitute a basis of $\mathscr{S}$, then

$$\Lambda_+ \phi_\pm \equiv \phi_\pm, \tag{3.6a}$$

$$\Lambda_- \chi_\pm \equiv \chi_\pm, \tag{3.6b}$$

$$\Sigma_\pm \phi_\pm \equiv \phi_\pm, \tag{3.6c}$$

$$\Sigma_\pm \chi_\mp \equiv \chi_\mp, \tag{3.6d}$$

where, again, the choice of signs will be justified by the physical interpretation. The dual, or adjoint, basis spinors are obtained from (2.17) and (3.1):

$$\bar{\phi}_\pm \Lambda_+ = \bar{\phi}_\pm, \tag{3.7a}$$

$$\bar{\chi}_\pm \Lambda_- = \bar{\chi}_\pm, \tag{3.7b}$$

$$\bar{\phi}_\pm \Sigma_\pm = \bar{\phi}_\pm, \tag{3.7c}$$

$$\bar{\chi}_\mp \Sigma_\pm = \bar{\chi}_\mp, \tag{3.7d}$$

because, for example,

$$(\widetilde{\Sigma}_+)^* = \tfrac{1}{2}[1 + (-i)^*(-\gamma_5 SV)] = \Sigma_+. \tag{3.8}$$

If we now define $\bar{\phi}_+ \phi_+ \equiv 1$, we obtain

$$\bar{\phi}_r \phi_s = \delta_{rs}, \tag{3.9a}$$

$$\bar{\chi}_r \chi_s = -\delta_{rs}, \tag{3.9b}$$

$$\bar{\phi}_r \chi_s = 0 = \bar{\chi}_r \phi_s, \tag{3.9c}$$

where $r,s = \pm$. It now follows that[8]

$$\phi_\pm \bar{\phi}_\pm = \Lambda_+ \Sigma_\pm, \tag{3.10a}$$

$$\chi_\pm \bar{\chi}_\pm = -\Lambda_- \Sigma_\mp, \tag{3.10b}$$

so that (summing over repeated subscripts)

$$\phi_r \bar{\phi}_r - \chi_r \bar{\chi}_r = 1, \tag{3.11a}$$

$$\phi_r \bar{\phi}_r + \chi_r \bar{\chi}_r = V, \tag{3.11b}$$

in the sense that operations on an arbitrarily selected spinor $\psi = \alpha_r \phi_r + \beta_r \chi_r$ (α_r, β_r complex numbers) are independent of which side of (3.10) or (3.11) is used. These of course, are the "completeness" or "closure" relations. The reader would have no trouble constructing a number of other identities from (3.10), such as

$$\phi_+ \bar{\phi}_+ - \chi_- \bar{\chi}_- = \Sigma_+, \tag{3.12}$$

or

$$\phi_+ \bar{\phi}_+ - \phi_- \bar{\phi}_- + \chi_+ \bar{\chi}_+ - \chi_- \bar{\chi}_- = -i\gamma_5 SV. \tag{3.13}$$

Indeed, the whole of the Dirac algebra can be constructed from elements of the form $\phi\bar{\chi}$; in the next section I will show how this can be accomplished in general.

Rotations and Lorentz transformations among spinors are accomplished using the multivectors R, L, and H of Sec. 2. If the reference frame (γ_μ) is rotated or boosted or, in general, a combination of the two, into the reference frame (γ'_μ), then $\gamma'_\mu = L\gamma_\mu L^{-1}$ with $\widetilde{L} = L^{-1}$, to repeat (2.21). It follows that the basis spinors θ_r ($= \phi_r$ or χ_r) associated with (γ_μ) transform into

$$\theta'_r \equiv L\theta_r, \tag{3.14a}$$

$$\bar{\theta}'_r = \bar{\theta}_r \widetilde{L} = \bar{\theta}_r L^{-1}, \tag{3.14b}$$

which therefore satisfy the same equations with respect to the reference frame (γ'_μ) as the basis θ_r does with respect to (γ_μ). Thus a general spinor

$$\psi = \alpha_r \phi_r + \beta_r \chi_r = \alpha'_r \phi'_r + \beta'_r \chi'_r \tag{3.15}$$

is manifestly a Lorentz covariant object provided its components transform as, for example,

$$\alpha'_r = \bar{\phi}'_r(\alpha_s \phi_s + \beta_s \chi_s) = \alpha_s \bar{\phi}_r L^{-1}\phi_s + \beta_s \bar{\phi}_r L^{-1}\chi_s. \tag{3.16}$$

The next section will consider how such expressions are to be explicitly evaluated.

A spatial rotation of ϕ_+, for example, gives, using (2.14),

$$\phi'_+ \equiv R\phi_+ = \alpha\phi_+ + \beta\phi_-, \tag{3.17}$$

because if $V \equiv \gamma_0$, $\gamma_5 N$ commutes with γ_0. Thus the basis sets ϕ_r, χ_r each remain in their own $\Lambda_\pm$ subspaces under spatial rotations, and can, with this restriction, be considered to be two-component Pauli spinors.[30]

The Lorentz boost (2.23a), however, completely scrambles the components:

$$\phi'_+ \equiv H\phi_+ = \alpha_r \phi_r + \beta_r \chi_r, \tag{3.18}$$

with none of α_r, β_r vanishing in general. I postpone further discussion of these transformations to Sec. 5A, where their physical interpretation will be considered.

It is now easy to see that if $\psi' = L\psi$, $\bar{\psi}' = \bar{\psi}L^{-1}$, then

$$\bar{\psi}'\psi' = \bar{\psi}\psi, \tag{3.19}$$

a Lorentz scalar, and

$$\psi'\bar{\psi}' = L\psi\bar{\psi}L^{-1}, \tag{3.20}$$

as elements of $\mathscr{D}$, such as vectors, must transform under a Lorentz transformation. Further, if $L: (\phi_r, \chi_r) \to (\phi'_r, \chi'_r)$, we may write

$$L = \phi'_r\bar{\phi}_r - \chi'_r\bar{\chi}_r, \tag{3.21a}$$

with

$$L^{-1} = \tilde{L} = \phi_r\bar{\phi}'_r - \chi_r\bar{\chi}'_r. \tag{3.21b}$$

4. THE "FUNDAMENTAL THEOREM" OF SPINOR ALGEBRA

I would now like to introduce a fundamental relation that considerably simplifies practical calculations using the spinor algebra, one that eliminates whatever apparent need may remain to refer to specific matrix representations of the Dirac algebra. It is easy to state: Let ψ be any Dirac spinor, $\bar{\phi}$ any adjoint spinor, and M any multivector of the Dirac algebra; then[32]

$$\bar{\phi}M\psi = 4S[M\psi\bar{\phi}], \tag{4.1}$$

where $S[\cdots]$ refers to the scalar part, in the sense of (2.3), of the multivector enclosed in the square brackets. In (4.1), $4S[\cdots]$ replaces[32,33] Tr $[\cdots]$ in what I consider to be obsolete matrix representations of the Dirac algebra, the factor 4 being required because the trace of the 4×4 unit matrix is 4.

From (4.1) easily follows what some call the "Fierz decomposition"[34,35]:

$$\psi\bar{\phi} = \tfrac{1}{4}(\bar{\phi}\psi) + \tfrac{1}{4}(\bar{\phi}\gamma^\mu\psi)\gamma_\mu - \tfrac{1}{8}(\bar{\phi}\gamma^\mu\wedge\gamma^\nu\psi)\gamma_\mu\wedge\gamma_\nu$$
$$+ \tfrac{1}{4}(\bar{\phi}\gamma_5\gamma^\mu\psi)\gamma_5\gamma_\mu - \tfrac{1}{4}(\bar{\phi}\gamma_5\psi)\gamma_5, \tag{4.2}$$

for any spinors ψ and ϕ. For example, the second term in (4.2) is obtained from (4.1) and

$$\begin{aligned}\bar{\phi}\gamma_\mu\psi &= 4S[\gamma_\mu\psi\bar{\phi}]\\ &= \tfrac{1}{4}(\bar{\phi}\gamma^\nu\psi)4S[\gamma_\mu\gamma_\nu]\\ &= (\bar{\phi}\gamma^\nu\psi)\eta_{\mu\nu},\end{aligned} \tag{4.3}$$

from (2.1) and (2.4). The important result (4.2) makes particularly clear the interesting connection between the four-component Dirac spinors on the one hand and the elements (2.3) and (2.5) of the Dirac algebra on the other, and clearly shows the tensor properties of the components of the multivectors in the brackets in (4.2).

I claim, without, I think, any exaggeration, that any "trace" calculation in relativistic quantum mechanics can be more easily and quickly accomplished by the use of (4.1) than by following the rules of trace evaluations and manipulations. Even the permutation rule is valid: If A, B, C, $\cdots$ are multivectors, then

$$\begin{aligned}S[AB\cdots CD] &= S[DAB\cdots C]\\ &= S[B\cdots CDA].\end{aligned} \tag{4.4}$$

For example, to evaluate $\bar{\phi}_+\gamma^\mu\frac{1}{2}(1-i\gamma_5)\phi_+$ one would use (3.10a) and (4.1) to obtain

$$\begin{aligned}&\bar{\phi}_+\gamma^\mu\tfrac{1}{2}(1-i\gamma_5)\phi_+\\ &= 4S[\gamma^\mu\tfrac{1}{2}(1-i\gamma_5)\tfrac{1}{2}(1+V)\tfrac{1}{2}(1-i\gamma_5SV)]\\ &= \tfrac{1}{2}S[\underset{(a)}{(\gamma^\mu} \underset{(b)}{- i\gamma^\mu\gamma_5)}(\underset{(c)}{1} \underset{(d)}{+ V} \underset{(e)}{- i\gamma_5SV} \underset{(f)}{- i\gamma_5S})]\\ &= \tfrac{1}{2}(V+S)\cdot\gamma^\mu,\end{aligned} \tag{4.5}$$

because one can see at a glance that only the products $a\cdot d$ and $b\cdot f$ have a scalar part; the rest do not. However, to evaluate an expression such as $\bar{\phi}_+\gamma^\mu\frac{1}{2}(1-i\gamma_5)\chi_-$, which is

$$\bar{\phi}_+\gamma^\mu\tfrac{1}{2}(1-i\gamma_5)\chi_- = 4S[\tfrac{1}{2}\gamma^\mu(1-i\gamma_5)\chi_-\bar{\phi}_+], \tag{4.6}$$

one would need the vector and pseudovector parts of (4.2), and $\bar{\phi}_+\gamma^\mu\chi_-$, $\bar{\phi}_+\gamma_5\gamma^\mu\chi_-$. Before these can be evaluated, one would have to tabulate the effects of the vectors (γ_μ) on the basis spinors. If $V\equiv\gamma_0$, $S\equiv\gamma_3$ in (3.6), one sees that, for example, $\gamma_1\phi_+ \propto \chi_+$, so with a few choices of phase this can readily be done. Thus all expressions of the form $\bar{\phi}M\psi$, such as those in the Lorentz transformation of (3.16), can be explicitly calculated.

5. APPLICATIONS TO THE PHYSICAL INTERPRETATION

The definitions and results of the previous sections, because of their simplicity and clarity, facilitate the physical interpretation of the Dirac equation and its solutions, particularly in regard to the components of the Dirac spinor $\psi(x)$ and transformations among solutions to the Dirac equation. In this section I will consider a number of simple examples.

A. Spin projection probabilities

If one takes the basis spinors for an electron (or positron) of mass m and 4-momentum $p = p^\mu\gamma_\mu = mV$ to be $u_r(p)$, $v_r(p)$ $(r = \pm)$, defined, as usual, by (3.6), as

$$\tfrac{1}{2}(1+V)\tfrac{1}{2}[1\pm(-i)\gamma_5SV]u_\pm \equiv u_\pm, \tag{5.1a}$$

$$\tfrac{1}{2}(1-V)\tfrac{1}{2}[1\pm(-i)\gamma_5SV]v_\mp \equiv v_\mp, \tag{5.1b}$$

where $SV = S\wedge V = \hat{s}$ is the spin 3-vector in the electron's rest frame, then the usual expansion in plane waves that satisfies the free-particle Dirac equation

$$(i\partial - m)\psi(x) = 0, \tag{5.2}$$

where $\partial \equiv \gamma^\mu\partial_\mu$, is[36]

$$\psi(x) = (2\pi)^{-3/2}\int d^3p\left(\frac{m}{\epsilon}\right)^{1/2} \times(e^{-ip\cdot x}a_ru_r + e^{ip\cdot x}b_r^*v_r), \tag{5.3}$$

where a_r^*, b_r^*, a_r, b_r are the usual Fock-space creation and annihilation operators and $\epsilon \equiv (m^2 + |\mathbf{p}|^2)^{1/2}$. The field 4-momentum P and angular momentum J have the components

$$P^\lambda = \int d^3x\, T^{0\lambda} = \int d^3x\, i\bar{\psi}\gamma^0\partial^\lambda\psi, \tag{5.4a}$$

$$J^{\lambda\mu} = \int d^3x\,[(x^\lambda T^{0\mu} - x^\mu T^{0\lambda}) + \tfrac{1}{2}i\bar{\psi}\gamma^0(\gamma^\lambda\wedge\gamma^\mu)\psi], \tag{5.4b}$$

where T is the canonical energy–momentum tensor and where the bivector $(\gamma^\lambda\wedge\gamma^\mu)$ appears in the "spin" part of the angular momentum because of the form for the infinitesimal Lorentz transformation (2.22b). The expansion (5.3) leads to the following 4-momentum and angular momentum in the direction of the electrons' motion ($\mathbf{J}\cdot\hat{P}$):

$$P = \int d^3p\, p(a_r^* a_r + b_r^* b_r), \tag{5.5a}$$

$$\mathbf{J}\cdot\hat{P} = \frac{1}{2}\int d^3p\,(a_+^* a_+ - a_-^* a_- + b_+^* b_+ - b_-^* b_-), \tag{5.5b}$$

where normal ordering has been assumed, results that justify the choices of sign made in (5.1) and (3.6). Here I have used

$$S = [\,|\mathbf{p}|/m,(\epsilon/m)\hat{p}\,] \tag{5.6}$$

with the 4-momentum $p = (\epsilon,|\mathbf{p}|\hat{p})$; that is, the spin was resolved into components in the direction of the electron's motion. These results are quite familiar,[8] but the spin assignments of the signs on the spinors $v_\pm$ apply only to the field-quantized theory.

In relativistic quantum mechanics (as opposed to the field-quantized version) the solution to the Dirac equation for a free, spin-up electron (or positron—they cannot, of course, be distinguished in the absence of an electromagnetic field) at rest in the reference frame (γ'_μ) is

$$\psi(x) = \psi(t') = e^{-imt'}\phi'_+, \tag{5.7}$$

where

$$\tfrac{1}{2}(1+\gamma'_0)\tfrac{1}{2}[1+(-i)\gamma_5 S\gamma'_0]\phi'_+ \equiv \phi'_+, \tag{5.8}$$

with $\hat{s}' = S\wedge\gamma'_0$ representing the spin direction in the electron's rest frame. (An equally valid solution,[37] to which I will return later, is

$$\psi(t') = e^{+imt'}\chi'_-, \tag{5.9}$$

where

$$\tfrac{1}{2}(1-\gamma'_0)\tfrac{1}{2}[1+(-i)\gamma_5 S\gamma'_0]\,\chi'_- \equiv \chi'_-. \tag{5.10}$$

A linear combination of (5.7) and (5.9), each parity eigenstates in the rest frame, is not a parity eigenstate (Sec. 5 D) and is therefore not acceptable. Here χ_- is associated with spin-up, the opposite of (5.1b), because the anticommutation relations, not present here, cause these roles to become reversed, a manifestation of the notorious "negative energy" problem.[37]) Now, as seen from the reference frame (γ_μ) related, we assume, to the electron's rest frame (γ'_μ) by a pure boost, so that

$$V \equiv p/m = \gamma'_0 = V^\mu\gamma_\mu = H\gamma_0 H^{-1}, \tag{5.11}$$

the solution (5.7) is

$$\psi(x) = e^{-ip\cdot x}\phi'_+ \equiv e^{-ip\cdot x}\psi(p), \tag{5.12}$$

where

$$H = (1+V\gamma_0)[2(1+V\cdot\gamma_0)]^{-1/2} \tag{5.13}$$

arises from (2.23a) and where the spinor $\psi(p)$ satisfies

$$\psi(p) = \tfrac{1}{2}(1+V)\tfrac{1}{2}[1+(-i)\gamma_5 SV]\psi(p). \tag{5.14}$$

Equations (5.12) and (5.14), of course, merely express (5.7) and (5.8) in manifestly covariant form.

The problem now is that $\psi(p)$ has four components in the reference frame (γ_μ) with the spinor basis (ϕ_r,χ_r), although it remains in the two-dimensional subspace of the projection operator $\tfrac{1}{2}(1+V)$. What do these components now represent? If we write

$$\psi(p) \equiv \alpha_r(p)\phi_r + \beta_r(p)\chi_r, \tag{5.15}$$

then

$$\bar{\psi}\psi = \bar{\phi}'_+\,\phi'_+ = 1 = |\alpha_r|^2 - |\beta_r|^2, \tag{5.16}$$

so that $|\alpha_r|^2$ and $|\beta_r|^2$ cannot be interpreted as probabilities. However, since $\gamma_0\chi_r = -\chi_r$, we have

$$\bar{\psi}\gamma_0\psi = V\cdot\gamma_0 = \epsilon/m \equiv \gamma = (1-v^2)^{-1/2}$$
$$= |\alpha_r|^2 + |\beta_r|^2, \tag{5.17}$$

where v is the electron's speed, which leads me to propose the following interpretation: $(|\alpha_+|^2 + |\beta_-|^2)/\gamma$ and $(|\alpha_-|^2 + |\beta_+|^2)/\gamma$ are the probabilities, respectively, of the electron having spin up or down in the reference frame (γ_μ), where

$$\alpha_r = \bar{\phi}_r\,\psi, \tag{5.18a}$$

$$\beta_r = -\bar{\chi}_r\,\psi, \tag{5.18b}$$

and where the up direction is defined by the spacelike vector in the "spin" projection operator for (ϕ_r,χ_r); for example, if (3.3) is used, the up direction is the 3-direction as $\gamma_3\gamma_0 = \sigma_3$. It is important to note that, in general, there is no spin-up *amplitude*, for example, as there is in nonrelativistic quantum mechanics.

The spinor components α_r,β_r can be readily calculated using (4.1) and (3.10), with (3.3) used as the spin projection operators:

$$|\alpha_+|^2 = |\bar{\phi}_+\psi|^2 = \bar{\phi}_+\psi\bar{\psi}\phi_+ = 4S[\psi\bar{\psi}\phi_+\bar{\phi}_+]$$
$$= 4S[\tfrac{1}{2}(1+V)\tfrac{1}{2}(1-i\gamma_5 SV)\tfrac{1}{2}(1+\gamma_0)\tfrac{1}{2}(1-i\gamma_5\gamma_3\gamma_0)]$$
$$= \tfrac{1}{4}(1+V\cdot\gamma_0 - S\cdot\gamma_3 - S\cdot\gamma_3\,V\cdot\gamma_0 + S\cdot\gamma_0\,V\cdot\gamma_3) \tag{5.19a}$$

$$|\alpha_-|^2 = |\bar{\phi}_-\psi|^2 = \tfrac{1}{4}(1+V\cdot\gamma_0 + S\cdot\gamma_3$$
$$+ S\cdot\gamma_3\,V\cdot\gamma_0 - S\cdot\gamma_0\,V\cdot\gamma_3) \tag{5.19b}$$

$$|\beta_+|^2 = |\bar{\chi}_+\psi|^2 = \tfrac{1}{4}(-1+V\cdot\gamma_0 + S\cdot\gamma_3$$
$$- S\cdot\gamma_3\,V\cdot\gamma_0 + S\cdot\gamma_0\,V\cdot\gamma_3) \tag{5.19c}$$

$$|\beta_-|^2 = |\bar{\chi}_-\psi|^2 = \tfrac{1}{4}(-1+V\cdot\gamma_0 - S\cdot\gamma_3$$
$$+ S\cdot\gamma_3\,V\cdot\gamma_0 - S\cdot\gamma_0\,V\cdot\gamma_3), \tag{5.19d}$$

with (5.16) and (5.17) clearly satisfied. The electron's spin vector in its rest frame (γ'_μ) is $S \equiv S_0^i\,\gamma'_i$ (as $S\cdot\gamma'_0 = 0$) and the Lorentz transformation in the form[38]

$$\gamma'_\mu = H\gamma_\mu H^{-1}$$
$$= \gamma_\mu + (1+\gamma'_0\cdot\gamma_0)^{-1}[-\gamma'_0\cdot\gamma_\mu(\gamma_0+\gamma'_0)$$
$$+ \eta_{\mu 0}(\gamma'_0 - \gamma_0 + 2\gamma'_0\gamma'_0\cdot\gamma_0)] \tag{5.20}$$

results in the following spin-up (P_u) and spin-down (P_d) probabilities:

$$P_u = \gamma^{-1}(|\alpha_+|^2 + |\beta_-|^2)$$
$$= \tfrac{1}{2}(1+S_0^3/\gamma) + \tfrac{1}{2}[(\gamma-1)/\gamma]S_0^i\hat{p}\cdot\sigma_i\,\hat{p}\cdot\sigma_3, \tag{5.21a}$$

$$P_d = \gamma^{-1}(|\alpha_-|^2 + |\beta_+|^2)$$
$$= \tfrac{1}{2}(1-S_0^3/\gamma) - \tfrac{1}{2}[(\gamma-1)/\gamma]S_0^i\hat{p}\cdot\sigma_i\,\hat{p}\cdot\sigma_3, \tag{5.21b}$$

where $\hat{p}$ is a unit 3-vector in the direction of the electron's momentum. In the nonrelativistic limit $(\gamma\to 1,\beta_r\to 0)$ (5.21) clearly reduces to the usual expressions,[30] the first terms of (5.21). As a check on signs it follows from (5.21) that a spin-up electron, boosted in the up direction, remains spin-up.

The solution for the spin $\hat{s}$ electron in its rest frame could equally well be represented by (5.9) or, in covariant form,

$$\psi(x) = e^{+ip\cdot x}\psi(p), \qquad (5.22)$$

where rather than (5.14) we have

$$\psi(p) = \tfrac{1}{2}(1 - V)\tfrac{1}{2}[1 + (-i)\gamma_5 SV]\,\psi(p). \qquad (5.23)$$

This solution differs from the one given above only in the replacements $\phi_+ \leftrightarrow \chi_-$, $\phi_- \leftrightarrow \chi_+$ in (5.19a)–(5.19d); there is no change in the physical interpretation. Indeed, the two solutions given here are spin-reversed, charge conjugates of each other as will be shown in Sec. 5D.

B. Electron in a magnetic field

The manifestly covariant formalism of Secs. 2 and 3 provides for an especially simple solution for the electron spin components in a magnetic field. The Dirac equation in the presence of an electromagnetic field is

$$(i\partial - m - eA(x))\psi(x) = 0, \qquad (5.24)$$

where the vector potential A is related to the electromagnetic field $F = \frac{1}{2}F^{\mu\nu}\gamma_\mu \wedge \gamma_\nu$ by[39]

$$F = \partial \wedge A = \mathbf{E} + \gamma_5 \mathbf{B}, \qquad (5.25)$$

where the electric and magnetic fields depend on a specific reference frame (γ_μ) having been singled out. If the electron is assumed to be spin-up, at rest, at time $t = 0$ in a uniform magnetic field, then a general solution to (5.24) is

$$\psi(x) = \psi(t) = e^{-imt}R(t)\phi_+, \qquad (5.26)$$

where $R(t)$ is a rotation operator or multivector, as in (2.14), with $R(0) \equiv 1$. This assumed solution gives

$$i\frac{d\psi}{dt} \equiv i\dot{\psi} = (m + i\dot{R}R^{-1})\psi = (m - \tfrac{1}{2}i\gamma_5\Omega\gamma_0)\,\psi, \qquad (5.27)$$

where (2.29) has been employed, with Ω being a spacelike angular velocity 4-vector.

From (5.24) follows

$$(\partial^2 + m^2 - e^2A^2 + 2ieA\cdot\partial + ieF)\psi = 0, \qquad (5.28)$$

and when (5.26) is inserted here, one obtains

$$2m\dot{R}R^{-1}\psi + e\gamma_5\gamma_0\Omega\cdot A\psi + i\frac{d(\dot{R}R^{-1})}{dt}\psi = eF\psi, \qquad (5.29)$$

where a pure magnetic field has been assumed $(A\cdot\gamma_0 = 0)$. The second and third terms are small relativistic corrections (in the weak field limit); if ignored, (5.27) and (5.29) lead to

$$\dot{R}R^{-1} = -\tfrac{1}{2}\gamma_5\Omega\gamma_0 = -\tfrac{1}{2}\gamma_5\boldsymbol{\omega}$$
$$\approx \tfrac{1}{2}(e/m)F = \tfrac{1}{2}(e/m)\gamma_5\mathbf{B}, \qquad (5.30)$$

where $\boldsymbol{\omega} = \Omega \wedge \gamma_0$. Thus, if a Hamiltonian $\hat{H}$ is defined as

$$i\dot{\psi} = \hat{H}\psi, \qquad (5.31)$$

we find

$$(\hat{H} - m)\psi = -\tfrac{1}{2}i\gamma_5\Omega\gamma_0\psi \approx \tfrac{1}{2}i(e/m)\gamma_5\mathbf{B}\psi, \qquad (5.32)$$

which in the transition to the Pauli algebra $(\gamma_5 \to i)$ becomes

$$(\hat{H} - m)\psi = -\tfrac{1}{2}(e/m)\mathbf{B}\psi, \qquad (5.33)$$

which is the Pauli equation.[30] If the magnetic field is constant as well as uniform the solutions given above are exact (the second and third terms of (5.29) vanish) and the solution from (2.14) is

$$R(t) = \exp(-\tfrac{1}{2}\gamma_5\Omega\gamma_0 t) = \exp(-\tfrac{1}{2}i\omega t). \qquad (5.34)$$

Since $R\gamma_0 = \gamma_0 R$, the magnetic field only induces transitions between states ϕ_+ and ϕ_- (not χ_+ or χ_- if these states are initially absent); thus the two-component Pauli spinors are sufficient to express the solution. The spin-up amplitude at time t would be $e^{-imt}\bar{\phi}_+R(t)\phi_+$ and the spin-down amplitude is $e^{-imt}\bar{\phi}_-R(t)\phi_+$.

C. Basis spinors for massless spin $-\frac{1}{2}$ particles

The projection operators (3.2) and (3.4) cannot be used for massless Dirac particles because no 4-vector of the form $V = p/m$, $V^2 = 1$, exists for such particles. It makes no difference if the normalization $\bar{\phi}_+\phi_+ = 1$, for example, is changed to the commonly used $\bar{\phi}_+\phi_+ \equiv 2m$ because $(m \pm p)$ are not ("unnormalized") projection operators if $m = 0$. The vector space of spinors is still four-dimensional so that new projection operators are needed.

The Dirac equation to be satisfied by free, massless particles is now simply

$$\partial\psi(x) = 0, \qquad (5.35)$$

which requires as basis spinors those $(\omega_r, r = \pm)$ that satisfy

$$p\omega_r(p) = 0 \qquad (5.36a)$$
$$= \epsilon\gamma_0(1 - \hat{p})\omega_r(p), \qquad (5.36b)$$

where the particle's 4-momentum p $(p^2 = 0)$ is written, as before, as $p = \epsilon(1, \hat{p})$. The spin 4-vector of Sec. 3 will be such that

$$S \wedge V = SV = \hat{s} \equiv \hat{p}, \qquad (5.37)$$

so that (5.36b) suggests that the spin projection operators become[40]

$$\Sigma_\pm = \tfrac{1}{2}[1 \pm (-i)\gamma_5], \qquad (5.38)$$

since $\hat{p}\omega_r(p) = \omega_r(p)$ from (5.36b). Thus the spinors $\omega_r(p)$, $\lambda_r(p)$ defined by

$$\tfrac{1}{2}(1 + \hat{p})\Sigma_\pm\,\omega_\pm(p) \equiv \omega_\pm(p), \qquad (5.39a)$$
$$\tfrac{1}{2}(1 - \hat{p})\Sigma_\pm\,\lambda_\mp(p) \equiv \lambda_\mp(p) \qquad (5.39b)$$

span spinor space and make a suitable basis. They satisfy

$$\bar{\omega}_r\omega_s = 0 = \bar{\lambda}_r\lambda_s, \qquad (5.40)$$

because (5.39a), for example, has as an adjoint equation

$$\bar{\omega}_\pm(p)\tfrac{1}{2}(1 - \hat{p})\Sigma_\mp = \bar{\omega}_\pm(p), \qquad (5.41a)$$

as well as (5.36a) and

$$p'\lambda_r(p) = 0 = p\lambda_r(p'), \qquad (5.41b)$$

where $p' \equiv \epsilon(1, -\hat{p})$ if $p = \epsilon(1, \hat{p})$.

An expansion of the massless spinor field in plane waves in the form

$$\psi(x) = (2\pi)^{-3/2}\int d^3p\,(e^{-ip\cdot x}a_r\omega_r(p) + e^{ip\cdot x}b_r^*\lambda_r(p')) \qquad (5.42)$$

satisfies the Dirac equation (5.35). The choices of sign in (5.39) are again made because of the physical interpretation. Note, however, that $\omega_r(p)$ and $\lambda_{-r}(p')$ satisfy the same defining equations, thus rendering the basis $\lambda_r(p)$ superfluous. The plane wave expansion expressed by

$$\psi(x) = (2\pi)^{-3/2}\int d^3p\,(e^{-ip\cdot x}a_r + e^{ip\cdot x}b_{-r}^*)\omega_r(p) \qquad (5.43)$$

leads to the correct field 4-momentum

$$P = \int d^3p\, p(a_r^* a_r + b_r^* b_r) \tag{5.44}$$

and spin in the direction of motion

$$\mathbf{J}\cdot\hat{P} = \frac{1}{2}\int d^3p\ (a_+^* a_+ - a_-^* a_- + b_+^* b_+ - b_-^* b_-). \tag{5.45}$$

Therefore, one needs only the spinors $\omega_\pm(p)$, $\bar{\omega}_\pm(p)$ in Feynman diagrams and in any other calculations concerning massless spin $-\frac{1}{2}$ particles. Practical calculations are aided by the definition

$$\bar{\omega}_+\gamma_0\omega_+ \equiv 1, \tag{5.46a}$$

which, along with (5.40), leads to

$$\bar{\omega}_r\gamma_0\omega_s = \delta_{rs}, \tag{5.46b}$$

$$\omega_\pm\bar{\omega}_\pm = (p/2\epsilon)\Sigma_\mp\,. \tag{5.46c}$$

Because of (5.46c) no factors such as $1/2\epsilon$ are put in the invariant amplitude when Feynman diagrams are being evaluated. Such factors, for unit reasons, are also absent from (5.43); they emerge naturally from (5.46c).

D. Manifestly covariant formalism for *CPT*

It seems to be a universal practice to exploit specific matrix representations of the Dirac algebra in order to construct C, P, and T transformations with the desired properties. This, clearly, is not acceptable if, in the spirit of the covariant space-time approach, only manifestly covariant equations are to be considered. It is therefore important to show that manifestly covariant C, P, and T transformations can be constructed. These transformations seem more important in the field-quantized version of the Dirac theory, so it will be within that context that they will be addressed here.

I will first propose that the charge conjugation (C) transformation be changed from the usual[41] matrix or representation dependent one to

$$\psi^c(x) = \mathscr{C}\psi(x)\mathscr{C}^{-1} \equiv \gamma_5\,\psi^*(x), \tag{5.47}$$

where $\mathscr{C}$ is the usual unitary Fock-space operator and where ψ^* represents the complex conjugate of complex numbers and the Hermitian conjugate of the Fock-space creation and annihilation operators, but which does not affect elements of the Dirac algebra ($\gamma_\mu^* \equiv \gamma_\mu$; vectors are *real*), although it does imply transformations among the basis spinors because of the unit imaginary factor in the spin projection operators. Apart from a negative sign, the phase in (5.47) is unique. One sees from (5.1) that

$$u_r^* \propto u_{-r}, \quad v_r^* \propto v_{-r} \tag{5.48a}$$

and

$$\gamma_5\, u_r^* \propto v_r, \quad \gamma_5\, v_r^* \propto u_r. \tag{5.48b}$$

One could set the phases, for example, as $\gamma_5\, u_+^* \equiv v_+$, but there is no advantage to be gained here by doing so as the C-transformation defined here does not differ all that much from the usual except that no matrix transpose is used here.

The solution to the free-particle Dirac equation for an electron of momentum $p = mV$ and spin S ($S \wedge V = \hat{s}$ = spin direction in rest frame), considered earlier in Sec. 5A, Eqs. (5.12) and (5.14),

$$\psi(x) = e^{-ip\cdot x}\psi(p),$$
$$\psi(p) = \tfrac{1}{2}(1+V)\tfrac{1}{2}[1 + (-i)\gamma_5 SV]\ \psi(p), \tag{5.49}$$

has as a charge conjugate

$$\psi^c(x) = e^{ip\cdot x}(\gamma_5\,\psi^*(p)), \tag{5.50a}$$

with

$$\gamma_5\,\psi^*(p) = \tfrac{1}{2}(1-V)\tfrac{1}{2}[1 - (-i)\gamma_5 SV]\gamma_5\,\psi^*(p), \tag{5.50b}$$

which is spin-opposite to the alternate solution (5.22) and (5.23). When applied to (5.3) or (5.44), for example, (5.47) interchanges particles and anti-particles (but not spin).

Since the Dirac equation and any conceivable interaction (that does not involve γ_5: the weak interaction is an important exception that is discussed at the end of this subsection) are to be written in manifestly covariant form, such expressions are automatically invariant under Lorentz transformations, including the improper transformations of space inversion (P) and time reversal (T), which therefore must be differently formulated if a nontrivial meaning is to be attached to such transformations.

The space-time transformation to be adopted is the "active" transformation

$$\begin{aligned} x = x^\mu\gamma_\mu \to x' &= x'^\mu\gamma_\mu \\ &\equiv LxL^{-1} \\ &\equiv x^\mu\gamma'_\mu, \end{aligned} \tag{5.51}$$

where x^μ and x'^μ are the coordinates of the old and new events, respectively, in the basis (γ_μ), and where (γ'_μ) can be thought of as a new basis. The coordinates of the *new* event in the *old* basis are defined to be the same as those of the *old* event in the *new* basis. It is easy to show that for infinitesimal (proper) Lorentz transformations and, at least, the improper Lorentz transformation of space inversion in the reference frame (γ_μ) $(L_p = \gamma_0)$,

$$\gamma'_\mu = L_p\,\gamma_\mu L_p^{-1} = (\gamma_0, -\gamma_i), \tag{5.52}$$

and time reversal $(L_t = \gamma_5\gamma_0)$

$$\gamma'_\mu = L_t\gamma_\mu L_t^{-1} = (-\gamma_0, \gamma_i), \tag{5.53}$$

that

$$\partial' \equiv \gamma^\mu\partial'_\mu = L\gamma^\mu L^{-1}\partial_\mu = L\partial L^{-1}, \tag{5.54}$$

so that the Dirac equation, for example,

$$(i\partial - eA(x) - m)\psi(x) = 0 \tag{5.55a}$$

becomes

$$(i\partial' - eA'(x') - m)\psi'(x') = 0, \tag{5.55b}$$

where $A'(x') \equiv LA(x)L^{-1}$ and $\psi'(x') \equiv L\psi(x)$ are *new* fields at the new event. Thus the space inversion (P) or parity transformation is, from (5.52),

$$\begin{aligned}\psi^P(t,\mathbf{r}) &= \mathscr{P}\psi(t,\mathbf{r})\mathscr{P}^{-1} \\ &\equiv \gamma_0\,\psi(t,-\mathbf{r}),\end{aligned} \tag{5.56}$$

which is identical in every way to the usual[42] transformation.

What one might now be tempted to define as the time-reversal (T) transformation is, however, complicated by the physical interpretation. The mapping (5.51) with the time-reversal Lorentz transformation (5.53) maps a timelike world line, say, one going into the future along the positive x^1 axis, into one going into the past also along the positive x^1 axis, but which *must* be physically interpreted as one going *into the future* along the negative x^1 axis, an interpretation that requires $\gamma_\mu \to -\gamma_\mu$ to make physical sense. Thus, although (5.53) and (5.55b) appear to map vectors, such as the momentum 4-vector p, as $p = (p^0, p^i) \to (-p^0, p^i)$, we require $\gamma_\mu \to -\gamma_\mu$ and $p \to -(-p^0, p^i)$ to make physical sense. This requires that the γ_5 factor in L_t be dropped and requires that the Fock-space operator $\mathscr{T}$ be *antilinear*:

$$\psi^t(t,\mathbf{r}) = \mathscr{T}\psi(t,\mathbf{r})\mathscr{T}^{-1} \equiv \gamma_0\,\psi(-t,\mathbf{r}), \tag{5.57}$$

which is to be contrasted with the usual[43] representation-dependent transformation. One should note that a *linear* $\mathscr{T}$ would have to accompany $L_t = \gamma_5\gamma_0$ if the physical interpretation is acceptable, which it is not.

The current components

$$j^\mu(x) = \bar\psi(x)\gamma^\mu\psi(x) \tag{5.58}$$

transform into

$$\begin{aligned} j^{t\mu}(t) &= \bar\psi^t(t)\gamma^\mu\psi^t(t) \\ &= \bar\psi(-t)\gamma_0\gamma^\mu\gamma_0\,\psi(-t) \\ &= (j^0(-t), -j^i(-t)), \end{aligned} \tag{5.59}$$

and as the electromagnetic field potential A transforms as

$$A^\mu(t) \to (A^0(-t), -A^i(-t)), \tag{5.60}$$

the electromagnetic interaction is seen to be T-invariant.

The space-time inversion is expressed by

$$\begin{aligned} \psi^{tp}(x) &= (\mathscr{T}\mathscr{P})\psi(x)(\mathscr{T}\mathscr{P})^{-1} \\ &= \psi(-x), \end{aligned} \tag{5.61}$$

a manifestly covariant result as it must be, and the (antilinear) CPT-transformation for spinor fields is thus (to within a sign)

$$\begin{aligned} \psi^{cpt}(x) &= (\mathscr{C}\mathscr{P}\mathscr{T})\psi(x)(\mathscr{C}\mathscr{P}\mathscr{T})^{-1} \\ &= \gamma_5\,\psi^*(-x) = \psi^c(-x). \end{aligned} \tag{5.62}$$

If the weak interaction can be illustrated by the Fermi four-point interaction Lagrangian

$$\mathscr{L}_I \propto \bar\nu\gamma^\mu\,\tfrac{1}{2}(1+i\gamma_5)e\bar e\gamma_\mu\,\tfrac{1}{2}(1+i\gamma_5)\nu, \tag{5.63}$$

where ν and e, respectively, are the neutrino and electron spinor fields, then, although γ_5, the unit pseudoscalar, reverses sign under the improper Lorentz transformations (5.52) and (5.53), it is (as usual) not invariant under P but is invariant under T because of the special nature of (5.57) (as γ_5 was dropped from L_t and $\mathscr{T}$ made antilinear in compensation). Nor, as is usual, is $\mathscr{L}_I$ invariant under C, but is under CP and CPT.

E. Majorana fields

The subject of Majorana fields[44–47] is of intense current interest because of the possibility that neutrinos may not be massless after all. Can there exist, one wonders, a two-component massive field within the context of quantized Dirac fields? In a word, no. Massive fields necessarily have four components. But it is possible to construct a field in which two components (right-handed particles and left-handed antiparticles, say) are suppressed and vanish in the high energy limit.

The charge conjugation transformation defined in the previous subsection, Eq. (5.47), does not permit self-conjugate fields because

$$(\psi^c)^c = \gamma_5(\gamma_5\,\psi^*)^* = -\psi. \tag{5.64}$$

Nevertheless, the "chirality" projection operators

$$\Omega_\pm \equiv \tfrac{1}{2}[1 \pm (-i\gamma_5)], \tag{5.65}$$

identical to the spin projection operators (5.38) for massless fields, can be used to project any Dirac field ψ onto two subspaces:

$$\psi = \psi_+ + \psi_-, \quad \psi_\pm \equiv \Omega_\pm\,\psi. \tag{5.66}$$

(The reader should note that the notation R, L, rather than $+$, $-$, is frequently used; this is unfortunate because helicity is not the same thing as chirality except in the zero-mass limit.)

Let us begin with the massive field:

$$\psi(x) = (2\pi)^{-3/2}\int d^3p\left(\frac{m}{\epsilon}\right)^{1/2} \times (e^{-ip\cdot x}a_r u_r + e^{ip\cdot x}b_r^* v_r), \tag{5.3}$$

where, to use a more correct notation, $r = R, L$, not $\pm$, provided the spin vector S satisfies (5.37). If in (5.48) the phases are defined to be

$$u_R^* \equiv u_L, \tag{5.67a}$$

$$\gamma_5\,u_R \equiv v_L, \tag{5.67b}$$

we find

$$(u_{R,L})^c \equiv \gamma_5(u_{R,L})^* = v_{R,L}, \tag{5.68a}$$

$$(v_{R,L})^c \equiv \gamma_5(v_{R,L})^* = -u_{R,L}. \tag{5.68b}$$

These equations yield a number of results such as

$$\Omega_- v_r = \Omega_- u_{-r}, \tag{5.69}$$

so that

$$\psi_- \equiv \Omega_-\,\psi = (2\pi)^{-3/2}\int d^3p\left(\frac{m}{\epsilon}\right)^{1/2} \times (e^{-ip\cdot x}a_r + e^{ip\cdot x}b_{-r}^*)\Omega_- u_r \tag{5.70}$$

and

$$(\psi_-)^c \equiv \Omega_+\,\psi^c = (2\pi)^{-3/2}\int d^3p\left(\frac{m}{\epsilon}\right)^{1/2} \times (e^{+ip\cdot x}a_r^* + e^{-ip\cdot x}b_{-r})(-\Omega_+)u_{-r}, \tag{5.71}$$

fields that curiously resemble the massless field (5.44), so that a field χ defined by

$$\chi \equiv (\tfrac{1}{2})^{1/2}[\,\psi_- + (\psi_-)^c], \tag{5.72}$$

with an assumed Lagrangian

$$\mathscr{L} = -\bar\chi(i\partial - m)\chi, \tag{5.73}$$

leads to a field 4-momentum in the high energy limit

$$P = \int d^3p\,p(a_-^*\,a_- + b_+^*\,b_+), \tag{5.74}$$

where $R, L \rightarrow +, -$. Thus χ represents a massive field, a Majorana field, but one with right-handed particles and left-handed antiparticles suppressed, and one that is identical, in the high energy limit, to the massless field ψ in (5.44).

6. DISCUSSION

The Dirac equation can now be regarded as one that is constructed of objects of geometric significance that can be considered to be independent of specific frames of reference, a requirement that must be met before it can be said to be a manifestly covariant equation. The main conceptual change required for this is in regarding the (γ_μ) as the basis 4-vectors of a reference frame, with γ_0 the reference frame's 4-velocity, not as a set of four fixed matrices devoid of physical meaning.

A further requirement was the construction of the vector spaces ($\mathscr{S}$ and $\overline{\mathscr{S}}$) of spinors on which act the (γ_μ) and other multivectors of the Dirac algebra $\mathscr{D}$. In symbolic form the structure is

$$\mathscr{D} \times \mathscr{S} = \mathscr{S}, \tag{6.1a}$$

$$\overline{\mathscr{S}} \times \mathscr{D} = \overline{\mathscr{S}}, \tag{6.1b}$$

with the spinors and multivectors connected via

$$\overline{\mathscr{S}} \times \mathscr{S} = \mathscr{C} \tag{6.2a}$$

and

$$\mathscr{S} \times \overline{\mathscr{S}} = \mathscr{D}, \tag{6.2b}$$

where $\mathscr{C}$ here reresents the complex field, relations that are explicitly expressed in the important relations (4.1) and (4.2), respectively, which show that the spinor component transformations (3.16) exist and can be calculated. This is the necessary prerequisite to making the claim that Dirac spinors have the required Lorentz transformation properties and that no matrix representation of the Dirac algebra is needed.

I hope to have shown that this formalism is an aid to the physical interpretation of the solutions of the Dirac equation, particularly in the case of spin projections and probabilities such as those considered in Sec. 5A, where an old problem in nonrelativistic quantum mechanics acquires new subtleties in its relativistic formulation, subtleties that are best considered using manifestly covariant concepts.

There are two results of this paper that could be of use to those doing calculations in weak interaction physics and its incorporation in unified theories. The charge conjugation transformation defined in (5.47), as well as being representation and matrix independent, is easier to use than the usual C-transformation, especially in conjunction with the chirality operators $\Sigma_\pm \equiv \frac{1}{2}[1 \pm (-i)\gamma_5]$ (which are the same as the helicity projection operators for massless fields). This is a consequence of taking the basis vectors (γ_μ) to be real, something that cannot be done with matrices. Secondly, the basis spinors of Sec. 5C for massless particles are particularly simple to work with, whether for massless particles or massive particles in the extreme relativistic limit, because there is only one two-component basis spinor that needs to be considered, and it obeys the simple rules (5.46a)–(5.46c).

Perhaps the most useful application of the present results is to the practical calculation of amplitudes, or "traces," where the results of Sec. 4 can be used to considerably simplify the frequent and lengthy calculations required in the evaluation of cross sections in which Dirac particles are participating. This alone may be considered justification for acquiring a facility in the techniques of the geometric algebra.

ACKNOWLEDGMENT

I would like to thank Dr. P. Kalyniak for convincing me of the existence of a very serious error in the first draft of this paper.

[1]A typical viewpoint seems to be "... the γ-matrices remain unaltered under [a] Lorentz transformation," in S. S. Schweber, *Introduction to Relativistic Quantum Field Theory* (Harper and Row, New York, 1961), Chap. 4.
[2]C. Itzykson and J. -B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980), Appendix A2.
[3]R. H. Good, Jr., Rev. Mod. Phys. **27**, 187 (1955).
[4]H. Feshbach and F. Villars, Rev. Mod. Phys. **30**, 24 (1958).
[5]W. L. Bade and H. Jehle, Rev. Mod. Phys. **25**, 714 (1953).
[6]I. M. Gel'fand, R. A. Minlos, and Z. Ya. Shapiro, *Representations of the Rotation and Lorentz Groups and Their Applications* (Macmillan, New York, 1963).
[7]E. Cartan, *Theory of Spinors* (MIT Press, Cambridge, MA, 1966).
[8]J. M. Jauch and F. Rohrlich, *The Theory of Photons and Electrons* (Springer-Verlag, New York, 1976), 2nd ed., Appendix A2.
[9]J. O. Eeg, J. Math. Phys. **21**, 170 (1980).
[10]Ref. 5, "View (A)," p. 715, but which is summarily dismissed.
[11]R. P. Feynman, *Quantum Electrodynamics* (Benjamin, New York, 1962), p. 44.
[12]N. Salingaros, J. Math. Phys. **22**, 226 (1981).
[13]D. Hestenes, *Space–Time Algebra* (Gordon and Breach, New York, 1966).
[14]D. Hestenes, J. Math. Phys. **8**, 798 (1967); **14**, 893 (1973); **16**, 556 (1975).
[15]D. Hestenes and R. Gurtler, J. Math. Phys. **16**, 573 (1975).
[16]D. Hestenes, Am. J. Phys. **39**, 1013 (1971); **47**, 399 (1979).
[17]D. Hestenes and R. Gurtler, Am. J. Phys. **39**, 1028 (1971).
[18]S. Teitler, J. Math. Phys. **7**, 1730, 1739 (1966).
[19]C. N. Ktorides, J. Math. Phys. **16**, 2123 (1975).
[20]K. Greider, Phys. Rev. Lett. **44**, 1718 (1980).
[21]N. Salingaros, J. Math. Phys. **23**, 1 (1982).
[22]C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973), Chap. 41.
[23]H. Flanders, *Differential Forms* (Academic, New York, 1963).
[24]It seems that many, if not most, authors today follow J. D. Bjorken and S. D. Drell, *Relativistic Quantum Mechanics* (McGraw-Hill, New York, 1964), Appendix A, and define $\gamma_5 = -i\gamma_0\gamma_1\gamma_2\gamma_3$. There are good reasons for our choice: see Eq. (2.10).
[25]The familiar notations (γ_μ) and (σ_i), rather than, say, (e_μ) and $(\hat{e}_i)$, used to denote bases in space-time and space, respectively, are chosen only to exploit the reader's familiarity with their respective algebras.
[26]J. D. Hamilton, Can. J. Phys. **59**, 213 (1981).
[27]P. Rastall, Can. J. Phys. **44**, 3109 (1966).
[28]W. E. Baylis, Am. J. Phys. **48**, 918 (1980).
[29]For other reasons this is the notation used in Ref. 1.
[30]J. D. Hamilton, Am. J. Phys. **52**, 56 (1984).
[31]Ref. 24, Chap. 3.
[32]Ref. 14 (third paper).
[33]J. Beckers, J. Harnad, and P. Jasselette, J. Math. Phys. **21**, 2491 (1980).
[34]P. Ramond, *Field Theory: A Modern Primer* (Benjamin-Cummings, Reading, MA, 1981), p. 58.
[35]Y. Takahashi, Phys. Rev. D **26**, 2169 (1982); J. Math. Phys. **24**, 1783 (1983).
[36]J. D. Bjorken and S. D. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965), Chap. 13.
[37]This is problematic. See Ref. 24, Sec. 3.3.
[38]Ref. 26, Sec. 3.
[39]Ref. 13, Chap. 8.
[40]Ref. 24, Sec. 7.10.
[41]Ref. 36, Sec. 15.12.
[42]Ref. 36, Sec. 15.11.
[43]Ref. 36, Sec. 15.13.
[44]K. M. Case, Phys. Rev. **107**, 307 (1957).
[45]D. L. Weaver, J. Math. Phys. **17**, 485 (1976).
[46]P. Langacker, Phys. Rep. **72**, 185 (1981), Sec. 2.4.
[47]L. F. Li and F. Wilczek, Phys. Rev. D **25**, 143 (1982).

Classical versus quantum integrability

Jarmo Hietarinta
Department of Physical Sciences, University of Turku, 20500 Turku 50, Finland

(Received 8 September 1983; accepted for publication 16 December 1983)

Classical integrability and quantum integrability are compared for two degrees of freedom Hamiltonian systems. We use c-number representatives for quantum operators and the Moyal bracket for the commutator. Three different cases are found: (i) the c-number representative of the quantum mechanical second invariant is identical to the classical second invariant, (ii) $O(\hbar^2)$ corrections are needed in the classical second invariant to obtain the quantum invariant, and (iii) also the potential must be deformed by an $O(\hbar^2)$ term. Several examples from the Henon–Heiles and Holt families of integrable potentials are included.

PACS numbers: 03.65.Fd, 03.20. + i, 02.30. + g

I. INTRODUCTION

In this paper we compare classical integrability (CI) and quantum integrability (QI). Such a comparison is perhaps most meaningful at the algebraic level, where classical mechanics and quantum mechanics are closest to each other.

A classical Hamiltonian system of n degrees of freedom is defined to be classically integrable if there are $n-1$ independent, well-defined, global functions, whose Poisson brackets with each other and with the Hamiltonian vanish. We carry this definition to quantum mechanics: A quantum mechanical Hamiltonian of n degrees of freedom is defined to be quantum integrable if there are $n-1$ independent, well-defined, global operators, which commute with each other and the Hamiltonian.

The comparison of classical and quantum integrability is made easier in many respects if we also represent the quantum operators by c-number functions. Such a correspondence can be well defined, it is discussed in Sec. II. It is clear that since classical and quantum mechanics are not algebraically isomorphic, differences must be present somewhere if we use c-number functions in both. Indeed, the commutator will become the Moyal bracket, which agrees with the Poisson bracket only when $\hbar \to 0$.

Using these tools we will compare CI and QI for systems of 2 degrees of freedom. In Sec. III, we discuss situations where the same Hamiltonian is both CI and QI. However, the classical second invariant may not qualify as the c-number representative of the quantum operator. If the classical invariant is at most second order in p it will also work for QI. For invariants of order p^3 and p^4 we derive the extra quantum conditions and compute the quantum corrections to the second invariant for several examples. We also discuss the possibility that the quantum corrections can be eliminated by adding αH^2 to I_4 or by going to another ordering rule.

In Sec. IV, a more complex situation is discussed. We find that in some cases it is necessary to deform the classical potential with an $O(\hbar^2)$ term to be able to construct a quantum invariant. For the family of Holt potentials $Cx^{4/3}+y^2x^{-2/3}$ the deformation is $-\frac{5}{72}\hbar^2x^{-2}$ and for the Fokas–Lagerstrom potential $(xy)^{-2/3}$ it is $-\frac{5}{72}\hbar^2(x^{-2}+y^{-2})$. It is interesting to note that the numerical factors are always the same and associated with the fractional power $-\frac{2}{3}$, also they cannot be scaled away.

II. C-NUMBER REPRESENTATION OF QUANTUM OPERATORS

In this paper we compare classical and quantum mechanical integrability and for that purpose it will be useful to also represent quantum mechanical operators by c-number functions. Of course, since classical mechanics is not quantum mechanics even at the algebraic level, there will be important differences, but c-number representation will be nevertheless useful in practical computations, especially when computer algebra is used.

The relationship between operators and c-number functions has been studied since the early days of quantum mechanics. Attempts were then made to construct quantum mechanics out of classical mechanics: different correspondence rules would give different operators from the same function of classical mechanics. We take another approach, for us the operator is fixed and only for convenience do we want a c-number representation for it. Different correspondence rules will then give different c-number representations.

Let us assume that we have n canonical coordinates and moments. In classical mechanics we have

$$\{x_i, p_j\}_{PB} = \delta_{ij}, \tag{2.1}$$

where

$$\{A, B\}_{PB} = \sum_{i=1}^{n} \left(\frac{\partial A}{\partial x_i}\frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i}\frac{\partial B}{\partial x_i} \right), \tag{2.2}$$

and in quantum mechanics

$$[\hat{x}_i, \hat{p}_j] = i\hbar\delta_{ij}, \tag{2.3}$$

$$[A,B] = AB - BA, \tag{2.4}$$

$$\begin{aligned} \hat{x}_i\psi &= x_i\psi, \\ \hat{p}_j\psi &= -i\hbar\frac{\partial\psi}{\partial x_j}. \end{aligned} \tag{2.5}$$

Many correspondence rules have been discussed in the literature.[1–3] (For an overview see Ref. 2.) The correspondence rules are usually characterized by a function of $2n$ variables $\mathscr{F}(x,y)$ and given by the (formal) integral[2,3]

$$\hat{\mathscr{A}}(\hat{p},\hat{q}) = \int d^np\, d^nq\, d^nx\, d^ny (2\pi\hbar)^{-2n}\, \mathscr{F}(x,y) A_{\mathscr{F}}(p,q)$$

$$\times \exp\left[\frac{i}{\hbar}(x\cdot(\hat{p}-p)+y\cdot(\hat{q}-q))\right]. \tag{2.6}$$

The subscript $\mathscr{F}$ is a reminder that for us the operator is fixed and in different ordering rules it will possibly have different c-number representations. For all practical purposes it is enough to assume

$$\mathscr{F}(x,y)=F\left(\frac{i}{\hbar}x\cdot y\right), \tag{2.7}$$

where

$$F(0)=1.$$

In the following we will mostly use only two ordering rules:

1) the Weyl rule, which is given by

$$\mathscr{F}_W(x,y)=1, \tag{2.8}$$

and 2) the standard rule

$$\mathscr{F}_S(x,y)=\exp\left(-\frac{i}{2\hbar}x\cdot y\right). \tag{2.9}$$

The Weyl rule will be convenient in computing the commutator, while the standard rule is easiest when we want to construct the c-number representation for a given operator. This construction goes as follows: For any given operator $\hat{A}$ let us commute in each monomial the $\partial/\partial x_i$ operators to the right and x_i's to the left (= the standard ordering). The representative $A_S(p,x)$ can now be obtained from this ordered form of the operator by the substitution

$$\hat{x}_i \to x_i,\qquad \frac{\partial}{\partial x_i}\to\frac{i}{\hbar}p_i. \tag{2.10}$$

The Weyl representative can finally be obtained from the standard one by

$$A_W(p,x)=\exp\left(\frac{i\hbar}{2}\sum_i\frac{\partial^2}{\partial x_i\partial p_i}\right)A_S(p,x). \tag{2.11}$$

In general the transformation from one ordering rule to another goes by

$$A_{\mathscr{F}_1}(p,q)=\int d^nx\, d^ny\, d^nk\, d^nl\,(2\pi\hbar)^{-2n}[\mathscr{F}_1(x,y)]^{-1}$$
$$\times\mathscr{F}_2(x,y)A_{\mathscr{F}_2}(k,d)\exp\frac{i}{\hbar}[x\cdot(p-k)+y\cdot(q-l)]. \tag{2.12}$$

As an example consider the operator

$$\hat{\mathscr{A}}=x\partial_x^2\partial_y+y\partial_x\partial_y^2+B\partial_x\partial_y. \tag{2.13}$$

This is in the standard ordering, therefore,

$$A_S=(i/\hbar)^3[xp_x^2p_y+yp_xp_y^2+B(\hbar/i)p_xp_y], \tag{2.14}$$

and from (2.11)

$$A_W=(i/\hbar)^3[xp_x^2p_y+yp_xp_y^2+(\hbar/i)(B-2)p_xp_y]. \tag{2.15}$$

Once the correspondence (2.6) is given (from now on we assume the Weyl rule $F\equiv 1$) the relevant algebraic question is the following: If the operators $\hat{\mathscr{A}}$, $\hat{\mathscr{B}}$, and $\hat{\mathscr{C}}$ are related according to $[\hat{\mathscr{A}},\hat{\mathscr{B}}]=i\hbar\hat{\mathscr{C}}$, how are the corresponding c-number representations A, B, and C related? The answer was given in 1947 by Moyal[4]:

$$\left[\frac{2}{\hbar}\sin\left\{\frac{\hbar}{2}\left(\frac{\partial}{\partial q_A}\cdot\frac{\partial}{\partial p_B}-\frac{\partial}{\partial p_A}\cdot\frac{\partial}{\partial q_B}\right)\right\}\times A(p_A,q_A)B(p_B,q_B)\right]_{\substack{p_A=p_B=p\\ q_A=q_B=q}}=C(p,q). \tag{2.16}$$

[For a straightforward derivation of (2.16) see also Sec. V.A of Ref. 5.] It is important to keep in mind that the derivatives in (2.16) must be computed before the indicated substitutions (cf. Ref. 6 footnote 1).

First thing to note of (2.16) is that as $\hbar\to 0$ we recover the Poisson bracket. Since in the present paper we will only be concerned about integrability, we will have $C=0$ and $A=H$, $B=I$ (say). If the second invariant is at most second order in p, like H is, then the Moyal bracket (2.16) reduces to the Poisson bracket. This means that the classical and quantum invariants are identical in this representation, and the results can be applied to both cases. For example the extensive quantum mechanical results of Makarov *et al.*[7] can immediately be converted to classical mechanics.

In quantum mechanics we will of course often get additional, purely quantal effects. One important property of the Moyal bracket (2.16) is that it is odd under $p\to -p$ like the Poisson bracket. (In many other ordering rules the bracket would not have this property.) Since our Hamiltonian always is even in p we see that also the Weyl representation of the quantum mechanical second invariant must have a definitive parity under $p\to -p$. This will be very useful in the search for invariants, and motivates our use of the Weyl rule in obtaining the c-number representatives.

To end this introduction to c-number representations of operators we will discuss a peculiarity all such rules have. The Poisson bracket has the property $\{f(H), g(H)\}_{PB}=0$ and for operators we similarly have $[f(\hat{H}),g(\hat{H})]=0$ (at least if f and g are polynomials with scalar coefficients). However, there is no ordering of type (2.6) that would give required correspondence $f(H)\leftrightarrow f(\hat{H})$.[2] For classical invariants we have freely added powers of H to make the invariant simpler, but for quantum invariants one must be more careful. As can be inferred from (2.16), any such extra terms must be accompanied with $O(\hbar^2)$-contributions to preserve commutativity with respect to the Moyal bracket. This is the price we have to pay for using c-number representation of operators.

III. CORRECTIONS TO THE SECOND INVARIANT

In this section we will discuss classically integrable models which are also quantum integrable. We assume the Hamiltonian to be of the form

$$H=\tfrac{1}{2}p_x^2+\tfrac{1}{2}p_y^2+V(x,y), \tag{3.1}$$

so there are no ordering ambiquities in H. If the second invariant is also second order in p, then the corresponding quantum operator can be obtained from it using Weyl rule on the classical invariant. We will now take a closer look on the possible modifications necessary if the second invariant is of order p^3 or p^4.

A. I is of order p^3

Let us assume

$$I_3=f_0p_x^3+f_1p_x^2p_y+f_2p_xp_y^2+f_3p_y^3+g_0p_x+g_1p_y, \tag{3.2}$$

where f_i and g_i are functions of x and y. Since the Poisson bracket of H and I_3 vanishes, the following set of equations is satisfied,[8,9] at order p^4

$$\begin{aligned} &\partial_x f_0 = 0 \\ \partial_y f_0 + {}&\partial_x f_1 = 0 \\ \partial_y f_1 + {}&\partial_x f_2 = 0 \\ \partial_y f_2 + {}&\partial_x f_3 = 0 \\ &\partial_y f_3 = 0, \end{aligned} \tag{3.3}$$

at order p^2

$$\begin{aligned} \partial_x g_0 &= 3f_0 \partial_x V + f_1 \partial_y V \\ \partial_y g_0 + \partial_x g_1 &= 2f_1 \partial_x V + 2f_2 \partial_y V \\ \partial_y g_1 &= f_2 \partial_x V + 3f_3 \partial_y V, \end{aligned} \tag{3.4}$$

and finally at order p^0

$$g_0 \partial_x V + g_1 \partial_y V = 0. \tag{3.5}$$

For quantum integrability we must require that the Moyal bracket of H and I_3 vanishes. Now since $p \to -p$ parity is conserved by the Moyal bracket we know that the c-number representative of I in Weyl correspondence rule will also have the general form (3.2) without order p^2 or p^0 terms. When I is of order p^3 (or p^4), only the second term in the expansion of (2.16) might contribute. Due to the form of the Hamiltonian (3.1) ∂_p^3, $\partial_p^2 \partial_x$, and $\partial_p \partial_x^2$ all annihilate it and therefore only the contribution

$$-\frac{2}{\hbar}\frac{1}{3!}\left(\frac{\hbar}{2}\right)^3 \left(\sum_i \frac{\partial}{\partial q_i^H}\frac{\partial}{\partial p_i^I}\right)^3 V(q_i^H) I(p^I, q^I)\Bigg|_{\substack{p^I = p \\ q^I = q^H = q}} \tag{3.6}$$

might be nonzero. It is of order p^0, thus for the quantum case we find that the Moyal bracket vanishes provided that equations (3.3) and (3.4) hold, and if instead of (3.5) also the equation

$$\begin{aligned} g_0 \partial_x V + g_1 \partial_y V - \frac{\hbar^2}{4}[f_0 \partial_x^3 V \\ + f_1 \partial_x^2 \partial_y V + f_2 \partial_x \partial_y^2 V + f_3 \partial_y^3 V] = 0 \end{aligned} \tag{3.5Q}$$

is satisfied.

The result can also be stated as follows: Assume that the Hamiltonian (3.1) is classically integrable and has the second invariant of type (3.2). If

$$f_0 \partial_x^3 V + f_1 \partial_x^2 \partial_y V + f_2 \partial_x \partial_y^2 V + f_3 \partial_y^3 V = 0, \tag{3.7}$$

then the corresponding quantum Hamiltonian is also integrable and has a second invariant operator $\hat{I}_3$, whose c-number representative in the Weyl rule is identical to the classical invariant.

[The question of quantum integrability at order p^3 has also been studied in Ref. 10. However, the corresponding theorem 2.2 in Ref. 10 appears to be wrong: In (2.20) of Ref. 10 there should be $B_{12} = d_{112x} + d_{221y} + \tilde{b}_{12}$, and furthermore $\hat{a}_i$ and $\hat{b}_{ij}$ should all be zero. With these corrections (2.21) of Ref. 10 is the standard ordering version of our (3.7).]

Example 3.1: The Toda-type potential[8]

$$V = e^{\sqrt{3}x + y} + e^{-\sqrt{3}x + y} + e^{-2y} \tag{3.8}$$

has the following second invariant:

$$\begin{aligned} I_3 = p_x^3 - 3p_x p_y^2 + 3(e^{\sqrt{3}x + y} + e^{-\sqrt{3}x + y} - 2e^{-2y})p_x \\ - 3\sqrt{3}(e^{\sqrt{3}x + y} - e^{-\sqrt{3}x + y})p_y. \end{aligned} \tag{3.9}$$

It is easy to see that (3.7) holds and quantum integrability through Weyl rule is obtained. However, we get an even stronger result. Writing the invariant in the form

$$\begin{aligned} I_3 = p_x^3 - 3p_x p_y^2 + 3e^{\sqrt{3}x + y}(p_x - \sqrt{3}p_y) \\ + 3e^{-\sqrt{3}x + y}(p_x + \sqrt{3}p_y) - 6e^{-2y}p_x \end{aligned} \tag{3.9'}$$

we see that in each monomial the factors commute as operators. Thus there is no ordering ambiquity and in every ordering rule the same representative is obtained.

Example 3.2: Take $V = \frac{1}{2}x^2 + \frac{1}{18}y^2$, then the following is a classical (third) invariant[10]

$$A_3 = (xp_y - yp_x)p_y^2 + \frac{y^3}{27}p_x - \frac{xy^2}{3}p_y. \tag{3.10}$$

Now (3.7) vanishes trivially since the potential is only second order in x and y, thus (3.10) is also the Weyl representative of the quantum operator. In other ordering rules the c-number representative would be different. Let us finally construct the actual quantum operator. Using the inverse of (2.11) we get the standard ordering version of A_3, and then we substitute (2.5) to obtain ($\hbar = 1$)

$$\begin{aligned} (-i)\hat{A}_3 = x\partial_y^3 - y\partial_x\partial_y^2 - \partial_x\partial_y \\ - \tfrac{1}{27}y^3\partial_x + \tfrac{1}{3}xy^2\partial_y + \tfrac{1}{3}xy. \end{aligned} \tag{3.10Q}$$

Note the extra terms $-\partial_x\partial_y$ and $\frac{1}{3}xy$. [The term $-\partial_x\partial_y$ was omitted in Ref. 10.]

The equation (3.7) is sufficient, but not always necessary. For although (3.4) determines the classical part of g_i there can also be a quantum part (a,b and c are constants)

$$\Delta_Q g_0 = \hbar^2(ay + b),$$

$$\Delta_Q g_1 = \hbar^2(-ax + c),$$

then [due to (3.5Q)] (3.7) will be replaced by the necessary condition

$$\begin{aligned} (ay + b)\partial_x V + (-ax + c)\partial_y V \\ - \tfrac{1}{4}[f_0\partial_x^3 V + f_1 \partial_x^2\partial_y V + f_2\partial_x\partial_y^2 V + f_3\partial_y^3 V] = 0. \end{aligned} \tag{3.7'}$$

Note however, that if f_i's are constant and V polynomial in x and y, then the generalization from (3.7) to (3.7') is useful only if the degrees of the various terms can be made to match. In particular, for homogeneous potentials (3.7') offers no help so that (3.7) is also necessary.

Example 3.3: The Holt Hamiltonian[8]

$$H = \tfrac{1}{2}p_x^2 + \tfrac{1}{2}p_y^2 + \tfrac{3}{4}x^{4/3} + y^2 x^{-2/3}$$

is classically integrable,

$$I_3 = p_y^3 + \tfrac{3}{2}p_x^2 p_y + \cdots$$

but now Eq. (3.7) does not hold. Thus it is not quantum integrable[6] at least at this order in p. We will return to this model in Sec. IV.

The above examples illustrate three different types of relationships between classical and quantum integrability and the corresponding second invariants (c-number representatives in the quantum case): 1) CI $\Rightarrow$ QI in any ordering rule, 2) CI $\Rightarrow$QI in a specific rule, and 3) CI $\nRightarrow$ QI at the same order. It is easy to imagine other types as well but we are not aware of actual examples (at order p^3) not falling in the categories 1)–3) above.

B. I is of order p^4

Let us next assume that the invariant is of the general form

$$I_4 = f_0 p_x^4 + f_1 p_x^3 p_y + f_2 p_x^2 p_y^2 + f_3 p_x p_y^3 + f_4 p_y^4 + g_0 p_x^2 + g_1 p_x p_y + g_2 p_y^2 + h. \tag{3.11}$$

For CI the equations are now the following[9]:
At order p^5

$$\begin{aligned}
\partial_x f_0 &= 0,\\
\partial_x f_1 + \partial_y f_0 &= 0,\\
\partial_x f_2 + \partial_y f_1 &= 0,\\
\partial_x f_3 + \partial_y f_2 &= 0,\\
\partial_x f_4 + \partial_y f_3 &= 0,\\
\partial_y f_4 &= 0,
\end{aligned} \tag{3.12}$$

at order p^3

$$\begin{aligned}
\partial_x g_0 &= 4f_0 \partial_x V + f_1 \partial_y V,\\
\partial_x g_1 + \partial_y g_0 &= 3f_1 \partial_x V + 2f_2 \partial_y V,\\
\partial_x g_2 + \partial_y g_1 &= 2f_2 \partial_x V + 3f_3 \partial_y V,\\
\partial_y g_2 &= f_3 \partial_x V + 4f_4 \partial_y V,
\end{aligned} \tag{3.13}$$

and finally at order p

$$\begin{aligned}
\partial_x h &= 2g_0 \partial_x V + g_1 \partial_y V,\\
\partial_y h &= g_1 \partial_x V + 2g_2 \partial_y V.
\end{aligned} \tag{3.14}$$

For quantum integrability the Moyal bracket still contributes only through the term (3.6). This extra term will now contribute at order p, therefore, in the quantum case (3.12) and (3.13) will be unchanged but (3.14) is replaced by

$$\partial_x h = 2g_0 \partial_x V + g_1 \partial_y V - \frac{\hbar^2}{4}\left[4f_0 \partial_x^3 V + 3f_1 \partial_x^2 \partial_y V + 2f_2 \partial_x \partial_y^2 V + f_3 \partial_y^3 V\right], \tag{3.14Q}$$

$$\partial_y h = g_1 \partial_x V + 2g_2 \partial_y V - \frac{\hbar^2}{4}\left[f_1 \partial_x^3 V + 2f_2 \partial_x^2 \partial_y V + 3f_3 \partial_x \partial_y^2 V + 4f_4 \partial_y^3 V\right].$$

The first result that can be obtained from (3.14Q) is the following: Assume that the Hamiltonian (3.1) is classically integrable and has the second invariant of type (3.11). If

$$\begin{aligned}
&4f_0 \partial_x^3 V + 3f_1 \partial_x^2 \partial_y V + 2f_2 \partial_x \partial_y^2 V + f_3 \partial_y^3 V = 0,\\
&f_1 \partial_x^3 V + 2f_2 \partial_x^2 \partial_y V + 3f_3 \partial_x \partial_y^2 V + 4f_4 \partial_y^3 V = 0,
\end{aligned} \tag{3.15}$$

then the corresponding quantum Hamiltonian is also integrable and has a second invariant operator $\hat{I}_4$, whose c-number representative in the Weyl rule is identical to the classical invariant.

Example 3.4: The Henon–Heiles type potential

$$V = \tfrac{16}{3} y^3 + x^2 y \tag{3.16}$$

is classically integrable,

$$I_4 = p_x^4 + 4yx^2 p_x^2 - \tfrac{4}{3} x^3 p_x p_y - \tfrac{4}{3} y^2 x^4 - (\tfrac{2}{9}) x^6. \tag{3.17}$$

Since only $f_0 \neq 0$ (3.15) is clearly satisfied and (3.17) is the second invariant in the Weyl rule.

Contrary to the previous p^3 case in Sec. III.A we can now also have quantum integrability with $\hbar^2$-corrections to the second invariant even in homogeneous cases. For even if Eqs. (3.15) do not hold it might be possible to solve for h from (3.14Q), if the $\hbar^2$-terms satisfy the integrability condition

$$\partial_y\left[4f_0 \partial_x^3 V + 3f_1 \partial_x^2 \partial_y V + 2f_2 \partial_x \partial_y^2 V + f_3 \partial_y^3 V\right] = \partial_x\left[f_1 \partial_x^3 V + 2f_2 \partial_x^2 \partial_y V + 3f_3 \partial_x \partial_y^2 V + 4f_4 \partial_y^3 V\right]. \tag{3.18}$$

Let us write

$$h_{\text{Quantum}} = h_{\text{Classical}} + \hbar^2 \Delta_Q h, \tag{3.19}$$

so that the quantum correction to I would be $\Delta_Q I = \hbar^2 \Delta_Q h$. If (3.18) holds $\Delta_Q h$ can be solved from

$$\partial_x \Delta_Q h = -\tfrac{1}{4}\left[4f_0 \partial_x^3 V + 3f_1 \partial_x^2 \partial_y V + 2f_2 \partial_x \partial_y^2 V + f_3 \partial_y^3 V\right], \tag{3.20}$$

$$\partial_y \Delta_Q h = -\tfrac{1}{4}\left[f_1 \partial_x^3 V + 2f_2 \partial_x^2 \partial_y V + 3f_3 \partial_x \partial_y^2 V + 4f_4 \partial_y^3 V\right].$$

Example 3.5: A generalization of (3.16) is[12] [the most general of type $\frac{16}{3}y^3 + yx^2 + f(x) + g(y)$, where $g(y)$ has no third-degree terms, and where translation in y is fixed by the absence of the term linear in y]

$$V = \tfrac{16}{3} y^3 + xy^2 + (a/2)(x^2 + 16y^2) + \mu x^{-2} + \nu x^{-6}, \tag{3.21}$$

$$\begin{aligned}
I_4 = {} & p_x^4 + (2ax^2 + 4x^2 y + 4\mu x^{-2} + 4\nu x^{-6}) p_x^2 - \tfrac{4}{3} x^3 p_x p_y\\
& - \tfrac{4}{3} a x^4 y - \tfrac{4}{3} x^4 y^2 + \tfrac{8}{3} \mu y + 8\nu y x^{-4} - \tfrac{2}{9} x^6 + a^2 x^4\\
& + 4(a\nu + \mu^2) x^{-4} + 8\mu\nu x^{-8} + 4\nu^2 x^{-12}.
\end{aligned} \tag{3.22}$$

Now (3.15) does not hold anymore, but (3.18) does and the additional term can be computed from (3.20). We find

$$\Delta_Q I_4 = -\hbar^2 (2 \cdot 3\mu x^{-4} + 6 \cdot 7\nu x^{-8}). \tag{3.23}$$

In particular if $\mu = \nu = 0$, we find that the c-number representative is identical to the classical invariant in the Weyl rule.

If one finds extra $\hbar^2$ terms esthetically unpleasing there are two ways one can try to eliminate them from the c-number representative of the quantum operator: (i) add αH^2 to I_4 (this trick often works only if the f_i's are constant), (ii) go to another ordering rule.

In classical mechanics we know that if I_4 is an invariant, so is $I_4 + \alpha H^2$. As discussed in Sec. II this does not hold for the c-number representatives of the quantum case due to the higher-order derivatives in the Moyal bracket. Instead, this change will necessitate the introduction of corrective $O(\hbar^2)$-terms. The addition αH^2 corresponds to the following change in the parameters f_i and g_i of I_4:

$$\begin{aligned}
f_0 &\to f_0 + \tfrac{1}{4}\alpha,\\
f_1 &\to f_1,\\
f_2 &\to f_2 + \tfrac{1}{2}\alpha,\\
f_3 &\to f_3,\\
f_4 &\to f_4 + \tfrac{1}{4}\alpha,
\end{aligned} \tag{3.24a}$$

and

$$\begin{aligned}
g_0 &\to g_0 + \alpha V,\\
g_1 &\to g_1,
\end{aligned} \tag{3.24b}$$

$$g_2 \to g_2 + \alpha V.$$

In the classical case (3.14) clearly gives

$$h_c \to h_c + \alpha V^2, \tag{3.25}$$

while in the quantum case we also get a change in the correction term $\Delta_Q h$ of (3.19)

$$\Delta_Q h \to \Delta_Q h - (\alpha/4)[\partial_x^2 V + \partial_y^2 V]. \tag{3.26}$$

Let us next find the effects of changing the ordering rule. It should be remarked again that this does not change the actual operator $\hat{I}_4$, only its c-number representative might change. As discussed before in Sec. II, the ordering rule is defined by a function $F((i/\hbar)x \cdot y)$. We must always have $F(0) = 1$, and now to preserve the $p \to -p$ parity we must also require that all odd derivatives of F vanish at 0. Let us denote $F''(0) = \Phi$ and discuss the effects of $\Phi \neq 0$. Using (2.12) we get

$$\begin{aligned} A_\Phi(p,q) &= \int d^n x\, d^n y\, d^n k\, d^n l\, (2\pi\hbar)^{-2n} \\ &\quad \times \Big(1 - \frac{1}{2}\Big(\frac{i}{\hbar} x \cdot y\Big)^2 \Phi + \cdots \Big) \\ &\quad \times A_W(k,l) \exp \frac{i}{\hbar} [x \cdot (p-k) + y \cdot (q-l)] \\ &= A_W(p,q) + \frac{1}{2}\hbar^2 \Phi \Big(\sum_i \partial_{q_i}\partial_{p_i} \Big)^2 A_W(p,q) + \cdots. \end{aligned} \tag{3.27}$$

For invariants of $O(p^4)$ the higher order terms in (3.27) do not contribute. Applying (3.27) to I_4 we see that the extra operator gives a nonzero contribution only when it operates on the p^2 part of I_4. This means that only $\Delta_Q h$ changes; we find

$$\Delta_Q h \to \Delta_Q h + \Phi(\partial_x^2 g_0 + \partial_x \partial_y g_1 + \partial_y^2 g_2). \tag{3.28}$$

Thus if we want to make $\Delta_Q h$ vanish, we have at our disposal two parameters α and Φ and the corresponding changes (3.26) and (3.28).

Example 3.5 continued: Let us apply the full α, Φ freedom to (3.22). The general form of the invariant has 70 terms and its quantum part is

$$\begin{aligned} \Delta_Q I &= \hbar^2 \{ (6\nu x^{-4} + 42\mu x^{-8})(-1 - \tfrac{1}{4}\alpha + 4\Phi) \\ &\quad + (a + 2y)(-\tfrac{17}{4}\alpha + 4\Phi) \}. \end{aligned} \tag{3.29}$$

The quantum correction vanishes if $\alpha = \frac{1}{4}$ and $\Phi = \frac{17}{64}$, but then the new invariant has 38 terms compared to the 16 in (3.22). So although it is possible to eliminate the $\hbar^2$-dependence from the c-number representation, it is hardly worth the effort.

Let us next take a look at the operator $\hat{I}_4$ corresponding to (3.22). We find

$$\begin{aligned} \hat{I} &= \hbar^4 \partial_x^4 + \hbar^2(-2ax - 4x^2 y - 4\mu x^{-2} - 4\nu x^{-6})\partial_x^2 \\ &\quad + \hbar^2 \tfrac{4}{3} x^3 \partial_x \partial_y \\ &\quad + \hbar^2(-4ax - 8xy + 8\mu x^{-3} + 24\nu x^{-6})\partial_x \\ &\quad + \hbar^2 2x^2 \partial_y \\ &\quad + \hbar^2(84\nu x^{-8} - a - 12\mu x^{-4} - 2y) \\ &\quad - \tfrac{4}{3} a x^4 g - \tfrac{4}{3} x^4 y^2 + \tfrac{8}{3}\mu y + 8\nu x^{-4} y \\ &\quad - \tfrac{2}{9} x^6 + a^2 x^4 + 4(a\nu + \mu^2) x^{-4} + 8\mu\nu x^{-8} + 4\nu^2 x^{-12}. \end{aligned} \tag{3.30}$$

This is in the standard ordering, and the terms linear in ∂_x and ∂_y are quantum effects as are the nonderivative $\hbar^2$-terms. For general α the corresponding expression has 74 terms. Using this remaining freedom one could try to eliminate, e.g., the linear terms by a proper choice of α. For ∂_x we need $\alpha = -4$ but for ∂_y no α will do (even if $\mu = \nu = a = 0$). For $\alpha = -4$, $\hat{I}$ would have 6 more terms compared to (3.30). So there is not much help for ordering rules, and in choosing the amount of α it is best to try to get as short a formula as possible.

Example 3.6: The potential $V = 8y^4 + 6x^2y^2 + x^4$ is integrable,[11] and its second invariant is of type $I_4 = p_x^4 + \cdots$. We have searched for all potentials of type

$$V = 6x^2y^2 + F_1(x) + F_2(y),$$

whose second invariant is of the same type. We found the following two integrable generalizations:

$$\begin{aligned} V_1 &= x^4 + 6x^2y^2 + 8y^4 + \kappa(x^2 + 4y^2) \\ &\quad + \mu x^{-2} + \nu x^{-6} + \lambda y^{-2}, \end{aligned} \tag{3.31}$$

which was found (except for the λ term) by Grammaticos *et al.*[12] Its second invariant is

$$\begin{aligned} I_4^1 &= p_x^4 + 4p_x^2(x^4 + 6x^2y^2 + \kappa x^2 + \mu x^{-2} + \nu x^{-6}) \\ &\quad - 16x^3 y p_x p_y + 4x^4 p_y^2 + 8\lambda x^4 y^{-2} + 4\mu^2 x^{-4} \\ &\quad + 8\mu\nu x^{-8} + 8\mu x^2 + 16\mu y^2 + 4\nu^2 x^{-12} \\ &\quad + 8\nu\kappa x^{-4} + 8\nu x^{-2} + 48\nu x^{-4} y^2 + 4\kappa^2 x^4 + 8\kappa x^6 \\ &\quad + 16\kappa x^4 y^2 + 4x^8 + 16x^6y^2 + 16x^4y^4 \\ &\quad - \hbar^2(6\mu x^{-4} + 42\nu x^{-8} + 12x^2). \end{aligned} \tag{3.32}$$

The other case is

$$V_2 = x^4 + 6x^2y^2 + 8y^4 + \kappa(x^2 + 4y^2) + \mu x^{-2} + \epsilon y, \tag{3.33}$$

where the μ-term generalizes the result of Ref. 13. Now the second invariant is

$$\begin{aligned} I_4^2 &= p_x^4 + 4p_x^2(x^4 + 6x^2y^2 + \kappa x^2 + \mu x^{-2} + \epsilon y) \\ &\quad - (16x^3 y + \epsilon 4x) p_x p_y + 4x^4 p_y^2 \\ &\quad + 4\mu^2 x^{-4} + 8\mu\epsilon x^{-2} y + 8\mu x^2 + 16\mu y^2 + 4\kappa^2 x^2 \\ &\quad - 8\kappa\epsilon x^2 y + 8\kappa x^6 + 16\kappa x^4 y^2 - 2\epsilon^2 x^2 - 8x^4 y\epsilon \\ &\quad - 16\epsilon x^2 y^3 + 4x^8 + 16x^6 y^2 + 16x^4 y^2 \\ &\quad - \hbar^2(6\mu x^{-4} + 12x^2). \end{aligned} \tag{3.34}$$

As in Sec. III.A we can have $\hbar^2$ contributions from the g_i's as well. From (3.13) we get

$$\Delta_Q g_0 = \hbar^2(ay^2 + by + c),$$

$$\Delta_Q q_1 = \hbar^2(-2axy - bx + dy + e),$$

$$\Delta_Q g2 = \hbar^2(ax^2 - dx),$$

and then (3.18) will be replaced by

$$\begin{aligned} &\partial_y \{ 2(ay^2 + by + c)\partial_x V + (-2ax - by + dy + e)\partial_y V \\ &\quad - \tfrac{1}{4}[4f_0 \partial_x^3 V + 3f_1 \partial_x^2 \partial_y V + 2f_2 \partial_x \partial_y^2 V + f_3 \partial_y^3 V] \} \\ &= \partial_x \{ (-2ax - by + dy + e)\partial_x V \\ &\quad + 2(ax^2 - dx)\partial_y V - \tfrac{1}{4}[f_1 \partial_x^3 V + 2f_2 \partial_x^2 \partial_y V \\ &\quad + 3f_3 \partial_x \partial_y^2 V + 4f_4 \partial_y^3 V] \}. \end{aligned} \tag{3.18'}$$

This is the necessary condition. Again the degrees do not match if V is homogeneous and the f_i's are constant.

Example 3.7: The Holt-type potential

$$V = \tfrac{9}{2}x^{4/3} + y^2 x^{-2/3}$$

is also classically integrable[12,14]

$$I_4 = p_y^4 + 2p_y^2 p_x^2 + \cdots$$

but as in Example 3.3 this potential is not quantum integrable.

For those classically integrable potentials that have an invariant four-order in p we have in the examples shown three different types of behavior: 1) CI $\Rightarrow$ QI in the Weyl rule (Example 3.5), 2) CI $\Rightarrow$ QI with additional $O(\hbar^2)$ terms in the invariant (Example 3.6, these corrections may vanish when one adds a suitable amount of H^2 and goes to a specific ordering rule), and 3) CI $\not\Rightarrow$ QI (Example 3.7). We also found that those extra terms in the potential that are allowed by classical integrability are also possible for the quantum case.

So far we have only considered particular examples of integrable potentials. One can also consider the problem of finding all integrable potentials having a second invariant of given order in p. For classical systems, Holt has derived from (3.3)–(3.5) a pair of PDE's for the potential [Ref. 8, Eqs. (170) and (171)]. For quantum integrability at the same order p^3 one should just add (3.7′) to the system of equations that are to be solved. If the potential is homogeneous, then (3.7) is solvable and then also the Holt equations might be solved. Similar results apply to higher-order invariants.

IV. DEFORMATIONS OF THE POTENTIAL

Examples of classically integrable potentials which do not satisfy the extra quantum integrability conditions were provided by the Holt Hamiltonian in Examples 3.3 and 3.7. In this section we will discuss the possibility that a classically integrable potential can be made quantum integrable when the potential is deformed by a term of $O(\hbar^2)$.

Let us write

$$H = \tfrac{1}{2}p_x^2 + \tfrac{1}{2}p_y^2 + V(x,y) + \hbar^2 \Delta(x,y). \tag{4.1}$$

We will now try to construct the second invariant order-by-order in $\hbar^2$. If the classical invariant is of order p^3, it turns out to be sufficient to write

$$I_3^Q = I_3^C + \hbar^2(p_x A(x,y) + p_y B(x,y)). \tag{4.2}$$

Now at order p^4 we get the old equation (3.3), at order $p^2\hbar^0$ (3.4) and at $p^2\hbar^2$ we get (3.4) for the extra terms, i.e.,

$$\begin{aligned} &\partial_x A = 3f_0 \partial_x \Delta + f_1 \partial_y \Delta, \\ &\partial_y A + \partial_x B = 2f_1 \partial_x \Delta + 2f_2 \partial_y \Delta, \\ &\partial_y B = f_2 \partial_x \Delta + 3f_3 \partial_y \Delta. \end{aligned} \tag{4.3}$$

Here f_i's are determined from classical integrability, and are not free. Finally at order p^0, we have (3.5) and the following new equations: at $O(\hbar^2)$

$$\begin{aligned} &A\partial_x V + B\partial_y V + g_0 \partial_x \Delta + g_1 \partial_y \Delta \\ &\quad - \tfrac{1}{4}[f_0 \partial_x^3 V + f_1 \partial_x^2 \partial_y V + f_2 \partial_x \partial_y^2 V + f_3 \partial_y^3 V] = 0, \end{aligned} \tag{4.4}$$

and at $O(\hbar^4)$

$$\begin{aligned} &A\partial_x \Delta + B\partial_y \Delta - \tfrac{1}{4}[f_0 \partial_x^3 \Delta + f_1 \partial_x^2 \partial_y \Delta \\ &\quad + f_2 \partial_x \partial_y^2 \Delta + f_3 \partial_y^3 \Delta] = 0. \end{aligned} \tag{4.5}$$

Equations (4.3) and (4.5) together just mean that the additional potential Δ should itself be quantum integrable for the given f_i's. Finally (4.4) is the critical equation that connects Δ and V.

Let us assume that the potential is homogeneous, i.e., $V(\lambda x, \lambda y) = \lambda^\alpha V(x,y)$, and denote this scaling dimension by $[V] = \alpha$. Then we know[13] that f_i's and g_i's have fixed dimension; if $[f_i] = n_c$, then $[g_i] = n_c + \alpha$. Assume further that also the extra term Δ is homogeneous, $[\Delta] = \beta$, then $[A] = [B] = n_c + \beta$ from (4.3). Now (3.5) introduces no constraints to the dimension but (4.4) and (4.5) do, we find that both are satisfied if $\beta = -2$. Thus at least for homogeneous potentials it is reasonable to write

$$\Delta(x,y) = x^{-2} W(x/y). \tag{4.6}$$

Example 4.1: The Holt Hamiltonian[8]

$$H = \tfrac{1}{2}p_x^2 + \tfrac{1}{2}p_y^2 + \tfrac{3}{4}x^{4/3} + y^2 x^{-2/3} + \delta x^{-2/3} \tag{4.7}$$

is classically integrable,

$$\begin{aligned} I_3 = p_y^3 + \tfrac{3}{2}p_y p_x^2 + (&-\tfrac{9}{2}x^{4/3} + 3x^{-2/3}y^2 \\ &+ 3\delta x^{-2/3})p_y + 9x^{1/3} y p_x, \end{aligned} \tag{4.8}$$

but not quantum integrable.[6] The square bracket in (4.4) gives rise to a term of type $yx^{-8/3}$, and a moments reflection on (4.3) and (4.4) then suggests that it is reasonable to try $\Delta = \text{const} \cdot x^{-2}$. Then for $A = 0$, $B \propto \Delta$ (4.5) is trivially satisfied, and (4.3), (4.4) determine the constants. We find[14]

$$\Delta_Q H \equiv \hbar^2 \Delta(x,y) = -\tfrac{5}{72}\hbar^2 x^{-2}, \tag{4.9}$$

$$\Delta_Q I = -\tfrac{5}{24}\hbar^2 x^{-2} p_y. \tag{4.10}$$

The above potential (4.7) is also the most general of type $y^2 x^{-2/3} + F_1(x) + F_2(y)$ having the leading part in I_3 as given in (4.8). The classically allowed extra terms have no effect on quantum integrability.

In Example (4.1) above we have witnessed a new kind of relationship between classical and quantum integrability: It was necessary to deform the potential with an $O(\hbar^2)$ term. This feature seems to be generic to the Holt-type Hamiltonians, and possibly others with rational powers in the potential.

Above we carried out the analysis when the second invariant was of order p^3. The computations can be generalized straightforwardly to situations where the Hamiltonian has a second invariant of order p^4. It suffices to say that again the deformation (4.6) is suggested.

Example 4.2: The Holt-type potential of Example 3.7 is not quantum integrable without deformations, but it turns out that precisely the same deformation (4.9) works.[14] The potential can be further generalized to

$$\begin{aligned} V = \tfrac{9}{2}x^{4/3} + y^2 x^{-2/3} + \delta x^{-2/3} + \mu x^{2/3} + \lambda y^{-2} \\ + a(9x^2 + 4y^2) - \tfrac{5}{72}\hbar^2 x^{-2} \end{aligned} \tag{4.11}$$

(classically this was obtained in Ref. 12 except for the a term). The second invariant is

$$\begin{aligned} I_4 = p_y^4 &+ 2p_x^2 p_y^2 + p_x^2(16ay^2 + 4\lambda y^{-2}) + 24x^{1/3} y p_x p_y \\ &+ 4p_y^2(x^{-2/3}y^2 - \tfrac{5}{72}\hbar^2 x^{-2} + \delta x^{-2/3} \\ &+ \mu x^{2/3} + a(9x^2 + 4y^2) + \lambda y^{-2}) \\ &+ 16\mu y^2 + 32\delta a x^{-2/3} y^2 + 8x^{-2/3} y^{-2} \nu\lambda \\ &+ 8\lambda x^{-2/3} + 32\mu a x^{2/3} y^2 + 8\mu\lambda x^{2/3} y^{-2} + 72x^{2/3} y^2 \\ &+ 72a\lambda x^2 y^{-2} + 4\lambda^2 y^{-4} + 32a x^{-2/3} y^2 (9x^2 + y^2) \end{aligned}$$

$$+ 32a^2y^2(9x^2 + 2y^2) - \hbar^2(2x^{-2/3} + \tfrac{20}{9}ax^{-2}y^2 + \tfrac{5}{9}\lambda x^{-2}y^{-2} - 6\lambda y^{-4}). \tag{4.12}$$

So far in this paper we have not discussed systems where the second invariant is of order higher than 4 in p. In such cases it is possible that the third term in the expansion of the Moyal bracket (2.16) also contributes. However, sometimes the first two terms are sufficient, as in the case below.

Example 4.3: The Holt family of potentials has also the following member[14,12]

$$V = 12x^{4/3} + y^2x^{-2/3}, \tag{4.13}$$

and then the second invariant can be chosen to start

$$I_6 = p_y^6 + 3p_y^4p_x^2 + O(p^4). \tag{4.14}$$

The form (4.14) shows that in the third term of the Moyal bracket only the derivatives $\partial^5_{p^I_y}\,\partial^5_{y^H}$, $\partial^4_{p^I_y}\,\partial_{p^I_x}\,\partial^4_{y^H}\,\partial_{x^H}$, and $\partial^3_{p^I_y}\,\partial^2_{p^I_x}\,\partial^3_{y^H}\,\partial^2_{x^H}$ could be nonzero, but they all annihilate the potential. As in the previous examples we have also searched for potentials of type $V = y^2x^{-2/3} + F_1(x) + F_2(y)$. This time we assume [in addition to (4.14)] that F_2 is a polynomial of degree less than 5 in y so that it is annihilated by the derivatives above. The result is

$$V = 12x^{4/3} + y^2x^{-2/3} + \delta x^{-2/3} - \tfrac{5}{72}\hbar^2x^{-2} \tag{4.15}$$

and

$$\begin{aligned} I_6 = {} & p_y^6 + 3p_x^2p_y^4 + 72x^{1/3}yp_xp_y^3 \\ & + 6p_y^4(\delta x^{-2/3} + 3x^{4/3} + x^{-2/3}y^2 - \tfrac{5}{72}\hbar^2x^{-2}) \\ & + 18p_y^2(32y^2x^{2/3} - x^{-2/3}\hbar^2) + 648y^4. \end{aligned} \tag{4.16}$$

This potential also has the generalization

$$V = 12x^{4/3} + y^2x^{-2/3} + \delta x^{-2/3} + \lambda y^{-2} - \tfrac{5}{72}\hbar^2x^{-2}, \tag{4.15'}$$

but now the next term in the Moyal bracket (2.16) contributes. The invariant for (4.15') has 34 terms.

In the previous examples the deformation const $\cdot\, x^{-2}$ was sufficient. A slightly more complicated deformation is needed in the example below.

Example 4.4: The Hamiltonian

$$H = \tfrac{1}{2}p_x^2 + \tfrac{1}{2}p_y^2 + (x^2 - y^2)^{-2/3}$$

was found classically integrable by Fokas and Lagerstrom,[15]

$$I = (p_x^2 - p_y^2)(xp_y - yp_x) - 4(yp_x + xp_y)(x^2 - y^2)^{-2/3}.$$

It turns out that it is simplest to study this system in the rotated form[12]

$$H = \tfrac{1}{2}p_x^2 + \tfrac{1}{2}p_y^2 + (xy)^{-2/3}, \tag{4.17}$$

$$I = p_xp_y(xp_y - yp_x) + 2(xy)^{-2/3}(p_xx - p_yy). \tag{4.18}$$

In this case the quantum deformations are[16]

$$\Delta_Q H = -\tfrac{5}{72}\hbar^2(1/x^2 + 1/y^2), \tag{4.19}$$

$$\Delta_Q I = -\tfrac{5}{36}\hbar^2(p_xx/y^2 - p_yy/x^2). \tag{4.20}$$

The corresponding quantum operator $\widehat{I}$ is given by ($\hbar = 1$)

$$\begin{aligned} (-i)\widehat{I} = {} & (x\partial_y - y\partial_x)\partial_x\partial_y + \tfrac{1}{2}(-\partial_x^2 + \partial_y^2) \\ & + (\tfrac{5}{36}xy^{-2} - 2y^{-2/3}x^{1/3})\partial x \\ & + (-\tfrac{5}{36}x^{-2}y + 2y^{1/3}x^{-2/3})\partial y \\ & + \tfrac{5}{72}(-x^{-2} + y^{-2}). \end{aligned} \tag{4.18Q}$$

(An incorrect operator was given in Ref. 10.) The potential cannot be further generalized in the form $(xy)^{-2/3} + F_1(x) + F_2(y)$ with the leading part given in (4.18).

In the previous section we had two ways to change the $\hbar^2$ dependence of the second invariant. Of course neither of these works if we try to eliminate the $\hbar^2$-deformations in the Hamiltonian. The common constants and the appearance of these deformations with rational powers in the potential suggest that canonical transformations (CT) might be helpful. Now the discussion of CT in quantum mechanics is very tricky. On one hand, integrability itself suggests that a CT to action-angle variables should be possible, on the other hand the construction of even simple CT's mixing p_x and x turns out to be very involved.[17] To be sure CT's can be dismissed by just noting that they are unitary transformations,[18] but the construction of the needed unitary transformation is another matter. In this paper we will, therefore, just consider point transformations, for which the situation is under control.

It is well known that a canonical point transformation

$$\begin{aligned} x &= F(X), \\ p_x &= P_X/F'(X), \end{aligned} \tag{4.21}$$

can give raise to $O(\hbar^2)$ terms to the potential.[5,19] For example the Hamiltonian $\frac{1}{2}p_x^2$ will transform to[5]

$$H' = \tfrac{1}{2}P_X^2F'(X)^{-2} + \tfrac{1}{8}\hbar^2F''(X)^2F'(X)^{-4}. \tag{4.22}$$

Let us now apply this to the Hamiltonian (4.7) with (4.9). If we make the canonical transformation (4.21), where now $F(x) = X^\alpha$, y, p_y unchanged) the $\hbar^2$-term in (4.22) will combine with the original x^{-2} term in (4.9). We find the transformed Hamiltonian to be

$$\begin{aligned} \widetilde{H} = {} & \alpha^{-2}\tfrac{1}{2}P_X^2X^{2(1-\alpha)} + \tfrac{1}{2}P_Y^2 + \tfrac{3}{4}X^{(4/3)\alpha} \\ & + Y^2X^{-(2/3)\alpha} + \delta X^{-(2/3)\alpha} \\ & + \hbar^2X^{-2\alpha}\{\tfrac{1}{8}((\alpha - 1)/\alpha)^2 - \tfrac{5}{72}\}. \end{aligned} \tag{4.23}$$

Now it would be natural to choose α so that only integer powers appear in $\widetilde{H}$, e.g., $\alpha = 3$, or $3/2$. However, such a choice does not fully eliminate the extra term.

The new Hamiltonian (4.23) is ordering dependent, so let us see how it looks in other ordering rules. Applying (3.27) to (4.23) gives the following $O(\hbar^2)$ part:

$$\hbar^2X^{-2\alpha}\{2\alpha^{-2}(1 - \alpha)(1 - 2\alpha)\Phi + \tfrac{1}{8}((\alpha - 1)/\alpha)^2 - \tfrac{5}{72}\}. \tag{4.24}$$

For example if $\alpha = \frac{3}{2}$, the extra term vanishes when $\Phi = \frac{1}{16}$, which is not a common ordering.

The use of point transformations and ordering rules does eliminate the extra $\hbar^2$-term from the Hamiltonian (or rather from its c-number representative), but this can hardly be taken as an explanation of the observed phenomenon. We leave the question open, and here just underline the fact that the deformations known so far are all associated with rational powers in the potential and are all of the same type (including numerical factors) $-\frac{5}{72}\hbar^2x^{-2}$.

V. DISCUSSION

In this paper we have studied the connection between quantum integrability and classical integrability. For quantum operators we used c-number representatives in the Weyl rule. For the two degree of freedom systems that were studied in this paper, we found the following types of relationships between the classical invariant and the c-number representative of the quantum operator: (i) they are identical in all ordering rules, (ii) $O(\hbar^2)$ corrections are needed for the quantum case, and (iii) also the potential needs an $O(\hbar^2)$ deformation.

Many n-dimensional systems studied before[20] fall into the category (i) above. These systems are associated with a Lax pair, L, M of classical matrices, where $iL = [L, M]$. The classical invariants are obtained as coefficients of the characteristic polynomial $\det(L - \lambda I)$. If now L has a proper form it turns out that the invariant depends on p_i and x_j in such a way that there is not ordering ambiquity. For quantum integrability one then needs to show only that for these systems the extra contributions from the Moyal bracket also vanish. Our example 3.1 is of this type.

In this paper we also found examples for types (ii) and (iii) (see also Refs. 13, 14, 16). For the Henon–Heiles and Holt type potentials we found the most general additive terms of form $F_1(x) + F_2(y)$. Some results are new even for the classical case. The classically allowed terms were also possible for quantum integrability.

Finally we studied various methods that could be used to eliminate the $O(\hbar^2)$-corrections or -deformations. It was found that there is no unique ordering rule which would be better or easier than the Weyl rule.

The $O(\hbar^2)$-deformations needed for the Holt and Fokas–Lagerstrom potentials were always of type $-\frac{5}{72}\hbar^2 x^{-2}$. The common numerical factors and association with fractional powers in the potential are intriguing and suggest to a common, perhaps geometric, origin.

ACKNOWLEDGMENTS

I would like to thank B. Grammaticos for useful remarks and B. Dorizzi for checking some computations.

[1] S. Agarwal and E. Wolf, Phys. Rev. D **2**, 2161 (1970).

[2] J. Shewell, Am. J. Phys. **27**, 16 (1959).

[3] L. Cohen, J. Math. Phys. **7**, 781 (1966); **17**, 597 (1976).

[4] J. Moyal, Proc. Cambridge Philos. Soc. **45**, 99 (1949).

[5] J. Hietarinta, Phys. Rev. D **25**, 2103 (1982).

[6] J. Hietarinta, Phys. Lett. A **93**, 55 (1982).

[7] A. Makarov, J. Smorodinsky, K. Valiev, and P. Winternitz, Nuovo Cimento A **52**, 1061 (1967).

[8] C. R. Holt, J. Math. Phys. **23**, 1037 (1982).

[9] B. Grammaticos, B. Dorizzi, and A. Ramani, J. Math. Phys. **24**, 2289 (1983).

[10] A. S. Fokas and P. A. Lagerstrom, J. Math. Anal. Appl. **74**, 359 (1980).

[11] A. Ramani, B. Dorizzi, and B. Grammaticos, Phys. Rev. Lett. **49**, 1539 (1982).

[12] B. Grammaticos, B. Dorizzi, and A. Ramani, CNET preprint A553.0583.

[13] J. Hietarinta, Phys. Lett. A **96** 273 (1983).

[14] J. Hietarinta, Phys. Rev. A **28**, 3670 (1983).

[15] A. S. Fokas and P. A. Lagerstrom, J. Math. Anal. Appl. **74**, 325 (1980).

[16] J. Hietarinta, B. Dorizzi, B. Grammaticos, and A. Ramani, C. R. Acad. Sci. Paris, Ser. II **297**, 791 (1983).

[17] P. A. Mello and M. Moshinsky, J. Math. Phys. **16**, 2017 (1975); P. Kramer, M. Moshinsky, and T. H. Seligman, J. Math. Phys. **19**, 683 (1978).

[18] P. A. M. Dirac, *Principles of Quantum Mechanics*, 4th ed. (Oxford University, London, 1958).

[19] J.-L. Gervais and A. Jevicki, Nucl. Phys. B **110**, 93 (1976).

[20] F. Calogero, O. Ragnisco, and C. Marchioro, Lett. Nuovo Cimento **13**, 383 (1975); P. O. Kulish, Theor. Math. Physics **26**, 132 (1976); A. M. Perelomov, Fiz. Elem. Chastits At. Yadra, **10**, 850 (1979). [Sov. J. Part. Nucl. **10**, 336 (1979)].

N-body quantum scattering theory in two Hilbert spaces. IV. Approximate equations

Colston Chandler
Department of Physics and Astronomy, University of New Mexico, Albuquerque, New Mexico 87131

A. G. Gibson
Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131

(Received 27 September 1983; accepted for publication 6 January 1984)

A rigorous mathematical theory of approximations is developed for the time-independent transition operators of N-body multichannel nonrelativistic quantum scattering theory. New basic dynamical equations are derived and shown to specify uniquely the approximate time-independent transition operators. These operator equations represent coupled integral equations with compact kernels, but it is not assumed that the equations that determine the exact transition amplitudes have compact kernels. Convergence of sequences of these approximate time-independent transition operators to the exact transition operator is established in appropriate limits. Stability of the basic dynamical equations is proved. Resolvent-type equations and their relation to the limiting absorption principle are investigated. The relation of this theory to the Petryshyn theory of A-proper operators and to the Feshbach unified theory of nuclear reactions is discussed.

PACS numbers: 03.65.Nk, 24.10. − i

I. INTRODUCTION

In Paper I of this series[1] we established a two-Hilbert-space formulation of nonrelativistic quantum mechanical scattering theory. The prior form $T(z)$ of the time-independent transition operator was shown to be, for Im $z \neq 0$, the unique solution of a linear equation of the Lippmann–Schwinger type [cf. Eq. (2.10) below]. The practical solution of this equation is the primary theoretical goal.

As we pointed out at the time, there were two difficulties with our basic dynamical equation.

(1) The inverse of a certain operator JJ^* appears in the kernel, considerably complicating calculations.

(2) Neither the kernel of the equation nor any of its iterates was compact. There was, consequently, no hope of applying Fredholm theory to obtain the solution of the equation.

The main purpose of this paper is to overcome these difficulties.

We circumvent the first difficulty by introducing equations (cf. Theorems 3.5 and 5.4) that do not contain the inverse of JJ^* in the kernel.

We could have circumvented the second difficulty by deriving another equation for $T(z)$ that does have a connected kernel. This has, in fact, been done.[2,3]

In our opinion, however, the practicality of connected kernel formulations of N-particle scattering theory (cf. the review papers[4,5] and references cited therein) is not yet firmly established for systems of more than four or five particles. The number of coupled integral equations represented by the operator equation for $T(z)$ increases very rapidly with particle number.[6,7] Accompanying this is a prodigious increase in mathematical complexity that requires truly heroic efforts to establish even partial control over the mathematics of the kernel.[8–12] In addition, the connection between the kernel and measurable quantities becomes much less transparent as the number of particles increases. Physical intuition is thus no longer an obviously reliable guide for choosing the good finite-rank approximations to the kernels that are the basis of all practical calculations.

In view of these comments we have elected, in this paper, to deal with the second difficulty of our equation by developing an alternative solution strategy that does not require compact kernels. The idea is to shift the emphasis from approximating the kernel to approximating the transition operator itself. The feasibility of such an approach is supported by proofs[13,14] that for N-particle systems with short-range interactions the prior form of the on-shell transition operator is compact if the initial state has only two fragments and the energy is restricted to a finite interval. A similar result for the off-shell transition operator $T(z)$ can be inferred from Lemma 3.4 of Ref. 15. In this case $T(z)$ itself admits, in principle, an accurate finite-rank approximation, even though the kernel of the dynamical equation may not. In this way the original equation can be well approximated by finite-dimensional equations without the necessity that the kernel be connected. Another important feature of our method is that a sequence of approximate transition operators $T^{(n)}(z)$ converges to the exact transition operator $T(z)$. These convergence results are of two types. First, a sequence of asymptotically complete approximate wave operators $\Omega^{(n)\pm}$ converges strongly to the exact wave operators $\Omega^{\pm}$ and the corresponding sequence of approximate scattering operators $S^{(n)}$ converges weakly to the exact scattering operator S. These results, which are proved in Ref. 16, can be interpreted as convergence of $T^{(n)}(z)$ to $T(z)$ on the energy shell. Second, for Im $z \neq 0$, the sequence of operators $T^{(n)}(z)$ converges strongly to $T(z)$ (cf. Theorems 4.1 and 4.7). This is convergence off the energy shell.

These convergence results, which we believe to be the most general currently available for general N-particle systems, do not establish a rate of convergence. There is no practical estimate of the accuracy of any given approximation. As with any other method, therefore, construction (via the method described in Sec. III) of good approximate spaces $\mathscr{H}_\pi$ and approximate Hamiltonians H_π is a matter of skill. Accuracy is to be judged by comparison of the results with data.

The theory presented in this paper is therefore to be construed as providing a rigorous general framework within which to construct approximations.

In Sec. II we recall our previous definition (cf. Assumption A) of an exact scattering system[16] and exact transition operators.[1,17,18]

In Sec. III we review the approximation method which was introduced in Ref. 16. The key idea is that the space $\mathscr{H}$ of asymptotic states of the exact theory is approximated by a subspace $\mathscr{H}^\pi \equiv \Pi\mathscr{H}$, where Π is an orthogonal projection operator. In addition, the exact asymptotic Hamiltonian H is replaced by an approximate Hamiltonian H^π. The necessary abstract properties of Π and H^π are given in Assumption Π. Our approximate time-independent theory is then considered. A new basic dynamical equation is proposed (Theorems 3.5 and 3.7), the solutions $M^\pi(z)$ of which can be used to construct the approximate transition operators $T^\pi(z)$. The kernel of the equation for $M^\pi(z)$ is compact for all z with $\operatorname{Im} z \neq 0$ (Theorem 3.12), and the operator $(JJ^*)^{-1}$ is not present in the equation. Finally, stability of our basic dynamical equation under certain perturbations is proved (Theorems 3.14 and 3.16).

In Sec. IV sequences of approximations are considered. The defining properties of the sequences are given in Assumption $\Pi^{(n)}$. The previously mentioned convergence of a sequence of approximate transition operators $T^{(n)}(z)$ to the exact transition operator $T(z)$ is then established. Another stability result (Theorem 4.9) is then given which establishes the uniform stability of the solutions of a sequence of our approximate equations.

In Sec. V resolvent-type equations are taken up, and a formula of the type encountered in the limiting absorption principle[8,12,19–22] is established (Theorem 5.2). These equations are not directly relevant to our main goal of laying a theoretical foundation for the approximation of the transition operator $T(z)$, and hence the scattering operator S. They are included here because they provide a theoretical starting point for certain analytical investigations of asymptotic completeness and spectral properties of the total Hamiltonian. In particular, some results of this type have already been obtained by Trucano[22] for certain approximation models. An exact analog of the equation for $M^\pi(z)$ and a differential equation form for the equations are also introduced. Some additional convergence results are also presented.

In Sec. VI we present a discussion of the most important features of our approximation method. We also comment on the relation between our method and the unified nuclear reaction theory of Feshbach.[23–25]

There are two appendices. Appendix A shows how our approximation strategy is related to the Petryshyn theory[26,27] of strong approximation solvability and A-proper operators. Appendix B gives an index of the major operators and Hilbert spaces used in the paper.

II. EXACT SCATTERING SYSTEM

A. Abstract assumption A

In the two-Hilbert-space formulation of nonrelativistic N-body quantum scattering theory (cf. Refs. 1, 2, 12, 16–18, 28–31) an *exact scattering system* is characterized by the sextuple

$$\mathfrak{S} \equiv \{\mathscr{H}_N, H_N, \mathscr{H}, H, J, I^a\}. \tag{2.1}$$

Here H_N is the *total Hamiltonian* (a self-adjoint operator) acting on the Hilbert space $\mathscr{H}_N$. H is the *asymptotic Hamiltonian* (a self-adjoint operator) acting on the Hilbert space $\mathscr{H}$. J is a bounded *identification operator* from $\mathscr{H}$ to $\mathscr{H}_N$. $\mathscr{H}$ has a decomposition $\mathscr{H} = \mathscr{H}^a \oplus \mathscr{H}^b$ that reduces H, and I^a is the orthogonal projection of $\mathscr{H}$ onto $\mathscr{H}^a$. The time-dependent *wave operators* $\Omega^\pm: \mathscr{H} \to \mathscr{H}_N$ are defined by

$$\Omega^\pm \equiv \Omega^\pm(H_N, H, JI^a) \equiv \operatorname*{s-lim}_{t \to \pm\infty} e^{iH_N t} J I^a e^{-iHt}, \tag{2.2}$$

and their adjoints are denoted by $\Omega^{\pm *}$. The *scattering operator* $S: \mathscr{H} \to \mathscr{H}$ is

$$S \equiv \Omega^{+*}\Omega^-. \tag{2.3}$$

We refer to the system characterized by the sextuple $\mathfrak{S}$ as the *exact scattering system*.

Our fundamental assumptions about the exact scattering system $\mathfrak{S}$ are contained in the following Assumption A.

Assumption A: The exact scattering system $\mathfrak{S} = \{\mathscr{H}_N, H_N, \mathscr{H}, H, J, I^a\}$ is said to satisfy Assumption A if the following five statements are true.[16]

(A1) $\mathscr{H}_N$ is a separable Hilbert space, and H_N: $\mathscr{D}(H_N) \subset \mathscr{H}_N \to \mathscr{H}_N$ is a self-adjoint operator that is bounded from below.

(A2) $\mathscr{H}$ is a separable Hilbert space, and H: $\mathscr{D}(H) \subset \mathscr{H} \to \mathscr{H}$ is a self-adjoint operator that is bounded from below. Its spectral family is denoted by $E(\lambda)$. There is a decomposition $\mathscr{H} = \mathscr{H}^a \oplus \mathscr{H}^b$ that reduces H. I^a is the orthogonal projection of $\mathscr{H}$ onto $\mathscr{H}^a$. The restriction H^a of H to $\mathscr{H}^a$ has only absolutely continuous spectrum consisting of a half-line.

(A3) J: $\mathscr{H} \to \mathscr{H}_N$ is a bounded linear operator. J maps $\mathscr{D}(H)$ into $\mathscr{D}(H_N)$, and J^*, the adjoint of J, maps $\mathscr{D}(H_N)$ into $\mathscr{D}(H)$. The operator JJ^*: $\mathscr{H}_N \to \mathscr{H}_N$ has a bounded inverse.

(A4) The operators V: $\mathscr{D}(V) \subset \mathscr{H} \to \mathscr{H}_N$ and V^*: $\mathscr{D}(V^*) \subset \mathscr{H}_N \to \mathscr{H}$,

$$V \equiv H_N J - JH \quad \text{and} \quad V^* \equiv J^* H_N - HJ^*, \tag{2.4}$$

satisfy $V \lessdot H$ and $V^* \lessdot H_N$, respectively, where $K \lessdot L$ means that K is infinitesimally small with respect to L.[16,31]

(A5) The wave operators $\Omega^\pm: \mathscr{H} \to \mathscr{H}_N$ defined by Eq. (2.2) exist and are partially isometric, i.e.,

$$\Omega^{\pm *}\Omega^\pm = I^a. \tag{2.5}$$

B. Exact transition operators

Resolvent operators defined on $\mathscr{H}_N$ and $\mathscr{H}$, respectively, are defined by

$$R_N = R_N(z) \equiv (z - H_N)^{-1} \quad \text{and} \quad R = R(z) \equiv (z - H)^{-1}. \tag{2.6}$$

The time-independent *transition operator* $T(z)$: $\mathscr{D}(H) \rightarrow \mathscr{H}$ is defined by

$$\begin{aligned} T(z) &\equiv (z - H)J^* R_N(z) J(z - H) - (z - H)J^*J \\ &= (z - H)J^* R_N(z) V, \end{aligned} \tag{2.7}$$

where $z \in \rho(H_N)$ (the resolvent set of H_N). The relation between $T(z)$ and the scattering operator S defined in Eq. (2.3) is given by the following theorem.

Theorem 2.1: Let $\mathfrak{S}$ satisfy Assumptions (A1)–(A3) and (A5), and let $\delta_\epsilon(x)$, x real, denote the function

$$\delta_\epsilon(x) \equiv (\epsilon/\pi)(x^2 + \epsilon^2)^{-1}. \tag{2.8}$$

Then, $S - I^a$ has on $\mathscr{H}$ the representation

$$\begin{aligned} S - I^a = \operatorname*{w-lim}_{\epsilon \rightarrow 0^+} (-2\pi i) \int_\lambda \int_\mu dE(\lambda) \delta_\epsilon(\lambda - \mu) \\ \times I^a T([\lambda + \mu + i\epsilon]/2) I^a \, dE(\mu), \end{aligned} \tag{2.9}$$

where the integral is a repeated spectral integral that may be evaluated in either order of integration.

Proof: Since I^a commutes with $E(\cdot)$ the result follows immediately from Theorem 4 of Ref. 17 and Theorem 3 of Ref. 18. The operator $I^a T(z) I^a$ in Eq. (2.9) is to be identified with the prior transition operator $T^{(-)}$ of Ref. 18. □

Numerical calculations of scattering probabilities are often based on dynamical equations for the operator $T(z)$. One such equation is given in the following theorem.

Theorem 2.2: Let $\mathfrak{S}$ satisfy Assumptions (A1)–(A4). Then the operator $T(z)$ is the unique solution defined on $\mathscr{D}(H)$ of the equation

$$T(z) = J^*V + V^*(JJ^*)^{-1} JR(z) T(z), \tag{2.10}$$

where $z \in \rho(H_N) \cap \rho(H)$.

Proof: See Ref. 1, Theorems 4 and 6. □

III. APPROXIMATE SCATTERING SYSTEMS

A. Assumption Π

In this paper we develop a time-independent approximation theory for nonrelativistic N-body quantum scattering in which an approximation to the exact scattering system $\mathfrak{S}$ is characterized by a sextuple[16]

$$\mathfrak{S}(\Pi, H^\pi) \equiv \mathfrak{S}\{\mathscr{H}_\pi, H_\pi, \mathscr{H}^\pi, H^\pi, J^\pi, \Pi^a\}. \tag{3.1}$$

$\mathscr{H}^\pi$ is a Hilbert space defined by $\mathscr{H}^\pi \equiv \Pi \mathscr{H}$, where

$$\Pi = \Pi^a \oplus \Pi^b: \; \mathscr{H} = \mathscr{H}^a \oplus \mathscr{H}^b \rightarrow \mathscr{H}^\pi = \mathscr{H}^{\pi a} \oplus \mathscr{H}^{\pi b} \tag{3.2}$$

is an orthogonal projection operator that commutes with I^a. The symbol Π^a (Π^b) will be used to denote the orthogonal projection of either $\mathscr{H}$, $\mathscr{H}^\pi$, or $\mathscr{H}^a$ ($\mathscr{H}^b$) onto $\mathscr{H}^{\pi a}$ ($\mathscr{H}^{\pi b}$). H^π is an *approximate asymptotic Hamiltonian* (a self-adjoint operator) acting on $\mathscr{H}^\pi$ that is constrained to be equal to H on $\mathscr{H}^{\pi a}$. J^π is the bounded *approximate identification operator* defined by

$$J^\pi \equiv J\Pi, \tag{3.3}$$

and its adjoint is denoted by $J^{\pi *}$. $\mathscr{H}_\pi$ is the closure of the range of J^π,

$$\mathscr{H}_\pi \equiv \overline{\mathscr{R}}(J^\pi). \tag{3.4}$$

H_π is the *approximate total Hamiltonian* defined by

$$H_\pi \equiv P_\pi H_N P_\pi, \tag{3.5}$$

where P_π is the orthogonal projection of $\mathscr{H}_N$ onto $\mathscr{H}_\pi$. We call the system represented by $\mathfrak{S}(\Pi, H^\pi)$ an *approximate scattering system*.

Our fundamental assumptions about the approximate scattering systems $\mathfrak{S}(\Pi, H^\pi)$ are contained in the following Assumption Π.

Assumption Π: The approximate scattering system $\mathfrak{S}(\Pi, H^\pi)$ is said to satisfy Assumption Π if the following statements are true.[16]

(Π0) The exact scattering system $\mathfrak{S}$ satisfies Assumptions (A1)–(A4).

(Π1) The orthogonal projection $\Pi = \Pi^a \oplus \Pi^b$: $\mathscr{H} = \mathscr{H}^a \oplus \mathscr{H}^b \rightarrow \mathscr{H}^\pi$ maps $\mathscr{D}(H)$ into $\mathscr{D}(H) \cap \mathscr{H}^\pi$.

(Π2) The approximate asymptotic Hamiltonian H^π is self-adjoint and has a domain that satisfies $\Pi \mathscr{D}(H) \subset \mathscr{D}(H^\pi) \subset \mathscr{D}(H) \cap \mathscr{H}^\pi$. The operator Π commutes with H^π on $\mathscr{D}(H^\pi)$. H^π is reduced by the decomposition $\mathscr{H}^\pi = \mathscr{H}^{\pi a} \oplus \mathscr{H}^{\pi b}$ and thus may be written as $H^\pi = H^{\pi a} \oplus H^{\pi b}$. On $\mathscr{D}(H^{\pi a}) = \mathscr{D}(H^a) \cap \mathscr{H}^{\pi a}$ the operator $H^{\pi a}$ is given by $H^{\pi a} = H^a$, where H^a is the restriction of H to the reducing subspace $\mathscr{H}^a$. The operator $H^{\pi b}$ is self-adjoint and bounded from below on $\mathscr{H}^{\pi b}$.

(Π3) The operator $J^\pi J^{\pi *}$ has a bounded inverse, denoted by $(J^\pi J^{\pi *})^{-1}$, on $\mathscr{H}_\pi$.

(Π4) The operators U^π: $\mathscr{D}(H^\pi) \rightarrow \mathscr{H}_N$ and $U^{\pi *}$: $\mathscr{D}(H_N) \rightarrow \mathscr{H}^\pi$,

$$U^\pi \equiv H_N J^\pi - J^\pi H^\pi \quad \text{and} \quad U^{\pi *} \equiv J^{\pi *} H_N - H^\pi J^{\pi *}, \tag{3.6}$$

satisfy $U^\pi \lessdot H^\pi$ and $U^{\pi *} \lessdot H_N$, respectively.

Remark 3.1: (a) The operator $(J^\pi J^{\pi *})^{-1} P_\pi$: $\mathscr{H}_\pi \oplus \mathscr{H}_\pi^\perp \rightarrow \mathscr{H}_\pi$ is an extension of the inverse defined in Assumption (Π3). It is the unique maximal generalized inverse[32] of $J^\pi J^{\pi *}$. The requirement in (Π3) that $J^\pi J^{\pi *}$ has a bounded inverse on $\mathscr{H}_\pi$ is equivalent to the requirement that it has closed range. Alternatively, there must exist a constant $c_\pi > 0$ such that $\|J^\pi J^{\pi *} \psi\| \geqslant c_\pi \|\psi\|$ for all $\psi \in \mathscr{H}_\pi$. A useful criterion that guarantees the existence and boundedness of this inverse was established in Theorem 3.4 and Remark 3.5 of Ref. 16.

(b) Assumptions (Π2) and (Π4) are largely statements about Π^b and $H^{\pi b}$. If, for example, $\mathscr{H}^{\pi b} \subset \mathscr{D}(H)$ and is finite dimensional, then Assumptions (Π2) and (Π4) are satisfied.[16]

Condition C: Condition C is said to be satisfied if there is a set $\mathscr{E}^a \subset \mathscr{D}(H) \cap \mathscr{H}^a$ dense in $\mathscr{H}^a$ such that $\Pi^a \mathscr{E}^a \subset \mathscr{E}^a$ and

$$\int_{|t| > t_0} dt \, \| V e^{-iHt} \Phi^a \| < \infty \tag{3.7}$$

for some finite $t_0 \geqslant 0$ and for all $\Phi^a \in \mathscr{E}^a$.

Condition TC: Condition TC is said to be satisfied if the operator $VE(\Delta)\Pi^a$: $\mathscr{H} \rightarrow \mathscr{H}_N$ is trace class for all finite intervals $\Delta \subset (-\infty, \infty)$.

The following theorem which was proved in Ref. 16

establishes conditions under which an approximate scattering system $\mathfrak{S}(\Pi,H^\pi)$ has the same abstract structure as the exact scattering system $\mathfrak{S}$.

Theorem 3.2: Let $\mathfrak{S}$ satisfy Assumption A, and let $\mathfrak{S}(\Pi,H^\pi)$ satisfy Assumption Π. Suppose, in addition, that either Condition C or Condition TC is satisfied. Then $\mathfrak{S}(\Pi,H^\pi)$ also satisfies Assumption A. In particular, there exist *approximate wave operators*

$$\Omega^{\pi\pm}\equiv \operatorname*{s-lim}_{t\to\pm\infty} e^{iH_\pi t}J^\pi\Pi^a e^{-iH^\pi t} \tag{3.8}$$

and an *approximate scattering operator*

$$S^\pi\equiv\Omega^{\pi+*}\Omega^{\pi-}, \tag{3.9}$$

where $\Omega^{\pi+*}$ denotes the adjoint of $\Omega^{\pi+}$.

For later use we define the *approximate potentials*

$$V^\pi\equiv P_\pi U^\pi = H_\pi J^\pi - J^\pi H_\pi$$

and

$$V^{\pi*}\equiv U^{\pi*}P_\pi = J^{\pi*}H_\pi - H^\pi J^{\pi*} \tag{3.10}$$

and the *approximate resolvent operators*

$$R_\pi(z)\equiv(z-H_\pi)^{-1} \quad\text{and}\quad R^\pi(z)\equiv(z-H^\pi)^{-1}. \tag{3.11}$$

B. Approximate transition operators

If $\mathfrak{S}(\Pi,H^\pi)$ satisfies Assumption A, then the time-independent scattering theory for that approximate system has the following form (cf. Theorems 2.1 and 2.2).

An *approximate transition operator* $T^\pi(z)$: $\mathscr{D}(H^\pi)\to\mathscr{H}^\pi$ is defined by

$$\begin{aligned}T^\pi(z)&\equiv(z-H^\pi)J^{\pi*}\{R_\pi(z)J^\pi(z-H^\pi)-J^\pi\}\\ &=(z-H^\pi)J^{\pi*}R_\pi(z)V^\pi,\end{aligned} \tag{3.12}$$

where $z\in\rho(H_\pi)$. The scattering operator S^π is then related to $T^\pi(z)$ by the following theorem.

Theorem 3.3: Suppose that $\mathfrak{S}(\Pi,H^\pi)$ satisfies Assumptions (A1)–(A3) and (A5). Then, $S^\pi-\Pi^a$ has on $\mathscr{H}$ the representation

$$\begin{aligned}S^\pi-\Pi^a=\operatorname*{w-lim}_{\epsilon\to0^+}(-2\pi i)\int_\lambda\int_\mu dE(\lambda)\,\delta_\epsilon(\lambda-\mu)\\ \times\Pi^aT^\pi([\lambda+\mu+i\epsilon]/2)\Pi^a\,dE(\mu),\end{aligned} \tag{3.13}$$

where $E(\cdot)$, $\delta_\epsilon(\cdot)$, and the spectral integrals are as in Theorem 2.1.

Proof: The theorem is a transcription of Theorem 2.1 into the notation of the system $\mathfrak{S}(\Pi,H^\pi)$. □

Theorem 3.4: Let $\mathfrak{S}(\Pi,H^\pi)$ satisfy Assumptions (A1)–(A4). Then, for $z\in\rho(H_\pi)\cap\rho(H^\pi)$ the operator $T^\pi(z)$ is the unique solution defined on $\mathscr{D}(H^\pi)$ of the equation

$$T^\pi(z)=J^{\pi*}V^\pi+V^{\pi*}(J^\pi J^{\pi*})^{-1}J^\pi R^\pi(z)T^\pi(z). \tag{3.14}$$

Proof: The theorem is a transcription of Theorem 2.2 into the notation of the system $\mathfrak{S}(\Pi,H^\pi)$. □

We previously[29] approached Eq. (3.14) by introducing a subsidiary equation for the operator $K^\pi\equiv V^{\pi*}(J^\pi J^{\pi*})^{-1}J^\pi$:

$$K^\pi J^{\pi*}J^\pi=V^{\pi*}J^\pi. \tag{3.15}$$

A similar two-step approach has also been advocated in Ref. 33. The solution to Eq. (3.15) may possibly not be unique on $\mathscr{H}^\pi$, but the desired solution K^π is given by $K^\pi=V^{\pi*}J^\pi Y^{\pi\dagger}$, where $Y^{\pi\dagger}$ is the Moore–Penrose operator inverse of $Y^\pi\equiv J^{\pi*}J^\pi$. The difficulty with this approach is, then, not a matter of principle but a matter of practicality (although a simple calculation has been made using this scheme[34]). In a multichannel quantum mechanical calculation, for example, the operator equation (3.15) would be realized as a large number of coupled integral equations. To need to solve this large number of equations in addition to the large number of equations that realize Eq. (3.14) seems undesirable. We have, therefore, abandoned this approach for the present.

In order to eliminate the operator $(J^\pi J^{\pi*})^{-1}$ in our basic dynamical equation, we define the operator $M^\pi(z)$: $\mathscr{D}(H^\pi)\to\mathscr{H}^\pi$ by

$$M^\pi(z)\equiv(z-H^\pi)J^{\pi*}(J^\pi J^{\pi*})^{-1}R_\pi(z)V^\pi, \tag{3.16}$$

where $z\in\rho(H_\pi)$. From Eqs. (3.12) and (3.16) follows

$$T^\pi(z)=Q^\pi(z)M^\pi(z), \tag{3.17}$$

where

$$Q^\pi(z)\equiv(z-H^\pi)J^{\pi*}J^\pi R^\pi(z). \tag{3.18}$$

The following theorems establish dynamical equations for the *approximate M-operator* $M^\pi(z)$ [cf. Ref. 30, Theorem 3 and Eqs. (30) and (31)].

Theorem 3.5: Let $\mathfrak{S}(\Pi,H^\pi)$ satisfy Assumptions (A1)–(A4). Then, for $z\in\rho(H_\pi)\cap\rho(H^\pi)$, the operator $M^\pi(z)$ is a solution defined on $\mathscr{D}(H^\pi)$ of each of the equations

$$J^{\pi*}(z-H_\pi)J^\pi R^\pi(z)M^\pi(z)=J^{\pi*}V^\pi, \tag{3.19}$$

$$J^{\pi*}[J^\pi-V^\pi R^\pi(z)]M^\pi(z)=J^{\pi*}V^\pi, \tag{3.20}$$

$$M^\pi(z)=J^{\pi*}V^\pi+[J^{\pi*}V^\pi R^\pi(z)-(J^{\pi*}J^\pi-\Pi)]M^\pi(z). \tag{3.21}$$

Proof: Substitution of $M^\pi(z)$ defined by Eq. (3.16) into the left side of Eq. (3.19) gives $J^{\pi*}V^\pi$, proving that $M^\pi(z)$ satisfies Eq. (3.19). Equation (3.20) follows from Eq. (3.19) upon substitution of the identity $H_\pi J^\pi=V^\pi+J^\pi H^\pi$ into the left side of Eq. (3.19). Equation (3.21) follows from Eq. (3.20) and the identity $\Pi M^\pi(z)=M^\pi(z)$. □

Remark 3.6: Equations (3.19)–(3.21) are all completely equivalent provided the range of any solution is required to lie in $\mathscr{H}^\pi$. Equation (3.21) automatically enforces this requirement and, hence, might be preferable as the basis of practical calculations.

Theorem 3.7: Let $\mathfrak{S}(\Pi,H^\pi)$ satisfy Assumptions (A1)–(A4). If $N^\pi(z)$, $z\in\rho(H_\pi)\cap\rho(H^\pi)$, is a solution on $\mathscr{D}(H^\pi)$ of any of Eqs. (3.19)–(3.21), then $T^\pi(z)=Q^\pi(z)N^\pi(z)$.

Proof: Suppose that $N^\pi(z)$ is a solution of Eq. (3.19) and consider $\Delta^\pi(z)\equiv M^\pi(z)-N^\pi(z)$. Then, since $M^\pi(z)$ is also a solution of Eq. (3.19),

$$J^{\pi*}(z-H_\pi)J^\pi R^\pi(z)\Delta^\pi(z)=0. \tag{3.22}$$

Multiplying Eq. (3.22) from the left by $(z-H^\pi)J^{\pi*}R_\pi(z)(J^\pi J^{\pi*})^{-1}J^\pi$ yields $Q^\pi(z)\Delta^\pi(z)=0$. The theorem now follows from Eq. (3.17). The proof if $N^\pi(z)$ is a solution of Eq. (3.20) or (3.21) is similar. □

Remark 3.8: Efficient algorithms for constructing the

Moore–Penrose operator inverse are known.[32] By Theorem 3.7, if $L^{\pi\dagger}(z)$ is the Moore–Penrose operator inverse of $L^{\pi}(z) \equiv J^{\pi *}(z-H_\pi)J^{\pi *}R^{\pi}(z)$, then $L^{\pi\dagger}(z)J^{\pi *}V^{\pi}$ is a solution of Eqs. (3.19)–(3.21) and $T^{\pi}(z)=Q^{\pi}(z)L^{\pi\dagger}(z)J^{\pi *}V^{\pi}$.

Under certain circumstances the multiplication by $Q^{\pi}(z)$ of the solution of Eqs. (3.19)–(3.21) is unnecessary.

Theorem 3.9: Let $\mathfrak{S}$ satisfy Assumption A and let $\mathfrak{S}(\Pi,H^{\pi})$ satisfy Assumption Π. Suppose, in addition, that either Condition C or Condition TC is satisfied, and assume that $J^{\pi *}J^{\pi}-\Pi$ is compact. Then on $\mathscr{H}^{\pi}$,

$$S^{\pi}-\Pi^{a}=\operatorname*{w-lim}_{\epsilon\to 0^{+}}(-2\pi i)\int_{\lambda}\int_{\mu}dE(\lambda)\,\delta_{\epsilon}(\lambda-\mu)\times\Pi^{a}M^{\pi}\Big(\frac{\lambda+\mu+i\epsilon}{2}\Big)\Pi^{a}\,dE(\mu), \qquad (3.23)$$

where $E(\cdot)$, $\delta_\epsilon(\cdot)$ and the spectral integrals are as in Theorem 2.1.

Proof: Let $W^{\pi}(\epsilon)$ denote the quantity to the right of w-lim in Eq. (3.23). Then, by working backwards through the proofs of Theorem 4 of Ref. 17 and Theorem 3 of Ref. 18, we obtain

$$W^{\pi}(\epsilon)=\epsilon\int_{0}^{\infty}dt\,e^{-\epsilon t}\Pi^{a}e^{iH^{\pi}t}J^{\pi *}(J^{\pi}J^{\pi *})^{-1}\times[e^{-2iH_{\pi}t}J^{\pi}-J^{\pi}e^{-2iH^{\pi}t}]e^{iH^{\pi}t}\Pi^{a}. \qquad (3.24)$$

From Theorem 3.12 of Ref. 16 and the existence of $\Omega^{\pi\pm}$ follows

$$\begin{aligned}\operatorname*{w-lim}_{\epsilon\to 0}W^{\pi}(\epsilon)&=[\Omega^{+*}(H_\pi,H^{\pi},(J^{\pi}J^{\pi *})^{-1}J^{\pi}\Pi^{a})\\&\quad-\Omega^{-*}(H_\pi,H^{\pi},(J^{\pi}J^{\pi *})^{-1}J^{\pi}\Pi^{a})]\Omega^{\pi -},\\&=[\Omega^{\pi +*}-\Omega^{\pi -*}]\Omega^{\pi -}. \end{aligned}\qquad (3.25)$$

The theorem now follows from the definition of S^{π} and the partial isometry of $\Omega^{\pi -}$. □

Proposition 3.10: If $J^{\pi *}J^{\pi}-\Pi$ is compact, then the null space $\mathscr{N}(J^{\pi})\subset\mathscr{H}^{\pi}$ is finite dimensional.

Proof: Let P^{π} denote the orthogonal projection of $\mathscr{H}$ onto $\overline{\mathscr{R}}(J^{\pi *})$. Then the operator

$$\Pi-P^{\pi}=(J^{\pi *}J^{\pi}-P^{\pi})-(J^{\pi *}J^{\pi}-\Pi) \qquad (3.26)$$

is the orthogonal projection of $\mathscr{H}^{\pi}$ onto $\mathscr{N}(J^{\pi})$. As both terms on the right side of Eq. (3.26) are compact, $\Pi-P^{\pi}$ is a compact, hence finite-dimensional, orthogonal projection. □

Now, in Theorem 3.9 the operator $J^{\pi *}J^{\pi}-\Pi$ is assumed to be compact. It follows from Proposition 3.10 that $\mathscr{N}(J^{\pi})$ is finite dimensional. It is reasonable[35] that in the practical case Π can be chosen so that $\mathscr{N}(J^{\pi})=\{0\}$, i.e., $\overline{\mathscr{R}}(J^{\pi *})=\mathscr{H}^{\pi}$. In that case the following theorem holds.

Theorem 3.11: Let $\mathfrak{S}(\Pi,H^{\pi})$ satisfy Assumptions (A1)–(A4) and suppose that $\mathscr{N}(J^{\pi})=\{0\}$. Then, for $z\in\rho(H_\pi)\cap\rho(H^{\pi})$, the operator $M^{\pi}(z)$ is the unique solution on $\mathscr{D}(H^{\pi})$ of each of Eqs. (3.19)–(3.21).

Proof: It remains only to prove the uniqueness. Let $N^{\pi}(z)$ be any solution of one of Eqs. (3.19)–(3.21), and define $\Delta^{\pi}(z)\equiv M^{\pi}(z)-N^{\pi}(z)$. Then $\Delta^{\pi}(z)$ satisfies Eq. (3.22). Multiplication of Eq. (3.22) from the left by $R_\pi(z)(J^{\pi}J^{\pi *})^{-1}J^{\pi}$ yields $J^{\pi}R^{\pi}(z)\Delta^{\pi}(z)=0$. Since $\mathscr{N}(J^{\pi})=\{0\}$, this implies that $R^{\pi}(z)\Delta^{\pi}(z)=0$ and, hence, that $\Delta^{\pi}(z)=0$. □

The following theorem and corollary establish sufficient conditions for the kernel of Eq. (3.21) to be a compact operator on $\mathscr{H}^{\pi}$.

Theorem 3.12: Let $\mathfrak{S}(\Pi,H^{\pi})$ satisfy Assumptions (A1)–(A4). Suppose, in addition, that $J^{\pi *}J^{\pi}-\Pi$ is a compact operator on $\mathscr{H}$ and that the operator $V^{\pi}R^{\pi}(iy)E^{\pi}(\Delta)$: $\mathscr{H}^{\pi}\to\mathscr{H}_{\pi}$ is compact for some nonzero y and all finite intervals $\Delta\subset(-\infty,\infty)$. Then the kernel of Eq. (3.21) is compact for all z with $\operatorname{Im} z\neq 0$.

Proof: Let $\Delta_n\equiv[-n,n]$ and $\Delta_n'\equiv(-\infty,\infty)\setminus\Delta_n$. By Assumption (A4), for every $\epsilon>0$ there is a $b>0$ such that for all $\Phi\in\mathscr{H}^{\pi}$,

$$\|V^{\pi}R^{\pi}(iy)E^{\pi}(\Delta_n')\Phi\|\leqslant\epsilon\|H^{\pi}R^{\pi}(iy)E^{\pi}(\Delta_n')\Phi\|+b\,\|R^{\pi}(iy)E^{\pi}(\Delta_n')\Phi\|. \qquad (3.27)$$

It follows [cf. Eq. (5.52) of Ref. 28] that

$$\|V^{\pi}R^{\pi}(iy)E^{\pi}(\Delta_n')\|\leqslant\epsilon+b\,(n^2+y^2)^{-1/2}. \qquad (3.28)$$

Since the right side of inequality (3.28) can be made as small as desired by choosing n sufficiently large, the operator $V^{\pi}R^{\pi}(iy)$ is a norm limit of the sequence $V^{\pi}R^{\pi}(iy)E^{\pi}(\Delta_n)$ of compact operators and is, therefore, compact. Using the first resolvent equation, the operator

$$J^{\pi *}V^{\pi}R^{\pi}(z)=J^{\pi *}V^{\pi}R^{\pi}(iy)+(iy-z)J^{\pi *}V^{\pi}R^{\pi}(iy)R^{\pi}(z) \qquad (3.29)$$

is then seen to be compact for all z with $\operatorname{Im} z\neq 0$. The compactness of the kernel of Eq. (3.21) follows. □

Corollary 3.13: Let $\mathfrak{S}(\Pi,H^{\pi})$ satisfy Assumption Π. Let Condition TC hold, and let Π^{b} and $J^{\pi *}J^{\pi}-\Pi$ be compact. Then the kernel of Eq. (3.21) is a compact operator on $\mathscr{H}^{\pi}$ for all z with $\operatorname{Im} z\neq 0$.

Proof: Theorem 3.12 applies since for all $\Delta\subset(-\infty,\infty)$ and $y>0$,

$$V^{\pi}R^{\pi}(iy)E^{\pi}(\Delta)=P_\pi VE(\Delta)\Pi^{a}R^{\pi}(iy)+V^{\pi}R^{\pi}(iy)E^{\pi}(\Delta)\Pi^{b} \qquad (3.30)$$

is compact (the sum of a trace-class operator and a compact operator). □

C. Stability

The equivalent Eqs. (3.19)–(3.21) are stable under certain perturbations, as the next two theorems establish.

Theorem 3.14: Let $\mathfrak{S}(\Pi,H^{\pi})$ satisfy Assumption Π. Let $\Psi^{\pi}\in\mathscr{D}(H^{\pi})$, and let Φ^{π} be a solution in $\mathscr{H}^{\pi}$ of the unperturbed equation

$$J^{\pi *}(z-H_\pi)J^{\pi}R^{\pi}(z)\Phi^{\pi}=J^{\pi *}V^{\pi}\Psi^{\pi}, \qquad (3.31)$$

where $z\in\rho(H_\pi)\cap\rho(H^{\pi})$. Let $F_\pi=F_\pi(z)$: $\mathscr{D}(H_\pi)\to\mathscr{H}_\pi$ satisfy $\|F_\pi(z)R_\pi(z)\|<1$, and let $\Upsilon^{\pi}\in\overline{\mathscr{R}}(J^{\pi *})$. Then the perturbed equation

$$J^{\pi *}[z-H_\pi+F_\pi(z)]\,J^{\pi}R^{\pi}(z)\Theta^{\pi}=J^{\pi *}V^{\pi}\Psi^{\pi}+\Upsilon^{\pi} \qquad (3.32)$$

has the solution

$$\Theta_0^\pi = (R^\pi)^{-1} J^{\pi*} (J^\pi J^{\pi*})^{-1} R_\pi [I_N + F_\pi R_\pi]^{-1} \times (J^\pi J^{\pi*})^{-1} J^\pi (J^{\pi*} V^\pi \Psi^\pi + \Upsilon^\pi) . \tag{3.33}$$

Moreover, if Θ^π is any solution in $\mathscr{H}^\pi$ of Eq. (3.32), then $Q^\pi\Theta^\pi = Q^\pi\Theta_0^\pi$.

Proof: Under the assumptions of the theorem $[I_N + F_\pi R_\pi]^{-1}$ exists and Θ_0^π is well defined. Direct substitution of Θ_0^π into the left side of Eq. (3.32) demonstrates that Θ_0^π is a solution. If Θ^π is any other solution of Eq. (3.32), then $\Theta^\pi - \Theta_0^\pi$ is a solution of the homogeneous equation. Multiplying the homogeneous equation from the left by $(R^\pi)^{-1} J^{\pi*} R_\pi [I_N + F_\pi R_\pi]^{-1} (J^\pi J^{\pi*})^{-1} J^\pi$ yields the desired result, $Q^\pi\Theta^\pi = Q^\pi\Theta_0^\pi$. □

Remark 3.15: (a) The perturbation $J^{\pi*} F_\pi(z) J^\pi R^\pi(z)$ is of the most general form with range contained in $\mathscr{R}(J^{\pi*})$ and null space containing that of the unperturbed operator $J^{\pi*}(z - H_\pi) J^\pi R^\pi(z)$. The range condition is imposed to avoid complicated consistency requirements on Υ^π. The null space requirement is imposed so that the perturbation can always be compared to the unperturbed operator.

(b) If $\mathscr{N}(J^\pi) = 0$, then the solution to Eq. (3.32) is unique (cf. proof of Theorem 3.11).

Theorem 3.16: Let the assumptions of Theorem 3.14 hold, and let Θ^π and Φ^π be solutions in $\mathscr{H}^\pi$ of Eq. (3.32) and Eq. (3.31), respectively. Let P denote the orthogonal projection of $\mathscr{H}$ onto $\mathscr{R}(J^*)$, and let $\Delta \subset (-\infty, \infty)$ be any bounded interval. Then,

$$\|E(\Delta) I^a Q^\pi(\Theta^\pi - \Phi^\pi)\| \leqslant \|E(\Delta)(z - H)\| \{c_1\|\Phi^\pi\| + c_2\|\Upsilon^\pi\|\} , \tag{3.34}$$

$$\|PR(z) I^a Q^\pi(\Theta^\pi - \Phi^\pi)\| \leqslant c_1\|\Phi^\pi\| + c_2\|\Upsilon^\pi\| , \tag{3.35}$$

where

$$c_1 \equiv |\operatorname{Im} z|^{-1} \|J^*\| (1 - \|F_\pi R_\pi\|)^{-1} \times \|F_\pi R_\pi\| (\|J\| + \|V^\pi R^\pi\|) , \tag{3.36}$$

$$c_2 \equiv |\operatorname{Im} z|^{-1} \|J^*\| \, \|(J^\pi J^{\pi*})^{-1} J^\pi\| (1 - \|F_\pi R_\pi\|)^{-1} . \tag{3.37}$$

Proof: The vector $\Theta^\pi - \Phi^\pi$ must satisfy the equation

$$J^{\pi*}[z - H_\pi + F_\pi(z)] J^\pi R^\pi(z)(\Theta^\pi - \Phi^\pi) = -J^{\pi*} F_\pi(z) J^\pi R^\pi(z) \Phi^\pi + \Upsilon^\pi . \tag{3.38}$$

Multiplying Eq. (3.38) by $(R^\pi)^{-1} J^{\pi*} R_\pi [I_N + F_\pi R_\pi]^{-1} (J^\pi J^{\pi*})^{-1} J^\pi$ yields

$$Q^\pi(\Theta^\pi - \Phi^\pi) = (R^\pi)^{-1} J^{\pi*} R_\pi [I_N + F_\pi R_\pi]^{-1} \{ -F_\pi J^\pi R^\pi \Phi^\pi + (J^\pi J^{\pi*})^{-1} J^\pi \Upsilon^\pi \} . \tag{3.39}$$

Consequences of Assumption (Π2) are $E(\Delta) I^a (R^\pi)^{-1}(z) J^{\pi*} = E(\Delta)(z - H) \Pi^a J^*$, and $PR(z) I^a (R^\pi)^{-1}(z) J^{\pi*} = P\Pi^a J^*$. Also,

$$F_\pi J^\pi R^\pi = F_\pi R_\pi (R^\pi)^{-1} J^\pi R^\pi = F_\pi R_\pi [J^\pi + V^\pi R^\pi] . \tag{3.40}$$

Equations (3.34)–(3.37) now follow in a straightforward way. □

IV. SEQUENCES OF APPROXIMATE SCATTERING SYSTEMS

A. Assumption $\Pi^{(n)}$

Consider a sequence $\{\Pi^{(n)}\}$ of orthogonal projection operators $\Pi^{(n)} = \Pi^{(n)a} \oplus \Pi^{(n)b}$: $\mathscr{H} \to \mathscr{H}^{(n)}$, where $\mathscr{H}^{(n)} \equiv \Pi^{(n)}\mathscr{H} \equiv \mathscr{H}^{(n)a} \oplus \mathscr{H}^{(n)b}$. For each value of n there is an *approximate asymptotic Hamiltonian* $H^{(n)}$, defined on $\mathscr{D}(H^{(n)}) \subset \mathscr{H}^{(n)}$, and a bounded *identification operator* $J^{(n)} \equiv J\Pi^{(n)}$: $\mathscr{H} \to \mathscr{H}_N$. In addition, there are the space $\mathscr{H}_{(n)} \equiv \mathscr{R}(J^{(n)}) = P_{(n)}\mathscr{H}_N$, where $P_{(n)}$ is the orthogonal projection of $\mathscr{H}_N$ onto $\mathscr{H}_{(n)}$, and the *approximate total Hamiltonian* $H_{(n)} \equiv P_{(n)} H_N P_{(n)}$. These entities together define a *sequence* $\{\mathfrak{S}(n)\}$ *of approximate scattering systems*

$$\mathfrak{S}(n) \equiv (\mathscr{H}_{(n)}, H_{(n)}, \mathscr{H}^{(n)}, H^{(n)}, J^{(n)}, \Pi^{(n)a}) . \tag{4.1}$$

For notational convenience in what follows, subscripts and superscripts (n) will be used in place of the more cumbersome $\Pi^{(n)}$ that would be demanded by the notation of the previous section. Also, the symbol $\Pi^{(n)a}$ ($\Pi^{(n)b}$) will be used to denote the orthogonal projection of either $\mathscr{H}$, $\mathscr{H}^{(n)}$, or $\mathscr{H}^a$ ($\mathscr{H}^b$) onto $\mathscr{H}^{(n)a}$ ($\mathscr{H}^{(n)b}$).

Our fundamental assumptions about the sequence $\{\mathfrak{S}(n)\}$ of approximate scattering systems are contained in the following Assumption $\Pi^{(n)}$.

Assumption $\Pi^{(n)}$: The sequence $\{\mathfrak{S}(n)\}$ of approximate scattering systems is said to satisfy Assumption $\Pi^{(n)}$ if the following statements are true.[16]

($\Pi^{(n)}$1) For each n, $\mathfrak{S}(n)$ satisfies Assumption Π.

($\Pi^{(n)}$2) If $n \geqslant m$, $\Pi^{(n)}\Pi^{(m)} = \Pi^{(m)}$.

($\Pi^{(n)}$3) The strong limit $\Pi^{(\infty)}$, which exists by Assumption ($\Pi^{(n)}$2), satisfies

$$\Pi^{(\infty)} \equiv \operatorname*{s-lim}_{n\to\infty} \Pi^{(n)} = I^a \oplus \Pi^{(\infty)b} . \tag{4.2}$$

In addition, $\mathscr{R}(J\Pi^{(\infty)}J^*) = \mathscr{H}_N$, $\Pi^{(\infty)}\mathscr{D}(H) \subset \mathscr{D}(H)$, and $[\Pi^{(\infty)}, H] \ll H$.

($\Pi^{(n)}$4) for every $\Phi \in \mathscr{D}(H)$, the sequence $\{H\Pi^{(n)b}\Phi\}$ converges.

B. Convergence results

The convergence of certain sequences of approximate wave and scattering operators to the exact ones was proved in Ref. 16. In this subsection we obtain some results concerning the convergence of a *sequence of approximate transition operators*

$$T^{(n)}(z) \equiv (z - H^{(n)}) J^{(n)*} R_{(n)}(z) V^{(n)} \tag{4.3}$$

to the exact transition operator $T(z)$. Here $V^{(n)} \equiv H_{(n)} J^{(n)} - J^{(n)} H^{(n)}$, $R_{(n)}(z) \equiv (z - H_{(n)})^{-1}$ and $J^{(n)*} \equiv \Pi^{(n)} J^*$.

Theorem 4.1: Let $\mathfrak{S}$ satisfy Assumptions (A1)–(A4), and let $\{\mathfrak{S}(n)\}$ satisfy Assumption $\Pi^{(n)}$. Then for all $\Phi \in \mathscr{D}(H)$, all z such that $\operatorname{Im} z \neq 0$, and all finite intervals $\Delta \subset (-\infty, \infty)$,

$$\lim_{n\to\infty} E(\Delta) I^a \{T^{(n)}(z) - T(z)\} I^a \Phi = 0 \tag{4.4}$$

and

$$\lim_{n\to\infty} PR(z) I^a \{T^{(n)}(z) - T(z)\} I^a \Phi = 0 . \tag{4.5}$$

Proof: Let $\Psi^{(n)} = \Psi^{(n)}(z) \equiv I^a\{T^{(n)}(z) - \Pi^{(n)a}T(z)\}I^a\Phi$ and let $\Phi \in \mathscr{D}(H)$. Then it is elementary that

$$\Psi^{(n)} = \Pi^{(n)a}R^{-1}J^*\{R_{(n)}P_{(n)} - R_N\}VI^a\Phi - \Pi^{(n)a}R^{-1}J^*R_{(n)}P_{(n)}VR\{I^a - \Pi^{(n)a}\}R^{-1}\Phi. \quad (4.6)$$

The operator $E(\Delta)\Pi^{(n)a}R^{-1}J^* = \Pi^{(n)a}E(\Delta)R^{-1}J^*$ is bounded for finite Δ. The operator $R_{(n)}P_{(n)}$ is uniformly bounded in n by $|\mathrm{Im}\, z|^{-1}$, and $\|VR(z)\| \leqslant \epsilon + b|\mathrm{Im}\, z|^{-1}$ [cf. Eq. (2.32) of Ref. 1]. Further, the difference $\{R_N - R_{(n)}P_{(n)}\} \to 0$ strongly (Theorem 4.5 of Ref. 16) and $\{I^a - \Pi^{(n)a}\} \to 0$ strongly [Eq. (4.2)]. It follows that $E(\Delta)\Psi^{(n)} \to 0$, and consequently Eq. (4.4) will be true provided that $E(\Delta)\{I^a - \Pi^{(n)a}\}T(z)I^a\Phi \to 0$ as $n \to \infty$. But this follows immediately from $\|E(\Delta)\| \leqslant 1$ and Eq. (4.2). In addition, the operator $PRI^aR^{-1}\Pi^{(n)a}J^* = P\Pi^{(n)a}J^*$ is bounded uniformly in n. It thus follows that $PR\Psi^{(n)} \to 0$ and that Eq. (4.5) is true if $PR(z)\{I^a - \Pi^{(n)a}\}T(z)I^a\Phi \to 0$. Again, this follows from $\|PR(z)\| \leqslant |\mathrm{Im}\, z|^{-1}$ and Eq. (4.2). □

Equation (4.5) is equivalent to convergence in the $\mathscr{M}$-norm defined in Appendix A (cf. Ref. 30). Equation (4.4) shows that this norm can be avoided if one restricts the energies to a finite interval. If the projection operators $\Pi^{(n)}$ are further restricted, the factors $E(\Delta)$ and PR in Theorem 4.1 can be discarded and a stronger result (Theorem 4.7) proved.

Condition Γ: A sequence $\{\Pi^{(n)}\}$ is said to satisfy Condition Γ if there exist $c_0 > 0$ and $n_0 \geqslant 1$ such that $\|J\Pi^{(n)}J^*\psi\| \geqslant c_0\|P_{(n)}\psi\|$ for all $\psi \in \mathscr{H}_N$ and all $n \geqslant n_0$.

Remark 4.2: Condition Γ is analogous to Condition A of the Petrov–Galerkin method (cf. Ref. 26).

In preparation for the proof of Theorem 4.7 we introduce the notion of uniform infinitesimal smallness and prove some auxiliary results that involve it.

Definition 4.3: Let $\{L_n\}$, $L_n: \mathscr{X} \to \mathscr{X}$, be a sequence of linear operators, each densely defined on a Hilbert space $\mathscr{X}$. Let $\{K_n\}$, $K_n: \mathscr{X} \to \mathscr{Y}$, be a sequence of linear operators from $\mathscr{X}$ to a Hilbert space $\mathscr{Y}$. If the domains satisfy $\mathscr{D}(K_n) \supset \mathscr{D}(L_n)$ for each n, and if for all $\epsilon > 0$ there exists a finite $b = b(\epsilon)$, independent of n, such that

$$\|K_n\psi_n\| \leqslant \epsilon\|L_n\psi_n\| + b\|\psi_n\| \quad (4.7)$$

for all $\psi_n \in \mathscr{D}(L_n)$ and all n, then K_n is said to be *uniformly infinitesimally small* with respect to L_n. This will be written $K_n \lessdot L_n$ uniformly.

Lemma 4.4: Let L_n be self-adjoint for each n. Then $K_n \lessdot L_n$ uniformly if and only if $\|K_n(iy - L_n)^{-1}\| \to 0$ as $|y| \to \infty$ uniformly in n.

Proof: We omit the elementary proof. □

Lemma 4.5: Let $\mathfrak{S}$ satisfy Assumptions (A1)–(A4), let $\{\mathfrak{S}(n)\}$ satisfy Assumption $\Pi^{(n)}$ and let $\{\Pi^{(n)}\}$ satisfy Condition Γ. Further, let

$$U^{(n)} \equiv H_N J^{(n)} - J^{(n)}H^{(n)}$$

and

$$U^{(n)*} \equiv J^{(n)*}H_N - H^{(n)}J^{(n)*}, \quad (4.8)$$

and assume that $U^{(n)} \lessdot H^{(n)}$ uniformly and $U^{(n)*} \lessdot H_N$ uniformly. Then, for any fixed z with $\mathrm{Im}\, z \neq 0$ the operator $(z - H_N)R_{(n)}(z)P_{(n)}$ is bounded uniformly in n.

Proof: For any real y, $|y| > 0$, define $A_{(n)}(y): \mathscr{H}_N \to \mathscr{H}_N$ by

$$A_{(n)}(y) \equiv (iy - H_N)J^{(n)}J^{(n)*}R_N(iy) - J^{(n)}J^{(n)*} \quad (4.9)$$

[cf. Eq. (3.7) of Ref. 16]. Equation (3.11) of Ref. 16 and Lemma 4.4 imply that $\|A_{(n)}(y)\|$ can be made as small as desired, uniformly in n, by choosing $|y|$ large enough. Condition Γ implies that $\|(J^{(n)}J^{(n)*})^{-1}P_{(n)}\| \leqslant c_0^{-1}$ for $n \geqslant n_0$. It follows that the bound on $(z - H_N)P_{(n)}R_{(n)}(z)$ obtained in Theorem 3.9 of Ref. 16 is uniform in n. □

Lemma 4.6: Let the assumptions of Lemma 4.5 be satisfied. Then, for any fixed z with $\mathrm{Im}\, z \neq 0$,

$$\operatorname*{s-lim}_{n\to\infty} (z - H_N)R_{(n)}(z)P_{(n)} = I_N. \quad (4.10)$$

Proof: Assumption $(\Pi^{(n)}3)$ implies that $Z \equiv J\Pi^{(\infty)}J^*$ has a bounded inverse on $\mathscr{H}_N$ and Proposition 4.4 of Ref. 16 implies that Z^{-1} maps $\mathscr{D}(H_N)$ onto $\mathscr{D}(H_N)$. For z with $\mathrm{Im}\, z \neq 0$ define

$$D_{(n)} = D_{(n)}(z) \equiv (z - H_N)R_{(n)}(z)P_{(n)} - I_N. \quad (4.11)$$

Since $D_{(n)}(z)(z - H_N)J^{(n)} = 0$, we have the identity

$$D_n = D_n[R_N^{-1}JR][R^{-1}(\Pi^{(\infty)} - \Pi^{(n)})R] \times R^{-1}J^*Z^{-1}R_N. \quad (4.12)$$

By Lemma 4.5 $D_n(z)$ is bounded uniformly in n. The operator $R_N^{-1}JR$ is also bounded (Ref. 17, Lemma 1). The remaining terms on the right side of Eq. (4.12) converge strongly to zero because of Assumption $(\Pi^{(n)}4)$ and Eq. (4.4) of Proposition 4.1 of Ref. 16. □

Theorem 4.7: Let the assumptions of Lemma 4.5 be satisfied. Then, for all $\Phi \in \mathscr{D}(H)$ and all z such that $\mathrm{Im}\, z \neq 0$,

$$\lim_{n\to\infty} I^a\{T^{(n)}(z) - T(z)\}I^a\Phi = 0. \quad (4.13)$$

Proof: Let $\Psi^{(n)}$ be defined as in the proof of Theorem 4.1 and let Eq. (4.6) be rewritten as

$$\Psi^{(n)} = \Pi^{(n)a}R^{-1}J^*R_N\{R_N^{-1}R_{(n)}P_{(n)} - I_N\}VI^a\Phi - \Pi^{(n)a}R^{-1}J^*R_NR_N^{-1}R_{(n)}P_{(n)}VR \times\{I^a - \Pi^{(n)a}\}R^{-1}\Phi. \quad (4.14)$$

The operators $\Pi^{(n)a}$, $R^{-1}J^*R_N$ and VR are bounded uniformly in n. By Lemma 4.5 the operator $R_N^{-1}R_{(n)}P_{(n)}$ is bounded uniformly in n. The first term on the right side of Eq. (4.14) thus converges to zero by Lemma 4.6, and the second term converges to zero because of Eq. (4.2). Equation (4.13) is therefore true if $\{I^a - \Pi^{(n)a}\}T(z)I^a\Phi \to 0$, which follows from Eq. (4.2). □

Remark 4.8: The hypotheses $U^{(n)} \lessdot H^{(n)}$ and $U^{(n)*} \lessdot H_N$ uniformly are satisfied if $(H\Pi^{(n)} - \Pi^{(n)}H^{(n)}) \lessdot H^{(n)}$ and $(\Pi^{(n)}H - H^{(n)}\Pi^{(n)}) \lessdot H$ uniformly. This in turn is satisfied (cf. Lemma 4.4) if there is a function $f(z)$ such that

$$\|(H\Pi^{(n)} - \Pi^{(n)}H^{(n)})R^{(n)}(z)\| < f(z) \quad (4.15)$$

and

$$\|(\Pi^{(n)}H - H^{(n)}\Pi^{(n)})R(z)\| < f(z), \quad (4.16)$$

where $f(z) \to 0$ as $\mathrm{Im}\, z \to \infty$. Here $R^{(n)}(z) \equiv (z - H^{(n)})^{-1}$.

C. Stability

Consider now a sequence $\{\mathfrak{S}(n)\}$ of approximate scattering systems. Then, for each fixed n, Theorems 3.14 and

3.16 hold, but the constants c_1 and c_2 in Theorem 3.16 may depend on n. A result that is uniform in n is the following.

Theorem 4.9: Let $\mathfrak{S}$ satisfy Assumptions (A1)–(A4), let $\{\mathfrak{S}(n)\}$ satisfy Assumption $\Pi^{(n)}$, and let $\{\Pi^{(n)}\}$ satisfy Condition Γ. Let $U^{(n)} \ll H^{(n)}$ uniformly. Let $\Psi^{(n)} \in \mathscr{D}(H^{(n)})$, and let $\Phi^{(n)}$ be solutions of the unperturbed equations

$$J^{(n)*}(z - H_{(n)})J^{(n)}R^{(n)}(z)\Phi^{(n)} = J^{(n)*}V^{(n)}\Psi^{(n)}, \qquad (4.17)$$

where $\operatorname{Im} z \neq 0$. Let $F_{(n)} = F_{(n)}(z)$: $\mathscr{D}(H_{(n)}) \to \mathscr{H}_{(n)}$ satisfy $\|F_{(n)}(z)R_{(n)}(z)\| < c_3$, where $c_3 < 1$ is independent of n, and let $\Upsilon^{(n)} \in \overline{\mathscr{R}}(J^{(n)*})$. Let $\Theta^{(n)}$ be solutions of the perturbed equations

$$J^{(n)*}(z - H_{(n)} + F_{(n)})J^{(n)}R^{(n)}(z)\Theta^{(n)} = J^{(n)*}V^{(n)}\Psi^{(n)} + \Upsilon^{(n)}. \qquad (4.18)$$

Then, there exist positive constants c_4 and c_5, independent of n, such that

$$\|E(\Delta)I^a Q^{(n)}(\Theta^{(n)} - \Phi^{(n)})\| \leqslant \|E(\Delta)(z - H)\|\{c_4\|\Phi^{(n)}\| + c_5\|\Upsilon^{(n)}\|\} \qquad (4.19)$$

for all bounded intervals $\Delta \subset (-\infty, \infty)$, and

$$\|PR(z)I^a Q^{(n)}(\Theta^{(n)} - \Phi^{(n)})\| \leqslant c_4\|\Phi^{(n)}\| + c_5\|\Upsilon^{(n)}\|. \qquad (4.20)$$

The constants c_4 and c_5 are given by

$$c_4 \equiv c_3(1 - c_3)^{-1}|\operatorname{Im} z|^{-1}\|J^*\|(\|J\| + \operatorname{Sup}_n \|U^{(n)}R^{(n)}(z)\|), \qquad (4.21)$$

$$c_5 \equiv c_0^{-1}(1 - c_3)^{-1}|\operatorname{Im} z|^{-1}\|J\|\,\|J^*\|. \qquad (4.22)$$

Proof: These results are almost immediate from Theorem 3.16. By Condition Γ, $\operatorname{Sup}_n \|(J^{(n)}J^{(n)*})^{-1}J^{(n)}\| \leqslant c_0^{-1}\|J\|$. Also,

$$\|V^{(n)}R^{(n)}(z)\| = \|P_{(n)}U^{(n)}R^{(n)}(z)\| \leqslant \operatorname{Sup}_n \|U^{(n)}R^{(n)}(z)\| < \infty \qquad (4.23)$$

by Lemma 4.4 and the assumption that $U^{(n)} \ll H^{(n)}$ uniformly. The values of c_4 and c_5 then follow from Eqs. (3.36) and (3.37). □

Remark 4.10: If $H^{(n)} = H$ and $\|F_{(n)}(z)\| < c_3|\operatorname{Im} z|$ for all n, then $\|F_{(n)}(z)R_{(n)}(z)\| < c_3$ and $\|U^{(n)}R^{(n)}(z)\| \leqslant \|VR(z)\|$. The uniform boundedness assumptions of Theorem 4.9 would therefore be satisfied under those conditions.

Remark 4.11: Suppose, in addition to the assumptions of Theorem 4.9 it is assumed that $U^{(n)*} \ll H_N$ uniformly. Then, by Lemma 4.5, the constant $c_6 \equiv \operatorname{Sup}\|(z - H_N)R_{(n)}(z)P_{(n)}\| < \infty$. Thus, if $\|F_{(n)}(z)R_N(z)\| \leqslant c_3 c_6^{-1}$, then $\|F_{(n)}(z)R_{(n)}(z)\| < c_3$. If, for example, the perturbation $F_{(n)}(z)$ consists of potentials that are uniformly small with respect to H_N, then uniform boundedness of $\|F_{(n)}(z)R_{(n)}(z)\|$ follows.

Remark 4.12: We announced stability results similar to those of Theorem 4.9 in a previous paper (Ref. 30, Theorem 8). In Ref. 30 $J^{(n)*}(z - H_{(n)})J^{(n)}R^{(n)}(z)$ was perturbed by the operator $\Pi^{(n)}F^{(n)}Q(z)\Pi^{(n)}$, where $Q(z)$ could be either $Q^T(z) \equiv (z - H)J^*JR(z)$ or $Q^M(z) \equiv (z - H)PR(z)$. It was assumed that $H^{(n)} = H$ and that $F^{(n)}$: $\mathscr{R}(Q) \to \overline{\mathscr{R}}(J^*)$. It can be verified by direct substitution that this previous form of the perturbation is obtained from our present form by setting $F_{(n)}(z) = P_{(n)}(JJ^*)^{-1}JF^{(n)}Q(z)(z - H)J^*(JJ^*)^{-1}$. The norm of $F^{(n)}$ appearing in Ref. 30 is, in the present notation, $\|F^{(n)}(z - H)P\|$ [cf. Lemma A5(b) of Appendix A]. Let $Q(z) = (z - H)J^*JR(z)$. Then

$$\|F_{(n)}(z)R_{(n)}(z)\| = \|P_{(n)}(JJ^*)^{-1}JF^{(n)}(z - H)PJ^*R_{(n)}(z)\| \leqslant c_2|\operatorname{Im} z|^{-1}\|J^*\|\,\|(JJ^*)^{-1}J\| < 1 \qquad (4.24)$$

if $\|F^{(n)}(z - H)P\| \leqslant c_2$, where c_2 is a sufficiently small positive constant. If $Q(z) = (z - H)PR(z)$, a similar result holds. Theorem 8 of Ref. 30 is therefore contained in Theorem 4.9.

V. OTHER TIME-INDEPENDENT RESULTS

A. Resolvent-type equations

Associated with the scattering system $\mathfrak{S}(\Pi, H^\pi)$ are some approximate resolvent equations.

Theorem 5.1: Let $\mathfrak{S}(\Pi, H^\pi)$ satisfy Assumptions (A1)–(A4). Then, for $z \in \rho(H_\pi) \cap \rho(H^\pi)$, the operator $R_\pi(z)$ is the unique solution on $\mathscr{H}_\pi$ of each of the equations

$$R_\pi(z)J^\pi = J^\pi R^\pi(z) + R_\pi(z)V^\pi R^\pi(z), \qquad (5.1)$$

$$J^{\pi*}R_\pi(z) = R^\pi(z)J^{\pi*} + R^\pi(z)V^{\pi*}R_\pi(z), \qquad (5.2)$$

and the operator $G^\pi(z) \equiv J^{\pi*}R_\pi(z)$ is the unique solution on $\mathscr{H}_\pi$ of

$$G^\pi(z) = R^\pi(z)J^{\pi*} + R^\pi(z)V^{\pi*}(J^\pi J^{\pi*})^{-1}J^\pi G^\pi(z). \qquad (5.3)$$

Proof: Equations (5.1) and (5.2) are the usual two-Hilbert-space resolvent equations written in the notation of the system $\mathfrak{S}(\Pi, H^\pi)$, and the proof of the existence of a unique solution is elementary. The existence of a unique solution of Eq. (5.3) follows from a transcription of Theorems 5 and 6 of Ref. 1 into the notation of the system $\mathfrak{S}(\Pi, H^\pi)$. □

Theorem 5.2: Let $\mathfrak{S}(\Pi, H^\pi)$ satisfy Assumptions (A1)–(A4). If $N^\pi(z)$, $z \in \rho(H_\pi) \cap \rho(H^\pi)$, is any solution on $\mathscr{D}(H^\pi)$ of Eqs. (3.19)–(3.21), then

$$R_\pi(z)J^\pi = J^\pi R^\pi(z) + J^\pi R^\pi(z)N^\pi(z)R^\pi(z). \qquad (5.4)$$

Proof: The validity of Eq. (5.4) with $N^\pi(z)$ replaced by $M^\pi(z)$ is immediate from Eqs. (3.16) and (5.1). For any other solution $N^\pi(z)$ of Eqs. (3.19)–(3.21) let $\Delta^\pi(z) \equiv M^\pi(z) - N^\pi(z)$. Then $\Delta^\pi(z)$ is a solution of Eq. (3.22). A multiplication of Eq. (3.22) on the left by $R_\pi(z)(J^\pi J^{\pi*})^{-1}J^\pi$ gives $J^\pi R^\pi(z)\Delta^\pi(z) = 0$. It follows that Eq. (5.4) is satisfied. □

Remark 5.3: Formula (5.4) is of the type encountered in the limiting-absorption principle.[8,12,19–22] For a particular choice of pair potentials V_{ij} Eq. (5.4) has been used by Trucano[22] (see also Sec. 15 of Ref. 30) to prove asymptotic completeness for a two-fragment approximation of the N-body scattering problem.

Equations (5.1)–(5.4) for the approximate system $\mathfrak{S}(\Pi, H^\pi)$ have analogs for the exact system $\mathfrak{S}$. The equations corresponding to Eqs. (5.1)–(5.3) appear in Ref. 1. For an analog of Eq. (5.4) we need an exact M-operator equation corresponding to Eqs. (3.19)–(3.21). The *exact M-operator* analogous to Eq. (3.16) is the operator $M(z)$: $\mathscr{D}(H) \to \mathscr{H}$ defined by

$$M(z) \equiv (z - H)J^*(JJ^*)^{-1}R_N(z)V. \qquad (5.5)$$

Then the following theorem holds (cf. Ref. 30, Theorem 3 and Corollary 4.1).

Theorem 5.4: Let $\mathfrak{S}$ satisfy Assumptions (A1)–(A4). Then, for $z\in\rho(H_N)$, the operator $M(z)$ is a solution of each of the equivalent equations

$$J^*(z-H_N)JR(z)M(z)=J^*V, \tag{5.6}$$

$$J^*[J-VR(z)]M(z)=J^*V, \tag{5.7}$$

$$M(z)=J^*V+[J^*VR(z)-(J^*J-I)]M(z). \tag{5.8}$$

Furthermore, if $N(z)$, $z\in\rho(H_N)$, is any solution of one of the Eqs. (5.6)–(5.8), then

$$R_N(z)J=JR(z)+JR(z)N(z)R(z) \tag{5.9}$$

and

$$T(z)=(z-H)J^*JR(z)N(z). \tag{5.10}$$

Proof: The theorem is a transcription of Theorems 3.5 and 5.2 into the notation of the system $\mathfrak{S}$, and those proofs apply if one sets $\Pi=I$. □

Equations (5.6)–(5.8) do not, in general, have a unique solution on the entire space $\mathscr{H}$. In order to obtain a unique solution we must restrict the solution space. Define

$$\mathscr{D}=\mathscr{D}(z)\equiv(z-H)PR(z)\mathscr{H}. \tag{5.11}$$

Then the following theorem holds (cf. Ref. 30, Theorem 4).

Theorem 5.5: Let $\mathfrak{S}$ satisfy Assumptions (A1)–(A4), and let $z\in\rho(H_N)$. Then $M(z)$ is the unique solution of Eqs. (5.6)–(5.8) with range lying in $\mathscr{D}$.

Proof: Since $M(z)$ defined by Eq. (5.5) is a solution with range in $\mathscr{D}$, it remains only to prove the uniqueness. Suppose that $M(z)$ and $N(z)$ are two solutions of one of the Eqs. (5.6)–(5.8) with range in $\mathscr{D}$. For a given $\Psi\in\mathscr{D}(H)$, let $\Phi\equiv[M(z)-N(z)]\Psi$. Since $\Phi\in\mathscr{D}$, $\Phi=(z-H_N)P\Theta$ for some $\Theta\in\mathscr{H}$. Thus

$$J^*(z-H_N)J\Theta=J^*(z-H_N)JR(z)\Phi=0. \tag{5.12}$$

A multiplication on the left by $J^*(JJ^*)^{-1}R_N(z)(JJ^*)^{-1}J$ gives $P\Theta=0$, which implies that $\Phi=0$. Since $\Psi\in\mathscr{D}(H)$ is arbitrary, it follows that $M(z)=N(z)$. □

As a final note in this subsection, we remark that Eqs. (5.6)–(5.8) may be written in a differential equation form by using $\widetilde{M}(z)\equiv R(z)M(z)$ as the unknown in the equations. For example, Eq. (5.6) becomes

$$J^*(z-H_N)J\widetilde{M}(z)=J^*V. \tag{5.13}$$

Similarly, letting $\widetilde{M}^{\pi}(z)\equiv R^{\pi}(z)M^{\pi}(z)$, the approximate Eq. (3.19) may be rewritten in the form

$$\Pi J^*(z-H_N)J\Pi\widetilde{M}^{\pi}(z)=\Pi J^*V\Pi. \tag{5.14}$$

All of our results in Secs. III B and V A pertaining to $M(z)$ and $M^{\pi}(z)$ have obvious analogs in terms of the operators $\widetilde{M}(z)$ and $\widetilde{M}^{\pi}(z)$.

B. Convergence results

The convergence results of Sec. IV B have analogs for the convergence of a sequence of approximate M-operators to the exact M-operator. Define the sequence $M^{(n)}(z)$: $\mathscr{D}(H^{(n)})\to\mathscr{H}^{(n)}$ by

$$M^{(n)}(z)\equiv(z-H^{(n)})J^{(n)*}(J^{(n)}J^{(n)*})^{-1}R_{(n)}(z)V^{(n)}. \tag{5.15}$$

Also define

$$Q^{M(n)}(z)\equiv(z-H)PR^{(n)}(z)\Pi^{(n)} \tag{5.16}$$

and

$$\Lambda^{(n)}(z)\equiv\{Q^{M(n)}(z)M^{(n)}(z)-M(z)\}I^a. \tag{5.17}$$

Theorem 5.6: (a) Let the assumptions of Theorem 4.1 hold. Then, for all $\Phi\in\mathscr{D}(H)$, all z such that $\operatorname{Im}z\neq0$, and all finite intervals $\Delta\subset(-\infty,\infty)$,

$$\lim_{n\to\infty}E(\Delta)\Lambda^{(n)}(z)\Phi=0 \tag{5.18}$$

and

$$\lim_{n\to\infty}PR(z)\Lambda^{(n)}(z)\Phi=0. \tag{5.19}$$

(b) Let the assumptions of Lemma 4.5 hold. Then, for all $\Phi\in\mathscr{D}(H)$ and all z such that $\operatorname{Im}z\neq0$,

$$\lim_{n\to\infty}\Lambda^{(n)}(z)\Phi=0. \tag{5.20}$$

Proof: The vector $\Lambda^{(n)}\Phi=\Lambda^{(n)}(z)\Phi$ may be written as

$$\begin{aligned}\Lambda^{(n)}\Phi&=R^{-1}J^*(JJ^*)^{-1}\{R_{(n)}P_{(n)}-R_N\}VI^a\Phi\\&\quad-R^{-1}J^*(JJ^*)^{-1}R_{(n)}P_{(n)}VR\\&\quad\times\{I^a-\Pi^{(n)a}\}R^{-1}\Phi.\end{aligned} \tag{5.21}$$

Consequently, Eqs. (5.18) and (5.19) follow by essentially the same arguments as those in the proof of Theorem 4.1. Equation (5.20) follows by an argument similar to the proof of Theorem 4.7. □

Finally, there is a corresponding convergence result for the $\widetilde{M}(z)$ operator. Define the sequence $\widetilde{M}^{(n)}(z)\equiv R^{(n)}(z)M^{(n)}(z)$.

Theorem 5.7: Let $\{\mathfrak{S}(n)\}$ satisfy Assumptions ($\Pi^{(n)}$1)–($\Pi^{(n)}$3), and let $\mathfrak{S}$ satisfy Assumptions (A1)–(A4). Then, for all $\Phi\in\mathscr{D}(H)$ and all z such that $\operatorname{Im}z\neq0$,

$$\lim_{n\to\infty}P\{\widetilde{M}^{(n)}(z)-\widetilde{M}(z)\}I^a\Phi=0. \tag{5.22}$$

If, in addition, $\{\Pi^{(n)}\}$ satisfies Condition Γ and $\Pi^{(n)}\to I$ strongly as $n\to\infty$, then

$$\lim_{n\to\infty}\{\widetilde{M}^{(n)}(z)-\widetilde{M}(z)\}I^a\Phi=0. \tag{5.23}$$

Proof: Since

$$P\{\widetilde{M}^{(n)}-\widetilde{M}\}I^a\Phi=R\Lambda^{(n)}\Phi, \tag{5.24}$$

Eq. (5.21) and the arguments of Theorem 4.1 imply Eq. (5.22). In order to prove Eq. (5.23), note the identity

$$\begin{aligned}\{\widetilde{M}^{(n)}-\widetilde{M}\}I^a\Phi\\=J^{(n)*}(J^{(n)}J^{(n)*})^{-1}P_{(n)}[R_{(n)}-R_N]VI^a\Phi\\-J^{(n)*}(J^{(n)}J^{(n)*})^{-1}R_{(n)}P_{(n)}VR[I^a-\Pi^{(n)a}]R^{-1}\Phi\\+[J^{(n)*}(J^{(n)}J^{(n)*})^{-1}P_{(n)}-J^*(JJ^*)^{-1}]R_NVI^a\Phi.\end{aligned} \tag{5.25}$$

By Condition (Γ) the operator $(J^{(n)}J^{(n)*})^{-1}P_{(n)}$ is bounded uniformly in n. Therefore, the first term on the right side of Eq. (5.25) converges to zero by Theorem 4.5 of Ref. 16. Since $J^{(n)*}(J^{(n)}J^{(n)*})^{-1}R_{(n)}P_{(n)}VR$ is bounded uniformly in n and $\Pi^{(n)a}\to I^a$ strongly, the second term also converges to zero. The third term converges to zero provided that

$$\operatorname*{s-lim}_{n\to\infty} J^{(n)*}(J^{(n)}J^{(n)*})^{-1}P_{(n)} = J^*(JJ^*)^{-1}. \tag{5.26}$$

But Eq. (5.26) follows from Condition Γ, $\Pi^{(n)}\to I$ strongly, and the identity

$$\begin{aligned} J^{(n)*}(J^{(n)}J^{(n)*})^{-1}P_{(n)} - J^*(JJ^*)^{-1} &= J^{(n)*}(J^{(n)}J^{(n)*})^{-1}P_{(n)}J(I-\Pi^{(n)})J^*(JJ^*)^{-1} \\ &\quad + (\Pi^{(n)}-I)J^*(JJ^*)^{-1}. \end{aligned} \tag{5.27}$$

□

Remark 5.8: In Secs. III and IV we presented our approximation scheme as a means for constructing approximations $T^\pi(z) = Q^\pi(z)M^\pi(z)$ to the exact transition operator $T(z)$. The development of this section shows that $M^\pi(z)$ can also be considered as an approximation to $M(z)$. It has been established in the context of nonrelativistic multichannel quantum scattering theory that the singularity structure of $M(z)$ can be markedly different from that of $T(z)$.[36] Thus a Π that gives an accurate approximation to $T(z)$ can be expected to be very different from a Π that gives an accurate approximation to $M(z)$.

VI. DISCUSSION

In this paper we have established a mathematically rigorous framework within which to calculate approximate solutions to our basic dynamical equation for the transition operator $T(z)$. The assumptions which we make about the exact scattering system $\mathfrak{S}$ are contained in the abstract Assumption A of Sec. II A. In that assumption the asymptotic Hilbert space $\mathscr{H}$ is decomposed into the two parts $\mathscr{H} = \mathscr{H}^a \oplus \mathscr{H}^b$. The assumptions of Ref. 30 are contained in Assumption A as the particular case $\mathscr{H}^a = \mathscr{H}$ and $I^a = I$. The additional generality in the present case allows, for example, the following three possibilities.

(1) $\mathscr{H}^a$ could now be the absolutely continuous subspace of H and I^a the orthogonal projection of $\mathscr{H}$ onto this subspace (cf. Refs. 28 and 31).

(2) For computational purposes it might be useful to incorporate a pseudochannel space $\mathscr{H}_p \subset \mathscr{H}_N$ into our previous formalism. The states of $\mathscr{H}_p$ could represent compound nuclear states or any other states not easily expressed in terms of the true asymptotic states. A Hamiltonian H_p, self-adjoint and bounded from below, must be defined on $\mathscr{H}_p$. This Hamiltonian must have the three properties $I_p\mathscr{D}(H_N)\subset\mathscr{D}(H_p)\subset\mathscr{D}(H_N)$, $(H_NI_p - I_pH_p)\ll H_p$, and $(I_pH_N - H_pI_p)\ll H_N$, where I_p is the orthogonal projection of $\mathscr{H}_N$ onto $\mathscr{H}_p$.

For example, if $\mathscr{H}_p$ is finite dimensional, then H_p is a (bounded, symmetric) finite-rank operator and $\mathscr{D}(H_p) = \mathscr{H}_p$. If $\mathscr{H}_p \subset \mathscr{D}(H_N)$, then $(H_NI_p - I_pH_p)$ and $(I_pH_N - H_pI_p)$ are bounded operators and all of the assumptions of the previous paragraph are satisfied.

The pseudochannel is incorporated into the abstract formalism in the following way. Define the bounded linear operator $J': \mathscr{H}\oplus\mathscr{H}_p\to\mathscr{H}_N$ by $J'(\Psi\oplus\psi_p)\equiv J\Psi + \psi_p$, and let I' denote the orthogonal projection of $\mathscr{H}\oplus\mathscr{H}_p$ onto $\mathscr{H}$. Then if $\{\mathscr{H}_N, H_N, \mathscr{H}, H, J, I\}$ satisfies Assumption A, the assumed properties of H_p and $\mathscr{H}_p$ imply that $\{\mathscr{H}_N, H_N, \mathscr{H}\oplus\mathscr{H}_p, H\oplus H_p, J', I'\}$ also satisfies Assumption A. Our abstract theory therefore applies in the case by identifying $\mathscr{H}^a$ with $\mathscr{H}$ and $\mathscr{H}^b$ with $\mathscr{H}_p$.

(3) The systems that we have previously studied can also be modified to accommodate resonances.[37,38] Let $\mathscr{H}_R$ be a finite-dimensional Hilbert space. Let H_R be a self-adjoint operator on $\mathscr{H}_R$, and let $W_R: \mathscr{H}_R\to\mathscr{H}_N$ be a linear operator with adjoint W_R^*. Define the Hamiltonian $\widetilde{H}_N$: $\mathscr{D}(H_N)\oplus\mathscr{H}_R\to\mathscr{H}_N\oplus\mathscr{H}_R$ by

$$\widetilde{H}_N(\psi_N\oplus\psi_R)\equiv(H_N\psi_N + W_R\psi_R)\oplus(H_R\psi_R + W_R^*\psi_N). \tag{6.1}$$

Further, define $J'': \mathscr{H}\oplus\mathscr{H}_R\to\mathscr{H}_N\oplus\mathscr{H}_R$ by

$$J''(\Psi\oplus\psi_R) = J\Psi\oplus\psi_R, \tag{6.2}$$

and define I'' to be the orthogonal projection of $\mathscr{H}\oplus\mathscr{H}_R$ onto $\mathscr{H}$. Then, if $\{\mathscr{H}_N, H_N, \mathscr{H}, H, J, I\}$ satisfies Assumption A, so does $\{\mathscr{H}_N\oplus\mathscr{H}_R, \widetilde{H}_N, \mathscr{H}\oplus\mathscr{H}_R, H\oplus H_R, J'', I''\}$. Thus resonances can be included in the abstract theory by identifying $\mathscr{H}^a$ with $\mathscr{H}$ and $\mathscr{H}^b$ with $\mathscr{H}_R$.

Our method for calculating approximate solutions does not require that the kernel of the exact equation be compact and hence provides an alternative to the widely used Fredholm theory. The relation of our method to others that also do not require compact kernels is illuminating.

The formal structure of our method has much in common with the Petryshyn theory[26,27] of strong approximation solvability and A-proper operators, as can be seen in Appendix A. Our theory is less general since we have focused on a specific linear equation in a Hilbert space context, as opposed to nonlinear equations in a Banach space context. On the other hand, we allow our approximating subspaces to be infinite dimensional, a feature that is essential to be able to interpret the approximate solutions as transition amplitudes for approximate scattering systems.

Consider next the unified nuclear reaction theory of Feshbach.[23–25] Suppose that $\mathfrak{S}$ satisfies Assumption A and that $\{\mathfrak{S}(n)\}$ satisfies Assumption $\Pi^{(n)}$. Let $P^a_{(n)}$ be the orthogonal projection of $\mathscr{H}_N$ onto $\overline{\mathscr{R}}(J\Pi^{(n)a})$, and define $Q^a_{N(n)}\equiv I_N - P^a_{(n)}$. Suppose the inverse of the operator $[z - Q^a_{N(n)}H_NQ^a_{N(n)}]$ exists. Then,

$$\Pi^{(n)a}T\Pi^{(n)a} = \Pi^{(n)a}R^{-1}\{J^*P^a_{(n)}R_NP^a_{(n)}JR^{-1} - J^*J\}\Pi^{(n)a}, \tag{6.3}$$

where[23,37]

$$P^a_{(n)}R_N(z)P^a_{(n)} = [z - P^a_{(n)}H_NP^a_{(n)} - W^a_{N(n)}(z)]^{-1} \tag{6.4}$$

and

$$W^a_{N(n)}(z)\equiv P^a_{(n)}H_NQ^a_{N(n)}[z - Q^a_{N(n)}H_NQ^a_{N(n)}]^{-1} \times Q^a_{N(n)}H_NP^a_{(n)}. \tag{6.5}$$

In the Feshbach theory the interesting resonant structure is generated by "eigenstates" of $Q^a_{N(n)}H_NQ^a_{N(n)}$. The fundamental problem is, then, how to approximate $Q^a_{N(n)}$ in a practical way.

The effect of our scheme is to approximate $Q^a_{N(n)}$ by the operator $Q^a_{(n)}\equiv P_{(n)} - P^a_{(n)}$. Equations (6.3) and (6.4) remain true if T is replaced by $T^{(n)}$, R_N by $R_{(n)}$, and $W^a_{N(n)}$ by $W^a_{(n)}$, where $W^a_{(n)}$ is defined by Eq. (6.5) with $Q^a_{N(n)}$ replaced by $Q^a_{(n)}$. The resonant structure must be generated, therefore, by states in $\mathscr{H}_{(n)}$ that are not in $\overline{\mathscr{R}}(J\Pi^{(n)a})$. Consequently, they are in $\mathscr{R}(J\Pi^{(n)b})$ and hence are themselves generated by states in $\mathscr{H}^{(n)b}$. It is clear that any finite number of resonant states can always be generated in this manner by suitable choice of $\mathscr{H}^{(n)b}$.

The important difference between our scheme and Feshbach theory is that we aim to calculate $T^{(n)}(z)$ directly using the equations of Sec. III B (or V A). Calculation of $P^a_{(n)}$, $Q^a_{(n)}$, or $W^a_{(n)}$ is never necessary. One advantage of our strategy is that we have no difficulty incorporating identical particle symmetries,[30,39] while this has been very troublesome in the past for multichannel Feshbach theories.[24] In addition, the nonorthogonality of asymptotic channel spaces is, in our scheme, incorporated in a simple and natural way.

Another method that does not require compact kernels is the coupled reaction channel (CRC) method. For an application of some of the present approximation ideas to the CRC method see Refs. 33 and 40.

We now turn to the question of the practical implementation of the abstract ideas of this series of papers. We assume that $\mathscr{H}$ has the form[1,30] $\mathscr{H} = \oplus_A \mathscr{H}_A$, that $\Pi = \oplus_A \Pi_A$, and that $J^\pi \oplus_A \psi_A = \Sigma_A \Pi_A \psi_A$. The spaces $\mathscr{H}_A$ are the usual cluster (partition or arrangement channel) spaces. The asymptotic Hamiltonian H has the form $H = \oplus_A H_A$, where the H_A are the usual cluster (partition or arrangement channel) Hamiltonians. The approximate asymptotic Hamiltonian H^π has the form $H^\pi = \oplus_A H^\pi_A$. Then the approximate transition operator [cf. Eqs. (3.6) and (3.12) and Ref. 1, Lemma 2 and Theorem 8]

$$T^\pi(z) = J^{\pi *} V^\pi + V^{\pi *} R_\pi(z) V^\pi, \tag{6.6}$$

$$= J^{\pi *} U^\pi + U^{\pi *} P_\pi R_\pi(z) U^\pi \tag{6.7}$$

has cluster matrix elements

$$T^\pi_{BA}(z) = \Pi_B \{ \overline{U}^\pi_A + \overline{U}^\pi_B P_\pi R_\pi(z) \overline{U}^\pi_A \} \Pi_A , \tag{6.8}$$

where $\overline{U}^\pi_A \equiv H_N - H^\pi_A$. The cluster matrix elements of $M^\pi(z)$ are

$$M^\pi_{BA}(z) = \Pi_B (z - H^\pi_B) \Big(\sum_C \Pi_C \Big)^{-1} P_\pi R_\pi(z) \overline{U}^\pi_A \Pi_A . \tag{6.9}$$

Equation (3.21) then reads [cf. Ref. 30, Eqs. (30) and (36)]

$$M^\pi_{BA}(z) = \Pi_B \overline{U}^\pi_A \Pi_A + \sum_C \Pi_B [\overline{U}^\pi_C R^\pi_C(z) - \bar{\delta}_{BC}] \Pi_C M^\pi_{CA}(z), \tag{6.10}$$

where $R^\pi_C(z) \equiv (z - H^\pi_C)^{-1}$ and $\bar{\delta}_{BC} \equiv 1 - \delta_{BC}$ with δ_{BC} the Kronecker delta.

In a previous paper[30] we sketched how to construct projection operators Π_A such that Assumption Π is satisfied if $\mathscr{H}^a = \mathscr{H}$. Denoting these projection operators by Π^a_A and letting Π^b_A be arbitrary compact projection operators, then $\Pi_A \equiv \Pi^a_A \oplus \Pi^b_A$ also satisfy Assumption Π. This construction also yields that $\Pi_A \Pi_B$, $A \neq B$, are compact and that $\overline{V}_A E_A(\Delta) \Pi^a_A = (H_N - H_A) E_A(\Delta) \Pi^a_A$, where Δ is any finite subinterval of $(-\infty, \infty)$, are trace class. Within this more concrete framework we can make the following comments about Eq. (6.10) and its solution.

(a) The number of coupled integral equations, and the dimensionality of the integrals, compares favorably with other schemes (Ref. 30, Tables 2–4).

(b) The kernel of Eq. (6.10) does not contain the troublesome operator $(J^\pi J^{\pi *})^{-1} P_\pi = (\Sigma_C \Pi_C)^{-1} P_\pi$. Moreover, the kernel of Eq. (6.10) is compact for $\operatorname{Im} z \neq 0$ (Corollary 3.13), thus permitting standard numerical methods to be used for the solution.

(c) The projection operators Π_A are incorporated into the unknown operators $M^\pi_{BA}(z)$ in Eq. (6.10). As a result, the locations of the thresholds are left undisturbed. Only the discontinuities across the cuts are changed.

(d) The operators $\Pi_B \overline{U}^\pi_C \Pi_C$ and $\Pi_B \bar{\delta}_{BC} \Pi_C$ in the kernel of Eq. (6.10) are independent of z. Moreover they have the form of Born approximations and overlap integrals, respectively. Such objects are well known in nuclear theory, at least for two-fragment clusterings.[35]

(e) Any solution $M^\pi_{BA}(z)$ of Eq. (6.10) generates $T^\pi_{BA}(z)$ via

$$T^\pi_{BA}(z) = \Pi_B (z - H^\pi_B) \sum_C (z - H^\pi_C)^{-1} M^\pi_{CA}(z) \tag{6.11}$$

[cf. Eq. (3.17) and Theorem 3.7]. In particular, the solution generated by the Moore–Penrose inverse technique[32] can be used for this purpose.

(f) If the Π_A are so chosen that the null space in $\mathscr{H}^\pi$ of J^π consists only of the zero vector, then Eq. (6.11) is unnecessary. The solution $M^\pi_{BA}(z)$ of Eq. (6.10) is unique and has the same on-shell limit as $T^\pi_{BA}(z)$ (Theorems 3.9 and 3.11).

(g) Equation (6.10) is stable under certain forms of perturbations (cf. Secs. III C and IV C).

(h) The on-shell values of the $T^\pi_{BA}(z)$ are the on-shell transition operators of an approximate scattering system $\mathfrak{S}(\Pi, H^\pi) = \{ \mathscr{H}_\pi, H_\pi, \mathscr{H}^\pi, H^\pi, J^\pi, \Pi^a \}$. The approximate total Hamiltonian H_π is self-adjoint and bounded from below by the lower bound k of H_N. Certain spectral properties of H_π have been obtained by Trucano[22] by proving an analog of the HVZ theorem (Ref. 31, Theorem XIII.17). In particular, let $\sigma(\Pi^a H)$ denote the spectrum of the asymptotic Hamiltonian H restricted to the subspace $\mathscr{H}^{\pi a}$. Then, the spectrum of H_π consists of $\sigma(\Pi^a H)$ and a discrete set of eigenvalues of finite multiplicity in $[k, \inf \sigma(\Pi^a H))$, which can possibly cluster only at $\inf \sigma(\Pi^a H)$. A consequence of Theorem 3.13 of Ref. 16 is that $\sigma(\Pi^a H)$ is equal to the spectrum of the absolutely continuous part of H_π. It also follows from Theorem 4.5 of Ref. 16 and Theorem VIII.24 of Ref. 31 that if λ belongs to the spectrum of H_N, then there exists a sequence $\{\lambda_n\}$, with λ_n belonging to the spectrum of $H_{(n)}$, such that $\lambda_n \to \lambda$.

(i) The approximate wave operators $\Omega^{\pi \pm}_A$ exist for clusterings A with $\mathscr{H}_A$ included in the direct sum $\mathscr{H}^a$ (Theorems 3.11 and 3.13 of Ref. 16). Moreover, they are asymptotically complete so that the scattering operator S^π is unitary on $\mathscr{H}^{\pi a}$ (Theorem 3.13 of Ref. 16 and also the related two-fragment result of Trucano[22] which was announced in Theorem 16 of Ref. 30).

(j) As $n \to \infty$, the approximate Hamiltonians $H_{(n)}$ converge (in the strong resolvent sense) to H_N, and the wave operators $\Omega^{(n)\pm}_A$ converge strongly to $\Omega^\pm_A$ (Theorems 4.5

and 4.6 of Ref. 16). As a consequence the approximate scattering operators converge weakly to the exact scattering operator. In this sense the on-shell approximate transition operators converge to the exact on-shell transition operator. There is also convergence off-shell (Theorems 4.1 and 4.7).

(k) The method can accommodate long-range Coulomb interactions, distorted asymptotic waves, and identical particle symmetries.[30]

Details of this more concrete scheme will be provided in a subsequent publication.

ACKNOWLEDGMENTS

We are indebted to E. O. Alt, H. Baumgärtel, J. M. Combes, F. S. Levin, E. F. Redish, W. Sandhas, E. Zeidler, and numerous other colleagues for their comments. Most especially we are indebted to F. Coester, W. N. Polyzou, I. H. Sloan, and T. Trucano for incisive discussions that have materially contributed to this work. Finally, we thank Betsy Poor for her outstanding typing of this manuscript.

C.C. gratefully acknowledges the generous support of the U.S.–German Fulbright Commission, the Minna–James–Heinemann–Stiftung in collaboration with the NATO Senior Scientists Programme, the German Academic Exchange Service (DAAD), and the University of Bonn during the beginning phases of this work. A.G.G. gratefully acknowledges the financial support of DGRST and CNRS in France during the beginning stages of this work. C.C. and A.G.G. gratefully acknowledge the financial support of a grant from the National Science Foundation (PHY-8103062) during the writing of this manuscript.

APPENDIX A: PETRYSHYN *A*-PROPER OPERATORS

In this Appendix our T- and M-operator equations are placed within the context of the theory of A-proper (approximation-proper) operators and strong approximation solvability. This theory has been reviewed by Petryshyn.[26] The theory is very general (cf. Ref. 26 and the 163 references cited therein) and subsumes Fredholm operator theory, ball condensing theory, monotone and accretive operator theory, Petrov–Galerkin methods, and others.

1. The abstract theory

Let $\mathcal{M}$ and $\mathcal{N}$ be Banach spaces, let $\mathcal{D}$ be a given subset of $\mathcal{M}$, and let $L: \mathcal{D} \subset \mathcal{M} \rightarrow \mathcal{N}$ be a (possibly unbounded) linear mapping.

Definition A 1: $\Gamma \equiv \{\mathcal{M}(n), \mathcal{N}(n), \Xi(n), Q(n), \Pi(n)\}$ is called a *suitable approximation scheme* for the operator L if

(i) $\{\mathcal{M}(n)\}$ and $\{\mathcal{N}(n)\}$ are two sequences of Banach spaces;

(ii) $\Xi(n): \mathcal{M} \rightarrow \mathcal{M}(n)$, $Q(n): \mathcal{M}(n) \rightarrow \mathcal{M}$, and $\Pi(n): \mathcal{N} \rightarrow \mathcal{N}(n)$ are three sequences of linear mappings;

(iii) $Q(n)$ and $\Pi(n)$ are uniformly bounded; and

(iv) $Q(n)\Xi(n)\Phi \rightarrow \Phi$ for each $\Phi \in \mathcal{M}$.

Let $\{L(n)\}$ be the sequence of mappings of $\mathcal{D}(n) \equiv \{\Phi \in \mathcal{M}(n) | Q(n)\Phi \in \mathcal{D}\}$ into $\mathcal{N}(n)$ defined by

$$L(n) \equiv \Pi(n) L Q(n). \tag{A1}$$

Then, an approximation to the exact equation

$$L\Phi = \Psi \tag{A2}$$

$(\Phi \in \mathcal{D}, \Psi \in \mathcal{N})$ is

$$L(n)\Phi(n) = \Pi(n)\Psi \tag{A3}$$

$[\Pi(n)\Psi \in \mathcal{N}(n), \Phi(n) \in \mathcal{D}(n)]$. Graphically the approximation scheme is the following:

$$\begin{tikzcd} \mathcal{M} \arrow[d, shift right, "\Xi(n)"'] & \supset \mathcal{D} \arrow[r, "L"] & \mathcal{N} \arrow[d, "\Pi(n)"] \\ \mathcal{M}(n) \arrow[u, shift right, "Q(n)"'] & \supset \mathcal{D}(n) \arrow[r, "L(n)"] & \mathcal{N}(n) \end{tikzcd}$$

Remark A 2: Our definition of a suitable approximation scheme is similar to Petryshyn's definition[27] of an admissible approximation scheme but differs in the following respects. First, Petryshyn's definition is more general in the sense that it allows $L, \Xi(n), Q(n)$, and $\Pi(n)$ to be nonlinear mappings in a Banach space. Since our applications are to linear operators on a Hilbert space, we state only the simpler case. Second, our definition is more general in the sense that we do not always assume that $\mathcal{M}(n)$ and $\mathcal{N}(n)$ are finite dimensional. For theoretical purposes we want to allow some intermediate stage approximations which, so that approximate wave operators will exist, may not be finite dimensional.

Definition A 3: (cf. Ref. 27). L is said to be *A proper* with respect to the suitable approximation scheme Γ if for any bounded sequence $\{\Phi(n) | \Phi(n) \in \mathcal{D}(n)\}$ with $Q(n)\Phi(n)$ in $\mathcal{D}$ such that $\|L(n)\Phi(n) - \Pi(n)\Psi\| \rightarrow 0$ for some $\Psi \in \mathcal{N}$, then there exists an infinite subsequence $\{\Phi(n_k)\}$ and a $\Phi \in \mathcal{D}$ such that $\|Q(n_k)\Phi(n_k) - \Phi\| \rightarrow 0$ as $k \rightarrow \infty$ and $L\Phi = \Psi$.

The main reason that one wants to show that L is an A-proper operator is because it is a sufficient condition for Eq. (A2) to be strongly approximation solvable.[26]

Definition A 4: (cf. Ref. 27). For a given $\Psi \in \mathcal{N}$, Eq. (A2) is said to be *strongly approximation solvable* with respect to a suitable approximation scheme Γ if there is an integer $n_0 \geqslant 1$ such that the approximate Eq. (A3) has a solution $\Phi(n) \in \mathcal{D}(n)$ for each $n \geqslant n_0$ which are such that $\|Q(n)\Phi(n) - \Phi\| \rightarrow 0$ as $n \rightarrow \infty$ for some $\Phi \in \mathcal{D}$ satisfying Eq. (A2).

An operator L is known to be A proper in several interesting cases.[26] In particular, if $L = I - C$, where $C: \mathcal{M} \rightarrow \mathcal{M}$ is a compact operator, then L is A proper with respect to the scheme $\Gamma_1 \equiv \{\mathcal{M}(n), \mathcal{N}(n), \Xi(n), I, \Pi(n)\}$, where $\Xi(n)$: $\mathcal{M} \rightarrow \mathcal{M}(n)$ and $\Pi(n): \mathcal{M} \rightarrow \mathcal{N}(n)$ are sequences of projection operators satisfying $\Xi(n)\Phi \rightarrow \Phi$ and $\Pi(n)\Phi \rightarrow \Phi$ for each $\Phi \in \mathcal{M}$. In this case Eq. (A2) is a Fredholm operator equation and its approximation solvability with respect to Γ_1 is the well-known Petrov–Galerkin method for obtaining a converging sequence of approximate solutions. It is not so well known that these methods generalize to certain other noncompact kernel equations. For example, if $L = I - K$ with $\|K\| < 1$, or if L is a positive-definite operator on a Hilbert space, then Eq. (A2) is strongly approximation solvable with respect to $\Gamma_0 \equiv \{\mathcal{M}(n), \mathcal{M}(n), \Pi(n), I, \Pi(n)\}$, where $\mathcal{M} = \mathcal{N}$ and $\Pi(n): \mathcal{M} \rightarrow \mathcal{M}(n)$ is a sequence of projection operators satisfying $\Pi(n)\Phi \rightarrow \Phi$ for each $\Phi \in \mathcal{M}$.

2. Application to scattering theory

In order to put our T- and M-operator equations within the context of the abstract theory, we choose the spaces and operators of Sec. A1 in the following way. Let $\mathscr{N} \equiv P\mathscr{H}$, where $\mathscr{H}$ is the Hilbert space given in the exact scattering system $\mathfrak{S}$, and P is the orthogonal projection onto $\overline{\mathscr{R}}(J^*)$. For a fixed $z\in\rho(H_N)$, let $\mathscr{D} = \mathscr{D}(z) \equiv (z-H)PR(z)\mathscr{H}$, and for Φ, $\Psi\in\mathscr{D}$, define $(\Phi,\Psi)_{\mathscr{M}} \equiv (PR(z)\Phi, PR(z)\Psi)$. It is easy to verify that $\mathscr{D}$ is a linear vector space and $(\Phi,\Psi)_{\mathscr{M}}$ is an inner product on $\mathscr{D}$. Define the Hilbert space $\mathscr{M}$ to be the completion of $\mathscr{D}$ under the inner product $(\cdot,\cdot)_{\mathscr{M}}$. Let $\mathscr{D}(n) = \mathscr{M}(n) \equiv \mathscr{H}^{(n)}$, where $\mathscr{H}^{(n)} \equiv \Pi^{(n)}\mathscr{H}$ and $\{\Pi^{(n)}\}$ is the sequence of orthogonal projection operators in $\{\mathfrak{S}(n)\}$. Let $\mathscr{N}(n) \equiv \mathscr{N}^{(n)}$, where $\mathscr{N}^{(n)} \equiv \overline{\mathscr{R}}(\Pi^{(n)}\mathscr{N})$.

Define two L operators mapping $\mathscr{D}\subset\mathscr{M}$ into $\mathscr{N}$ by

$$L^T = L^T(z) \equiv I - V^*(JJ^*)^{-1}JR(z) \tag{A4}$$

and

$$L^M = L^M(z) \equiv J^*[J - VR(z)]\,. \tag{A5}$$

Define the sequence $\Pi(n)$: $\mathscr{N}\to\mathscr{N}^{(n)}$ by

$$\Pi(n) \equiv \Pi^{(n)}\,. \tag{A6}$$

Define two sequences of $Q(n)$-operators mapping $\mathscr{H}^{(n)}$ into $\mathscr{D}\subset\mathscr{M}$ by

$$Q^{T(n)} = Q^{T(n)}(z) \equiv (z-H)J^*J^{(n)}R^{(n)}(z) \tag{A7}$$

and

$$Q^{M(n)} = Q^{M(n)}(z) \equiv (z-H)J^*(JJ^*)^{-1}J^{(n)}R^{(n)}(z)\,. \tag{A8}$$

Finally, define two sequences of $\Xi(n)$-operators mapping $\mathscr{D}\subset\mathscr{M}$ into $\mathscr{H}^{(n)}$ by

$$\Xi^{T(n)} = \Xi^{T(n)}(z) \equiv (z-H^{(n)})J^{(n)*}Z^{-1}(JJ^*)^{-1}JR(z) \tag{A9}$$

and

$$\Xi^{M(n)} = \Xi^{M(n)}(z) \equiv (z-H^{(n)})J^{(n)*}Z^{-1}JR(z)\,, \tag{A10}$$

where $Z \equiv J\Pi^{(\infty)}J^*$.

The following lemma establishes some formulas for the norms of operators which map from and/or to the Hilbert space $\mathscr{M}$. The norm of an operator which maps from $\mathscr{H}$ into $\mathscr{H}$ will be denoted by $\|\cdot\|$ and other operator norms will be distinguished by appropriate subscripts.

Lemma A 5: Let $z\in\rho(H_N)$ be fixed, and let $F = F(z)$.

(a) If F: $\mathscr{H}\to\mathscr{M}$, then $\|F\|_{\mathscr{H},\mathscr{M}} = \|PRF\|$.

(b) If F: $\mathscr{M}\to\mathscr{H}$, then $\|F\|_{\mathscr{M},\mathscr{H}} = \|FR^{-1}P\|$.

(c) If F: $\mathscr{M}\to\mathscr{M}$, then $\|F\|_{\mathscr{M},\mathscr{M}} = \|PRFR^{-1}P\|$.

Proof: If Φ belongs to the dense subset $\mathscr{D}\subset\mathscr{M}$, then $R^{-1}PR\Phi = \Phi$. Thus

$$\|F\|_{\mathscr{M},\mathscr{M}} \equiv \sup_{\Phi\in\mathscr{M}} \frac{\|F\Phi\|_{\mathscr{M}}}{\|\Phi\|_{\mathscr{M}}} = \sup_{\Phi\in\mathscr{M}} \frac{\|PRFR^{-1}PR\Phi\|}{\|PR\Phi\|} \leqslant \|PRFR^{-1}P\|\,. \tag{A11}$$

On the other hand, $\Phi = R^{-1}P\Psi\in\mathscr{D}$ for all $\Psi\in\mathscr{H}$. Therefore,

$$\|F\|_{\mathscr{M},\mathscr{M}} \geqslant \sup_{\Psi\in\mathscr{H}} \frac{\|PRFR^{-1}P\Psi\|}{\|P\Psi\|} \geqslant \sup_{\Psi\in\mathscr{H}} \frac{\|PRFR^{-1}P\Psi\|}{\|\Psi\|} = \|PRFR^{-1}P\|\,. \tag{A12}$$

This proves (c). The proofs of (a) and (b) are similar. □

Theorem A6: Let $\mathfrak{S}$ satisfy Assumptions (A1)–(A4), and let $\{\mathfrak{S}(n)\}$ satisfy Assumptions ($\Pi^{(n)}$1)–($\Pi^{(n)}$3). Let $z\in\rho(H_N)$ be fixed. Then

(i) $\Gamma^T \equiv \{\mathscr{H}^{(n)}, \mathscr{N}^{(n)}, \Xi^{T(n)}, Q^{T(n)}, \Pi^{(n)}\}$ is a suitable approximation scheme for the operator L^T, and

(ii) $\Gamma^M \equiv \{\mathscr{H}^{(n)}, \mathscr{N}^{(n)}, \Xi^{M(n)}, Q^{M(n)}, \Pi^{(n)}\}$ is a suitable approximation scheme for the operator L^M.

Proof: $\{\Pi^{(n)}\}$ is uniformly bounded by 1. By Lemma A5(a) and Eq. (A7),

$$\|Q^{T(n)}\|_{\mathscr{H}^{(n)},\mathscr{M}} = \|PJ^*J\Pi^{(n)}R^{(n)}\| \leqslant |\mathrm{Im}\, z|^{-1}\|J^*J\|\,. \tag{A13}$$

This proves that $\{Q^{T(n)}\}$ is uniformly bounded. Similarly, $\{Q^{M(n)}\}$ is uniformly bounded in norm by $|\mathrm{Im}\, z|^{-1}$. Now, by Assumption ($\Pi^{(n)}$3),

$$\|Q^{T(n)}\Xi^{T(n)}\Phi - \Phi\|_{\mathscr{M}} = \|J^*J[\Pi^{(n)} - \Pi^{(\infty)}]J^*Z^{-1}(JJ^*)^{-1}JR\Phi\| \tag{A14}$$

converges to zero for all $\Phi\in\mathscr{M}$. Similarly,

$$\|Q^{M(n)}\Xi^{M(n)}\Phi - \Phi\|_{\mathscr{M}} = \|P[\Pi^{(n)} - \Pi^{(\infty)}]J^*Z^{-1}JR\Phi\| \tag{A15}$$

converges to zero for all $\Phi\in\mathscr{M}$. □

Let $z\in\rho(H_N)$ and $\Psi\in\mathscr{D}(H)\subset\mathscr{H}$. Then $J^*V\Psi\in\mathscr{N}$. The exact transition-operator equation (2.10) may be written in the form

$$L^T\Phi = J^*V\Psi \tag{A16}$$

with solution $\Phi = \Phi(z) = T(z)\Psi$ in $\mathscr{D}$. The exact M-operator equation (5.7) may be written in the form

$$L^M\Phi = J^*V\Psi \tag{A17}$$

with solution $\Phi = \Phi(z) = M(z)\Psi$ in $\mathscr{D}$.

Define the sequence of operators $L^{(n)}$: $\mathscr{H}^{(n)}\to\mathscr{H}^{(n)}$ by

$$L^{(n)} = L^{(n)}(z) \equiv J^{(n)*}(z - H_{(n)})J^{(n)}R^{(n)}\,. \tag{A18}$$

It is easy to verify that

$$L^{(n)} = \Pi^{(n)}L^TQ^{T(n)} = \Pi^{(n)}L^MQ^{M(n)}\,. \tag{A19}$$

Therefore, the approximate equation [cf. Eq. (A3)] associated with either Eq. (A16) or Eq. (A17) is

$$L^{(n)}\Phi^{(n)} = \Pi^{(n)}J^*V\Psi\,. \tag{A20}$$

Theorem A7: Let $\mathfrak{S}$ satisfy Assumptions (A1)–(A4), and let $\{\mathfrak{S}(n)\}$ satisfy Assumption $\Pi^{(n)}$. Then, for all $\Psi\in\mathscr{D}(H)$ and all z such that $\mathrm{Im}\, z\neq 0$:

(i) Eq. (A16) is strongly approximation solvable with respect to Γ^T; and

(ii) Eq. (A17) is strongly approximation solvable with respect to Γ^M.

Proof: Substitution of

$$\Phi_0^{(n)} \equiv (z - H^{(n)})J^{(n)*}(J^{(n)}J^{(n)*})^{-1}R_{(n)}(z)P_{(n)}V\Psi \tag{A21}$$

into Eq. (A20) verifies that $\Phi_0^{(n)}$ is a solution in $\mathscr{H}^{(n)}$ of Eq. (A20). Similarly,

$$\Phi^T \equiv (z-H)J^*R_N(z)V\Psi \tag{A22}$$

is a solution in $\mathscr{D}$ of Eq. (A16), and

$$\Phi^M \equiv (z-H)J^*(JJ^*)^{-1}R_N(z)V\Psi \tag{A23}$$

is a solution in $\mathscr{D}$ of Eq. (A17). Furthermore,

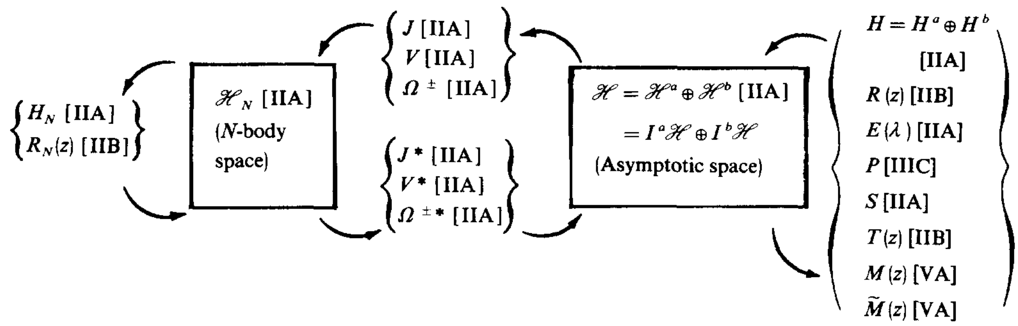


FIG. 1. Exact scattering system $\mathfrak{S}$.

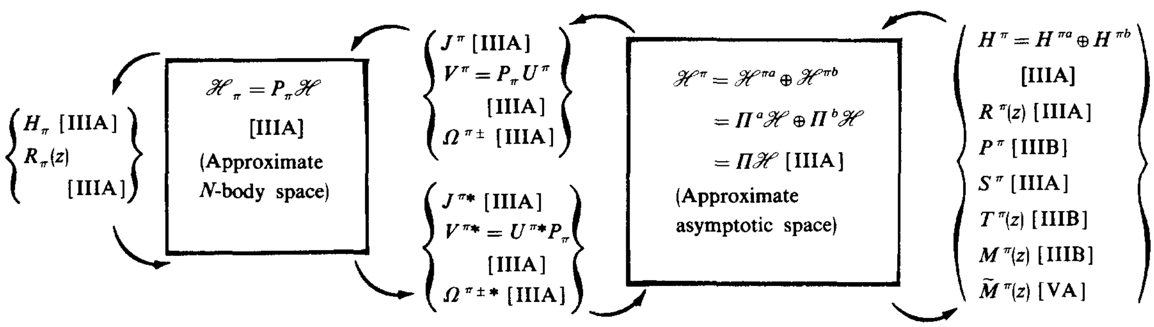


FIG. 2. Approximate scattering system $\mathfrak{S}(\Pi, H^\pi)$.

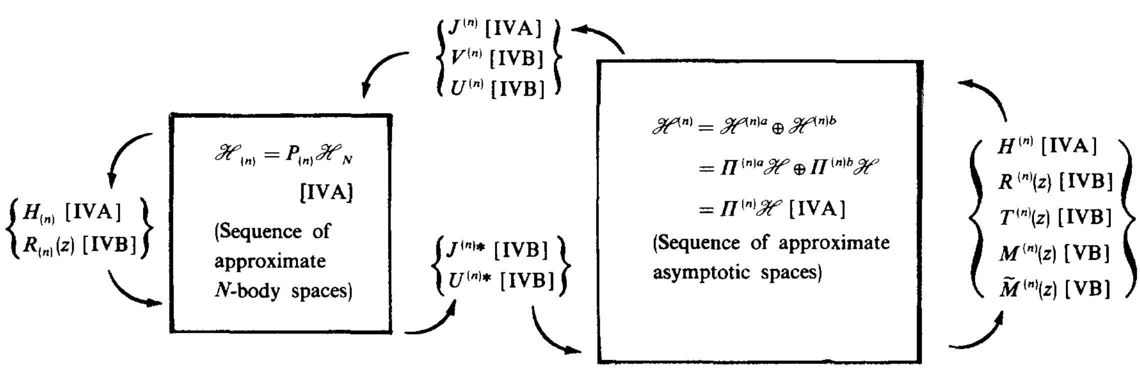


FIG. 3. Sequence $\{\mathfrak{S}(n)\}$ of approximate scattering systems. [The notation is the same as that of Fig. 2 with subscripts and superscripts (n) used in place of the more cumbersome $\Pi^{(n)}$ that would be demanded by the previous notation.]

$$\|Q^{T(n)}\Phi_0^{(n)} - \Phi^T\|_{\mathscr{M}} = \|J^*\{R_{(n)}P_{(n)} - R_N\}V\Psi\|, \quad \text{(A24)}$$

and the right side of Eq. (A24) converges to zero as $n\to\infty$ by Theorem 4.5 and Eq. (4.8) of Ref. 16. Similarly

$$\|Q^{M(n)}\Phi_0^{(n)} - \Phi^M\|_{\mathscr{M}} = \|J^*(JJ^*)^{-1}\{R_{(n)}P_{(n)} - R_N\}V\Psi\| \quad \text{(A25)}$$

converges to zero as $n\to\infty$. □

Remark A 8: (a) Theorem A7 contains as a special case (take $\mathscr{H}^a = \mathscr{H}$ and $H^{(n)} = H$) the result announced in Theorem 7 of Ref. 30.

(b) Suppose that $\Psi\in\mathscr{D}(H)\cap\mathscr{H}^{(n)a}$. Then $P_{(n)}V\Psi = V^{(n)}\Psi$ and $\Phi_0^{(n)}(z) = M^{(n)}(z)\Psi$ [cf. Eqs. (A21) and (5.15)]. In this case, Theorem A7(ii) and Eq. (5.19) of Theorem 5.6 are the same result. In addition, $\Pi^{(n)a}Q^{T(n)}(z)M^{(n)}(z)\Psi = I^aT^{(n)}(z)\Psi$ [cf. Eqs. (A7), (5.15), and (4.3)]. It follows that Theorem A7(i) and Eq. (4.5) of Theorem 4.1 correspond when the input and output states are restricted to $\mathscr{H}^{(n)a}$. The restriction to $\mathscr{H}^{(n)a}$ is necessary for the approximate wave operators $\Omega^{(n)\pm}$, and hence for the approximate scattering operators $S^{(n)}$, to be nonzero [cf. Eqs. (4.9)–(4.11) of Ref. 16].

In this paper we have proved approximation solvability directly rather than prove that the operators L^T and L^M are Petryshyn A proper. Consequently, the characterizations of A-proper operators in Refs. 26 and 27 have not been directly useful to us. We end this Appendix, however, by showing that, under the additional Condition Γ, the operators L^T and L^M are Petryshyn A proper and one-to-one.

Theorem A9: Let the assumptions of Theorem A7 hold. In addition, suppose that $\{\Pi^{(n)}\}$ satisfies Condition Γ. Then the operators L^T and L^M are Petryshyn A proper and one-to-one.

Proof: Let $\Psi\in\mathscr{D}(H)$ and $\operatorname{Im} z\neq 0$. Let $\{\Phi^{(n)}\}$ be a bounded sequence in $\mathscr{H}^{(n)}$ with $\|\Theta^{(n)}\|\to 0$ as $n\to\infty$, where

$$\Theta^{(n)} \equiv L^{(n)}\Phi^{(n)} - \Pi^{(n)}J^*V\Psi \quad \text{(A26)}$$

with $L^{(n)}$ given by Eqs. (A18) and (A19). Let $\Phi_0^{(n)}$, Φ^T, and Φ^M be defined as in Eqs. (A21)–(A23). Then

$$\|Q^{T(n)}\Phi^{(n)} - \Phi^T\|_{\mathscr{M}} \leqslant \|Q^{T(n)}[\Phi^{(n)} - \Phi_0^{(n)}]\|_{\mathscr{M}} + \|Q^{T(n)}\Phi_0^{(n)} - \Phi^T\|_{\mathscr{M}}. \quad \text{(A27)}$$

The first term on the right side of Eq. (A27) satisfies

$$\|Q^{T(n)}[\Phi^{(n)} - \Phi_0^{(n)}]\|_{\mathscr{M}} = \|J^*R_{(n)}(J^{(n)}J^{(n)*})^{-1}P_{(n)}J\Theta^{(n)}\| \leqslant c_0^{-1}|\operatorname{Im} z|^{-1}\|J\|\,\|J^*\|\,\|\Theta^{(n)}\| \quad \text{(A28)}$$

and converges to zero as $n\to\infty$. The second term on the right side of Eq. (A27) converges to zero by Theorem A7. It follows that L^T is A proper. Similarly, L^M is A proper since

$$\|Q^{M(n)}\Phi^{(n)} - \Phi^M\|_{\mathscr{M}} \leqslant \|J^*(JJ^*)^{-1}R_{(n)}(J^{(n)}J^{(n)*})^{-1}P_{(n)}J\Theta^{(n)}\| + \|Q^{M(n)}\Phi_0^{(n)} - \Phi^M\|_{\mathscr{M}} \quad \text{(A29)}$$

converges to zero as $n\to\infty$.

The operator L^T: $\mathscr{D}\subset\mathscr{M}\to\mathscr{N}$ is one-to-one by the uniqueness of the solution of Eq. (2.10) in Theorem 2.2. The operator L^M: $\mathscr{D}\subset\mathscr{M}\to\mathscr{N}$ is one-to-one by the uniqueness of the solution of Eq. (5.7) in Theorem 5.5. □

APPENDIX B: INDEX OF MAJOR OPERATORS AND HILBERT SPACES

The major operators and Hilbert spaces used in this paper are shown in the Figs. 1–3. Here [#] indicates the section number where the description of the symbol is found. The arrows indicate the domain and range spaces (with unbounded operators defined on only a subspace).

[1] C. Chandler and A. G. Gibson, "N-body quantum scattering theory in two Hilbert spaces. I. The basic equations," J. Math. Phys. **18**, 2336–2347 (1977).
[2] C. Chandler and A. G. Gibson, "Some N-body transition operator equations containing channel projection operators," in *Atomic Scattering Theory*, edited by J. Nuttall (University of Western Ontario, London, Ontario, 1978), pp. 189–193.
[3] W. N. Polyzou, A. G. Gibson, and C. Chandler, "Connected Chandler–Gibson equations and few body collisions," Phys. Rev. C **26**, 1878–1892 (1982).
[4] K. L. Kowalski, "N-body systems," in *Few Body Systems and Nuclear Forces II, Lecture Notes in Physics*, Vol. 87, edited by H. Zingl *et al.* (Springer, New York, 1978), pp. 393–413.
[5] F. S. Levin, "Some recent developments in n-particle scattering theory," Nucl. Phys. A **353**, 143c–156c (1981).
[6] Gy. Bencze, "Combinatorial problems in N-particle scattering," Phys. Lett. B **72**, 155–158 (1977).
[7] B. R. Karlsson, "A combinatoric result related to the N-body problem," J. Math. Phys. **23**, 421–422 (1982).
[8] G. A. Hagedorn, "Asymptotic completeness for classes of two, three, and four particle Schrödinger operators," Trans. Amer. Math. Soc. **258**, 1–75 (1980).
[9] K. Hepp, "On the quantum mechanical N-body problem," Helv. Phys. Acta **42**, 425–458 (1969).
[10] F. Riahi, "On the analyticity properties of the N-body scattering amplitude in nonrelativistic quantum mechanics," Helv. Phys. Acta **42**, 299–329 (1969).
[11] A. Schtalheim, "Multi-particle quantum systems below the four-particle threshold," Helv. Phys. Acta **44**, 642–661 (1971).
[12] I. M. Sigal, "Scattering Theory for Many Particle Systems. I and II," Princeton University preprints, 1979.
[13] W. O. Amrein, D. B. Pearson, and K. B. Sinha, "Bounds on the total scattering cross-section for N-body systems," Nuovo Cimento A **52**, 115–131 (1979).
[14] T. A. Osborn and D. Bolle, "Integral bounds for N-body total cross sections," J. Math. Phys. **20**, 1059–1063 (1979).
[15] B. Simon, "Geometric methods in multiparticle quantum systems," Comm. Math. Phys. **55**, 259–274 (1977).
[16] C. Chandler and A. G. Gibson, "N-body quantum scattering theory in two Hilbert spaces. III. Theory of approximations," J. Funct. Anal. **52**, 80–105 (1983).
[17] C. Chandler and A. G. Gibson, "Transition from time-dependent to time-independent multichannel quantum scattering theory," J. Math. Phys. **14**, 1328–1335 (1973).
[18] C. Chandler and A. G. Gibson, "N-body quantum scattering theory in two Hilbert spaces. II. Some asymptotic limits," J. Math. Phys. **19**, 1610–1616 (1978).
[19] J. Ginibre and M. Moulin, "Hilbert space approach to the quantum mechanical three-body problem," Ann. Inst. H. Poincaré **21**, 97–145 (1974).
[20] K. Gustafson, "Scattering theory and the principle of limiting absorption," in *Proceedings of the International Conference on Operator Algebras, Ideals, and Their Applications in Theoretical Physics*, edited by H. Baumgärtel *et al.* (Teubner-Texte, Leipzig, 1978), pp. 335–349.
[21] J. S. Howland, "Abstract stationary theory of multichannel scattering," J. Funct. Anal. **22**, 250–282 (1976).
[22] T. G. Trucano, "Asymptotic Completeness in an Approximate Model of Quantum Mechanical Scattering Theory," Ph.D. Dissertation, University of New Mexico, 1980.
[23] H. Feshbach, "A unified theory of nuclear reactions. I and II," Ann. Phys. (NY) **5**, 357–390 (1958); **19**, 287–313 (1962).
[24] P. Grassberger and W. Sandhas, "Relation between the projection operator formalism and the Faddeev theory," Z. Physik **220**, 29–44 (1969).
[25] W. M. MacDonald, "A unified nuclear reaction theory for phenomenolo-

gical analyses," Nucl. Phys. **54**, 393–404 (1963); "Antisymmetrization of a unified nuclear reaction theory," Nucl. Phys. **56**, 636–646 (1964).

[26]W. V. Petryshyn, "On the approximation-solvability of equations involving A-proper and pseudo-A-proper mappings," Bull. Am. Math. Soc. **81**, 223–312 (1975).

[27]W. V. Petryshyn, "Nonlinear equations involving noncompact operators," in *Proceedings of the Symposium on Pure Mathematics*, Vol. 18, Part 1 (Am. Math. Soc., Providence, RI, 1970), pp. 206–233.

[28]W. O. Amrein, J. M. Jauch, and K. B. Sinha, *Scattering Theory in Quantum Mechanics* (Benjamin, Reading, MA, 1977).

[29]C. Chandler and A. G. Gibson, "Two Hilbert space scattering theory: Recent progress," in *Few Body Systems and Nuclear Forces I, Lecture Notes in Physics*, Vol. 82, edited by H. Zingl *et al.* (Springer, New York, 1978), pp. 356–359.

[30]C. Chandler and A. G. Gibson, "A two-Hilbert-space formulation of multichannel scattering theory," in *Mathematical Methods and Applications of Scattering Theory, Lecture Notes in Physics*, Vol. 130, edited by J. A. DeSanto *et al.* (Springer, New York, 1980), pp. 134–148.

[31]M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vol I, *Functional Analysis*; Vol. II, *Fourier Analysis, Self-Adjointness*; Vol. III, *Scattering Theory*, Vol. IV, *Analysis of Operators* (Academic, New York, 1972, 1975, 1979, 1978).

[32]*Generalized Inverses and Applications*, edited by M. Z. Nashed (Academic, New York, 1976).

[33]M. C. Birse and E. F. Redish, "Non-orthogonality and overcompleteness in direct nuclear reactions," University of Maryland, preprint, December 1982.

[34]W. N. Polyzou (private communication).

[35]S. R. Cotanch and C. M. Vincent, "Channel coupling and nonorthogonality in elastic and transfer processes," Phys. Rev. C **14**, 1739–1762 (1976).

[36]I. H. Sloan (private communication).

[37]F. Coester and L. Schlessinger, "Unstable states produced in multiparticle reactions," Ann. Phys. (NY) **78**, 90–131 (1973).

[38]K. O. Friedrichs, "On the perturbation of continuous spectra," Comm. Pure Appl. Math. **1**, 361–406 (1948).

[39]Gy. Bencze and C. Chandler, "Time-dependent scattering theory for identical particles," Phys. Rev. C **25**, 136–145 (1982).

[40]Gy. Bencze, C. Chandler, and A. G. Gibson, "Multiparticle scattering theory and the method of coupled reaction channels," Nucl. Phys. A **390**, 461–485 (1982).

Velocity and density of a two-dimensional acoustic medium from point source surface data

Shimon Coen
Department of Materials Science and Mineral Engineering, University of California, Berkeley, California 94720

Margaret Cheney
Department of Mathematics, Stanford University, Stanford, California 94305

Arthur Weglein
Geophysical Research, Sohio Petroleum Company, 1 Lincoln Center, 5400 LBJ Freeway, Dallas, Texas 75240

(Received 8 November 1983; accepted for publication 30 December 1983)

An inverse acoustic scattering theory and algorithm is presented for the reconstruction of a two-dimensional inhomogeneous acoustic medium from surface measurements. The measurements of the surface pressure due to a harmonically oscillating surface point source at two arbitrary frequencies allows the separate reconstruction of the density and velocity of the subsurface. This is a first step towards solving the inverse problem of exploration geophyiscs.

PACS numbers: 03.65.Nk, 91.30.Rz, 43.20. + g

I. INTRODUCTION

The problem we consider is that of reconstructing the subsurface velocity and density from surface measurements of pressure waves emanating from an oscillating point source on the surface. A natural approach to this problem is to transform the acoustic equation into the Schrödinger equation in order to take advantage of the great body of knowledge on inverse scattering. Ware and Aki[1] transformed the one-dimensional elastodynamics equation into the Schrödinger equation. Coen[2] transformed various inverse problems for a one-dimensional acoustic or elastic medium with plane waves, line or point sources into the one-dimensional Schrödinger equation whose potential is real and energy independent.

There are, in addition, a number of multidimensional inverse scattering methods. These have suggested the possibility of applying inverse scattering techniques to the multidimensional acoustic equation; however, for a number of years attempts to transform the higher-dimensional acoustic equation into the Schrödinger equation have been unsuccessful. This paper overcomes this difficulty and goes on to apply state-of-the-art inverse scattering techniques to the three-dimensional acoustic equation with a two-dimensional medium.

We transform the three-dimensional acoustic inverse scattering problem for a two-dimensional acoustic medium into an equivalent inverse quantum scattering problem in two dimensions with a potential which is real and has compact support. The solution procedure contains four parts: (1) Transformation of the acoustic equation to the two-dimensional Schrödinger equation. (2) Transformation of the acoustic data into equivalent quantum scattering data—the scattering amplitude. (3) Reconstruction of the potential from the scattering amplitude. (4) Reconstruction of the velocity and density of the acoustic medium from the potential.

II. TRANSFORMATION OF THE ACOUSTIC EQUATION INTO THE SCHRÖDINGER EQUATION

The pressure $p(\mathbf{r},\mathbf{r}_s,\omega)$ in a fluid with density $\rho(\mathbf{r})$ and local wave velocity $c(\mathbf{r})$, due to a harmonically oscillating point source at $\mathbf{r}_s$ with frequency ω, is governed by the inhomogeneous acoustic equation

$$(\nabla^2 - \rho^{-1}\nabla\rho\cdot\nabla + \omega^2 c^{-2})p(\mathbf{r},\mathbf{r}_s,\omega) = \delta(\mathbf{r}-\mathbf{r}_s). \qquad (1)$$

Let (x, y, z) denote the Cartesian coordinates of $\mathbf{r}$. We assume that:

(I) The density and local wave velocity are uniform in the y direction, so that $\rho = \rho(x,z)$ and $c = c(x,z)$.

(II) The density and velocity are uniform exterior to the circle $x^2 + z^2 = a^2$ so that $\rho = \rho_0$ and $c = c_0$ for $x^2 + z^2 \geqslant a^2$, a being a finite constant.

(III) The exterior velocity c_0 satisfies $c_0 \leqslant c(x,z)$ for all x and z. (This condition excludes bound states. See Appendix A.)

We also assume either of the two following geometries:

(IV) The point sources are placed on a circle at positions $\mathbf{r}_s = (x_s,0,z_s)$ with $x_s^2 + z_s^2 = a^2$.

(IV′) The point sources are placed on two lines at positions $\mathbf{r}_s = (x_s,0,a)$ and $\mathbf{r}_s = (x_s,0,-a)$.

If we define $\mathbf{r}_t = (x,0,z)$ and $g(\mathbf{r}_t) = [\rho_0/\rho(\mathbf{r}_t)]^{1/2}$ then

$$\Phi(\mathbf{r}_t,\mathbf{r}_s,k,\omega) = g(\mathbf{r}_t)\int_{-\infty}^{\infty} p(\mathbf{r},\mathbf{r}_s,\omega)e^{-iky}dy \qquad (2)$$

transforms Eq. (1) into

$$(\nabla_t^2 + \xi^2)\psi(\mathbf{r}_t,\mathbf{r}_s,\xi) = V(\mathbf{r}_t;\omega)\psi(\mathbf{r}_t,\mathbf{r}_s,\xi) + \delta(\mathbf{r}_t - \mathbf{r}_s), \qquad (3)$$

where $\xi^2 = \omega^2/c_0^2 - k^2$, $\psi(\mathbf{r}_t,\mathbf{r}_s,\xi) = \phi(\mathbf{r}_t,\mathbf{r}_s,k,\omega)$,

$$V(\mathbf{r}_t;\omega) = \omega^2 c_0^{-2}[1 - c_0^2 c^{-2}(\mathbf{r}_t)] + g^{-1}(\mathbf{r}_t)\nabla_t^2 g(\mathbf{r}_t), \qquad (4)$$

and $\nabla_t^2 = \partial^2/\partial x^2 + \partial^2/\partial z^2$ is the transverse Laplacian. Equation (3) for $\mathbf{r}_t \neq \mathbf{r}_s$ is a two-dimensional Schrödinger equation where ξ plays the role of energy and V plays the role of potential which in this case is energy independent with ω being a fixed parameter. We assume that the velocity and density are such that V is square integrable.

To obtain data for the inverse problem under Assumption IV, we measure the pressure at the boundary of the cylinder of radius a, thus obtaining $p(\mathbf{r},\mathbf{r}_s,\omega)$ for all $\mathbf{r}$ of the form $\mathbf{r} = (x, y,z)$ with $x^2 + z^2 = a^2$ and for all $\mathbf{r}_s$ satisfying Assumption IV above. To obtain data under Assumption IV′, we measure the pressure at $\mathbf{r} = (x, y,a)$ for sources at

$(x_s,0,a)$ and $(x_s,0,-a)$ and at $\mathbf{r}=(x,y,-a)$ for sources at $(x_s,0,-a)$. In either case the transformation (2) transforms this data into $\psi(\mathbf{r}_t,\mathbf{r}_s,\xi)$ for Eq. (3), which is known for all $\mathbf{r}_t$ and $\mathbf{r}_s$ as above and for ξ of the form $\xi^2=\omega^2/c_0^2-k^2$ for all real k. In the ξ variable, we therefore have data on the positive imaginary axis and on the interval $(0,\omega/c_0)$. Since the problem is now wholly two dimensional, we shall drop the subscripts "t" and write $\mathbf{r}=(x,z)$. We will use carets to denote the two-dimensional unit vectors.

III. TRANSFORMATION TO PLANE-WAVE INCIDENCE

Equation (3) may be cast into the Lippmann–Schwinger integral equation

$$\psi(\mathbf{r},\mathbf{r}_s,\xi)=G(\mathbf{r},\mathbf{r}_s,\xi)+\int_{|r'|<a}V(\mathbf{r}';\omega)\psi(\mathbf{r}',\mathbf{r}_s,\xi)G(\mathbf{r},\mathbf{r}',\xi)d^2\mathbf{r}', \tag{5}$$

where $G(\mathbf{r},\mathbf{r}',\xi)=-\frac{1}{4}iH_0^{(1)}(\xi|\mathbf{r}-\mathbf{r}'|)$ is the two-dimensional Green's function. We write Eq. (1) as $\psi=\psi^{\rm inc}+\psi^{\rm sc}$, where $\psi^{\rm inc}$ and $\psi^{\rm sc}$ denote the incident and scattered wave functions, respectively.

If the incident field were the plane wave $e^{i\boldsymbol{\xi}\cdot\mathbf{r}}$ with $\boldsymbol{\xi}=\xi\hat{\xi}$ rather than the two-dimensional Green's function $G(\mathbf{r},\mathbf{r}_s,\xi)$, then Eq. (3) would be

$$(\nabla^2+\xi^2)\phi(\mathbf{r},\hat{\xi},\xi)=V(\mathbf{r};\omega)\phi(\mathbf{r},\boldsymbol{\xi},\xi), \tag{6}$$

with a corresponding Lippmann–Schwinger integral equation

$$\phi(\mathbf{r},\hat{\xi},\xi)=\exp(i\boldsymbol{\xi}\cdot\mathbf{r})+\int_{|r'|<a}V(\mathbf{r}';\omega)\phi(\mathbf{r}',\hat{\xi},\xi)G(\mathbf{r},\mathbf{r}',\xi)d^2\mathbf{r}', \tag{7}$$

which may be rewritten as $\phi=\phi^{\rm inc}+\phi^{\rm sc}$, where inc and sc denote the incident and scattered wave functions, respectively.

For Geometry IV, an argument modeled on that of Berezanskii[3] (see Appendix B) allows us to relate $\phi^{\rm sc}$ to $\psi^{\rm sc}$ by means of the formula

$$\phi^{\rm sc}(\mathbf{r},\hat{\xi},\xi)=\sum_{n=0}^{\infty}{}'\frac{4i^{n+1}}{\pi H_n^{(1)}(\xi a)}\int_0^{2\pi}\psi^{\rm sc}(\mathbf{r},\mathbf{r}_s,\xi)\cos(n\theta_{s\xi})d\theta_s \tag{8a}$$

for $|\mathbf{r}|=a=|\mathbf{r}_s|$. The prime on the summation sign in (8a) indicates that the $n=0$ term is halved. For Geometry IV′, for reflection from above, we use

$$\phi^{\rm sc}(\mathbf{r},\hat{\xi},\xi)=i\xi\,2^{-1}\cos\theta_\xi\int_{-\infty}^{\infty}\psi^{\rm sc}(\mathbf{r},\mathbf{r}_s,\xi)\times\exp(ix_s\xi\sin\theta_\xi)dx_s, \tag{8b}$$

where θ_ξ is the angle between $\mathbf{r}_s$ and $\boldsymbol{\xi}$; for the reflection from below and for transmission there are similar formulas.

The transformation (8) allows us to transform our point source data into plane-wave-incident data; however, we still have the problem that our data is known on the imaginary axis and not on the whole real axis. Following Stickler and Deift,[4] we deal with this problem as follows.

We know from the Lippmann–Schwinger equation [Eq. (8)] that in the absence of bound and half-bound states $\Phi=\exp(-i\boldsymbol{\xi}\cdot\mathbf{r})\phi^{\rm sc}$ is analytic in the upper half ξ plane. Moreover, a generalization of a proof in Cheney[5] shows that Φ is in the Hardy space H^{2+}, which is the space of functions analytic in the upper half-plane that are uniformly square integrable on each horizontal line. In order to obtain Φ on the real axis, we now apply a theorem of Van Winter's.[6] Her theorem allows us to determine Φ for real ξ by means of two successive transforms. We first form the Mellin transform

$$\tilde{\Phi}(\mathbf{r},\hat{\xi},t)=\exp\left(\frac{-\pi t}{2}+\frac{i\pi}{4}\right)\int_0^{\infty}\Phi(\mathbf{r},\hat{\xi}is)s^{it-1/2}ds; \tag{9a}$$

then we perform the inverse Mellin transform

$$\exp(-\boldsymbol{\xi}\cdot\mathbf{r})\phi^{\rm sc}(\mathbf{r},\hat{\xi},\xi)=(2\pi)^{-1}\int_{-\infty}^{\infty}\tilde{\Phi}(\mathbf{r},\hat{\xi},t)\xi^{-it-1/2}dt. \tag{9b}$$

Alternatively, if we wish to avoid analytic continuation altogether, we can use only the data we have on the real axis, namely $\phi^{\rm sc}(\mathbf{r},\hat{\xi},\xi)$ for $0\leqslant\xi\leqslant\omega/c_0$. Use of this data alone should lead to an answer with limited resolution. For the one-dimensional inverse scattering case, a theory of inversion using limited data has been worked out by Mel'nikov[7]; the higher-dimensional version has yet to be investigated.

IV. THE SCATTERING AMPLITUDE

The scattering amplitude $A(\xi,\hat{\xi}',\hat{\xi})$, corresponding to plane-wave incidence in the $\hat{\xi}$ direction and scattered cylindrical wave in direction $\hat{\xi}'$, must next be determined from $\phi^{\rm sc}(\mathbf{r},\hat{\xi},\xi)$ for $|\mathbf{r}|=a$ (for Assumption IV) or $\mathbf{r}=(x,a)$ and $\mathbf{r}=(x,-a)$ (for Assumption IV′). First we consider the case of Assumption IV. Because $\phi^{\rm sc}$ for $|\mathbf{r}|\geqslant a$ satisfies the Schrödinger equation with zero potential and also satisfies the radiation condition at $|\mathbf{r}|\to\infty$, we can write the solution for $|\mathbf{r}|\geqslant a$ as $\phi^{\rm sc}=\Sigma_{m=-\infty}^{\infty}a_mH_m^{(1)}(\xi,r)e^{-im\theta}$; at $r=a$ the coefficients $a_mH_m^{(1)}(\xi a)$ must be the Fourier coefficients of $\phi^{\rm sc}|_{r=a}$. This gives us

$$\phi^{\rm sc}(\mathbf{r},\hat{\xi},\xi)=\sum_{m=-\infty}^{\infty}\frac{H_m^{(1)}(\xi|\mathbf{r}|)e^{-im\theta}}{2\pi H_m^{(1)}(\xi a)}\times\int_0^{2\pi}\phi^{\rm sc}(\mathbf{r}',\hat{\xi},\xi)\Big|_{|r'|=a}e^{im\theta'}d\theta', \tag{10a}$$

from which, by using the asymptotic expansion of the Hankel functions for large argument, we conclude that

$$A(\xi,\hat{\xi}',\hat{\xi})=\sum_{m=-\infty}^{\infty}\frac{4e^{-im(\theta'+\pi/2)}}{2\pi H_m^{(1)}(\xi a)}\times\int_0^{2\pi}\phi^{\rm sc}(\mathbf{r},\hat{\xi},\xi)\Big|_{|\mathbf{r}|=a}e^{im\theta}d\theta, \tag{10b}$$

where $\hat{\xi}'=(\sin\theta',\cos\theta')$.

Now, under Assumption IV′, and for the case of reflection from above,

$$A(\xi,\hat{\xi}',\hat{\xi})=-2i\xi\sin\theta'\exp(i\xi a\sin\theta')\times\int_{-\infty}^{\infty}\phi^{\rm sc}((x,a),\hat{\xi},\xi)\exp(-i\xi x\cos\theta')dx. \tag{10c}$$

The results for reflection from below and transmission are similar. Weglein, Boyse, and Anderson,[3] Weglein and Silvia,[9] and Stolt and Jacobs[10] provided various equivalent methods for extracting $A(\xi,\hat{\alpha},\hat{\beta})$ from $\psi^{\rm sc}$ for situations such as IV and IV′.

V. RECONSTRUCTION OF THE POTENTIAL FROM THE SCATTERING AMPLITUDE

We now use $A(\xi,\hat{\alpha},\hat{\beta})$ as input into any of the existing two-dimensional inverse scattering theories, each of which requires different assumptions on the potential. Knowledge of the large ξ behavior of A gives us an inversion via the Born method: the Fourier transform $\hat{V}$ is computed via

$$\hat{V}(\tau;\omega)=\lim_{\substack{\xi\to\infty\\ \xi(\hat{\alpha}-\hat{\beta})=\tau}} A(\xi,\hat{\alpha},\hat{\beta}), \tag{11}$$

and then the Fourier transform is inverted.

A method with less dependence on large ξ data is the Newton–Cheney machinery.[11,5] We compute the kernel G of the Marchenko equation,

$$G_{\mathbf{r}}(s,\hat{\alpha},\hat{\beta})=2(2\pi)^{3/2}i\int_{-\infty}^{\infty}\exp[i\xi(s(\hat{\alpha}+\hat{\beta})\cdot\mathbf{r})]\times(\operatorname{sgn}\xi)A^*(\xi,\hat{\beta},\hat{\alpha})d\xi, \tag{12}$$

where the asterisk denotes complex conjugate. We insert $G_{\mathbf{r}}$ into the Marchenko equation,

$$\eta(s,\hat{\alpha},\mathbf{r})=\int_0^\infty\int_{S^1}G_{\mathbf{r}}(s+q,\hat{\alpha},\hat{\beta})\eta(q,-\hat{\beta},\mathbf{r})d\hat{\beta}\,dq+\int_{S^1}G_{\mathbf{r}}(s,\hat{\alpha},\hat{\beta})d\hat{\beta}, \tag{13}$$

where S^1 denotes the unit circle and $s>0$. For each $\mathbf{r}$, we must then solve (13) for $\eta(s,\hat{\alpha},\mathbf{r})$, from which the potential $V(\mathbf{r};\omega)$ can be computed via

$$V(\mathbf{r};\omega)=-2\hat{\alpha}\cdot\nabla_{\mathbf{r}}\eta(s,\hat{\alpha},\mathbf{r})|_{s=0^+}. \tag{14}$$

We remark that this solution must be miraculous, i.e., the right side of (14) must be independent of $\hat{\alpha}$. This method therefore applies only to the reconstruction problem.

Another method, which applies only to weak potentials, is an iterative scheme introduced by Moses[12] and Prosser.[13]

VI. RECONSTRUCTION OF THE VELOCITY AND DENSITY FROM THE POTENTIAL

If, as in Coen,[2] we perform the acoustic experiment with two arbitrary frequencies ω_1 and ω_2 the procedure described will reproduce the two potentials $V(\mathbf{r};\omega_l)$, $l=1,2$. Thus, the velocity is given by

$$c^2(\mathbf{r})c_0^{-2}=(\omega_1^2-\omega_2^2)\{\omega_1^2-\omega_2^2-c_0^2[V(\mathbf{r};\omega_1)-V(\mathbf{r};\omega_2)]\}^{-1}, \tag{15}$$

$$\{\nabla^2-\epsilon(\mathbf{r};\omega_1)\}g(\mathbf{r})=0,\quad g(\mathbf{r})=1\ \text{for}\ |\mathbf{r}|=a, \tag{16}$$

where $\epsilon(\mathbf{r};\omega_1)=V(\mathbf{r};\omega_1)-\omega_1^2c_0^{-2}[1-c_0^2c^{-2}(\mathbf{r})]$ and $\rho(\mathbf{r})\rho_0^{-1}=g^{-2}(\mathbf{r})$.

The solution of g of (16) is unique if it is positive, because if v were another solution then $(g-v)/g$ satisfies an equation to which the maximum principle applies. Moreover, g is guaranteed to be positive if we are doing a reconstructive problem.

ACKNOWLEDGMENTS

We are grateful to David Gilbarg, Roger Newton, and Joe Keller for some helpful comments. The work of Margaret Cheney was supported in part by the National Science Foundation, the Office of Naval Research, the Army Research Office, and the Air Force Office of Scientific Research. Arthur Weglein would like to thank Sohio Petroleum Company for permission to publish this work.

APPENDIX A

Here we briefly outline the proof that there are no bound or half-bound states when $c_0<c(x,z)$.

We multiply both sides of Eq. (7) by ϕ^*, change ϕ to Φ via $\phi=g\Phi$, and then use the expansion of $(\nabla\cdot g^2\Phi^*)\nabla\Phi$. Upon integration and use of the divergence theorem, this results in

$$\int_S g^2\Phi^*\frac{\partial\Phi}{\partial n}ds=-\xi^2\int_D g^2|\Phi|^2dx\,dz+\int_D g^2\frac{\omega^2}{c_0^2}\left(1-\frac{c_0^2}{c^2(x,z)}\right)|\Phi|^2dx\,dz+\int_D g^2|\nabla\Phi|^2dx\,dz, \tag{A1}$$

where S is the circle that bounds the area of the region D. If we let the radius of this circle go to infinity, then the left side of (A1) will vanish which implies that ξ^2 must be strictly positive provided that ω is real and $c_0\leqslant c(x,z)$. If ϕ is a bound or half-bound state, however, we must have $\xi^2\leqslant 0$. We thus conclude that the condition $c_0\leqslant c(x,z)$ guarantees the absence of bound and half-bound states.

Even without the condition $c_0\leqslant c(x,z)$, however, the presence or absence of bound states can be ascertained from the data, because bound states give rise to poles of $\phi(\mathbf{r}_t,\boldsymbol{\xi},\xi)$ that are located precisely on the imaginary axis.

APPENDIX B

In this Appendix we derive (8a). From (3) and (5) we deduce that ψ^{sc} satisfies the equation

$$[\nabla^2+\xi^2-V(\mathbf{r};\omega)]\psi^{sc}(\mathbf{r},\mathbf{r}_s,\omega)=V(\mathbf{r};\omega)G(\mathbf{r},\mathbf{r}_s,\xi). \tag{B1}$$

Similarly, from (6) and (7) we see that ϕ^{sc} satisfies

$$[\nabla^2+\xi^2-V(\mathbf{r};\omega)]\phi^{sc}(\mathbf{r},\hat{\xi},\xi)=V(\mathbf{r};\omega)\exp(i\boldsymbol{\xi}\cdot\mathbf{r}). \tag{B2}$$

We recall[14] that ∇^2-V is a self-adjoint operator with domain $W^{2,2}$ (the space of L^2 functions having two derivatives also in L^2) and range L^2. Moreover, if V has no bound states, then for ξ on the positive imaginary axis, ξ^2 is in the resolvent set. The inverse $[\nabla^2+\xi^2-V]^{-1}$ is therefore a bounded operator mapping L^2 onto $W^{2,2}$, and we can write (B1) and (B2) as

$$\psi^{sc}(\mathbf{r},\mathbf{r}_s,\xi)=[\nabla^2+\xi^2-V]^{-1}(-i/4)V(\mathbf{r};\omega)H_0^{(1)}(\xi|\mathbf{r}-\mathbf{r}_s|) \tag{B3}$$

and

$$\phi^{sc}(\mathbf{r},\hat{\xi},\xi)=[\nabla^2+\xi^2-V]^{-1}V(\mathbf{r};\omega)\exp(i\boldsymbol{\xi}\cdot\mathbf{r}), \tag{B4}$$

respectively. We note that the resolvent $[\nabla^2+\xi^2-V]^{-1}$ operates only in the $\mathbf{r}$ variable; the variables $\boldsymbol{\xi}$, $\mathbf{r}_s$, and ξ are merely parameters.

Next we use the well-known[15] Bessel function identities to express the exponential in (B4) in terms of Hankel functions. Roughly, we are adding up point sources to simulate a plane wave. In particular, we use the following identities:

$$\exp(i\xi|\mathbf{r}|\cos\theta_{r,\xi})$$
$$= J_0(\xi|\mathbf{r}|) + 2\sum_{n=1}^{\infty} i^n J_n(\xi|\mathbf{r}|)\cos n\theta_{r,\xi}, \qquad \text{(B5)}$$
$$H_0(\xi|\mathbf{r}-\mathbf{r}_s|)$$
$$= \sum_{k=-\infty}^{\infty} H_k(\xi|\mathbf{r}_s|)J_k(\xi|\mathbf{r}|)\cos(k\theta_{r,r_s}), \qquad \text{(B6)}$$
$$\int_0^{2\pi}\cos n\theta_{r,r_s}\cos k\theta_{r_s,\xi}\,d\theta_{r_s}$$
$$= \pi(\delta_{n,-k}+\delta_{n,k})\cos\theta_{r,\xi}, \qquad \text{(B7)}$$

where $\theta_{r,\xi} = \theta_r - \theta_\xi$ is the angle between the vectors $\mathbf{r}$ and $\boldsymbol{\xi}$. We multiply (B6) by $\cos n\theta_{r_s,\xi}$ and integrate with respect to θ_{r_s} to pick out a single term on the right side of (B6). We plug the resulting expression for $J_n(\xi|\mathbf{r}|)\cos n\theta_{r,\xi}$ into (B5) to obtain

$$\exp(i\boldsymbol{\xi}\cdot\mathbf{r}) = {\sum_{n=0}^{\infty}}' i^n \int_0^{2\pi} H_0(\xi|\mathbf{r}-\mathbf{r}_s|)$$
$$\times\cos n\theta_{r_s,\xi}\,d\theta_{r_s}[\pi H_n(\xi|\mathbf{r}_s|)]^{-1}, \qquad \text{(B8)}$$

where the prime on the sum indicates that the $n=0$ term is to be multiplied by $\frac{1}{2}$.

We now use (B8) in (B4); since $[\nabla^2+\xi^2-V]^{-1}$ is continuous, it can be brought inside the summation and integration signs. We then use (B3) to obtain (8a).

APPENDIX C

In this Appendix we derive (8b). We first compute the Fourier transform of the Hankel function $H_0^{(1)}$. We use the representation

$$-\frac{i}{4}H_0^{(1)}(\xi r) = (2\pi)^{-2}\int \exp(i\mathbf{k}\cdot\mathbf{r})(\xi^2-k^2)^{-1}d\,\mathbf{k}. \qquad \text{(C1)}$$

The Fourier transform can then be computed easily from (C1),

$$-\frac{i}{4}\int_{-\infty}^{\infty}\exp(-i\eta x)H_0^{(1)}(\xi r)dx$$
$$= (2\pi)^{-1}\int \exp(ik_2 z)(\xi^2-\eta^2+k_2^2)^{-1}dk_2$$
$$= (i/2)\exp[i(\xi^2-\eta^2)^{1/2}|z|](\xi^2-\eta^2)^{-1/2}. \qquad \text{(C2)}$$

For $r = |\mathbf{r}-\mathbf{r}_s|$, (C2) becomes

$$-\frac{i}{4}\int_{-\infty}^{\infty}\exp(-i\eta x_s)H_0^{(1)}(\xi|\mathbf{r}-\mathbf{r}_s|)dx_s$$
$$= (i/2)\exp[i(\xi^2-\eta^2)^{1/2}|z-z_s| + i\eta x](\xi^2-\eta^2)^{-1/2}. \qquad \text{(C3)}$$

In (C3) we then make the substitution $\eta = \xi\sin\theta$, $0<\theta<\pi$, and write $\hat{\theta} = (\cos\theta, \sin\theta)$,

$$-\frac{i}{4}\int_{-\infty}^{\infty}\exp(-i\xi\sin\theta x_s)H_0^{(1)}(\xi|\mathbf{r}-\mathbf{r}_s|)dx_s$$
$$= (i/2)\exp[i(x,z-z_s)\cdot\hat{\theta}](\xi\cos\theta)^{-1}. \qquad \text{(C4)}$$

We use (C4) in (B4), obtaining (8b).

APPENDIX D

In this Appendix we derive (10c). As in the case of Geometry IV, we know that ϕ^{sc} for $\mathbf{r} = (x,z)$ with $|z|>a$ satisfies the Schrödinger equation with zero potential. Moreover, the values of ϕ^{sc} for $r = (x, \pm a)$ are known, and at infinity ϕ^{sc} satisfies the radiation condition.

Because ϕ^{sc} is a generalized eigenfunction of Δ corresponding to $-\xi^2$, it can be written

$$\phi^{sc}(\mathbf{r},\hat{\xi},\xi) = \int_{-\pi/2}^{\pi/2} b(\boldsymbol{\xi},\hat{\theta})\exp(i\xi\hat{\theta}\cdot r)d\hat{\theta}, \qquad \text{(D1)}$$

where b is to be determined as follows. [The range of integration in (D1) is restricted in order to satisfy the radiation condition.] On the line $z=a$ we write $\hat{\theta}\cdot\mathbf{r} = x\cos\theta + a\sin\theta$, where θ is measured from the x axis. We then multiply (D1) by $\xi\exp(-ix\xi\cos\theta)$ and integrate over x:

$$\int_{-\infty}^{\infty}\phi^{sc}((x,a),\hat{\xi},\xi)\exp(-ix\xi\cos\phi)\xi\,dx$$
$$= 2\pi\int_{-\pi/2}^{\pi/2} b(\boldsymbol{\xi},\hat{\theta})\exp(i\xi a\sin\theta)\delta(\cos\theta-\cos\phi)d\theta$$
$$= -2\pi b(\boldsymbol{\xi},\hat{\theta})\exp(i\xi a\sin\theta)/\sin\theta. \qquad \text{(D2)}$$

This gives us b in terms of the known values of ϕ^{sc}:

$$b(\boldsymbol{\xi},\hat{\theta}) = -\xi(2\pi)^{-1}\sin\theta\int_{-\infty}^{\infty}\phi^{sc}((x,a),\hat{\xi},\xi)$$
$$\times\exp(-i\xi(x,-a)\cdot\hat{\phi})dx. \qquad \text{(D3)}$$

Use of (D3) in (D1) gives us

$$\phi^{sc}(\mathbf{r},\hat{\xi},\xi) = -\xi(2\pi)^{-1}\int_0^{2\pi}\sin\theta\int_{-\infty}^{\infty}\phi^{sc}((x,a),\hat{\xi},\xi)$$
$$\times\exp[i\xi\hat{\theta}\cdot(\mathbf{r}-(x,-a))]dx\,d\theta. \qquad \text{(D4)}$$

Our next task is to compute the $r^{-1/2}$ term of the large-r asymptotic expansion for ϕ^{sc}, because this term will give us the scattering amplitude.

An application of the stationary phase approximation to (D1) results in

$$\phi^{sc}(\mathbf{r},\hat{\xi},\xi) = \pi^{1/2}(2\xi r)^{-1/2}b(\boldsymbol{\xi},\hat{r})$$
$$\times\exp(i\xi r + i\pi/4) + O((\xi r)^{-1}). \qquad \text{(D5)}$$

The $(\xi r)^{-1/2}$ term of ϕ^{sc} can also be computed another way, namely by means of the Lippman–Schwinger equation. The $(\xi r)^{-1/2}$ term is

$$(8\pi)^{-1/2}\exp(-3\pi i/4)(\xi r)^{-1/2}A(\xi,\hat{r},\hat{\xi})\exp(i\xi r). \qquad \text{(D6)}$$

Comparison of (D5) and (D6) tells us that

$$A(\xi,\hat{r},\hat{\xi}) = -2b(\boldsymbol{\xi},\hat{r})$$
$$= -\xi\pi^{-1}\sin\theta_r\int_{-\infty}^{\infty}\phi^{sc}((x,a),\hat{\xi},\xi)$$
$$\times\exp(-i\xi(x,-a)\cdot\hat{r})dx,$$

where θ_r is the angle between $\mathbf{r}$ and the x axis.

[1] J. A. Ware and K. Aki, J. Acoust. Soc. Am. **45**, 911 (1969).
[2] S. Coen, J. Math. Phys. **22**, 2338 (1981); Geophysics **46**, 1244 (1981); J. Acoust. Soc. Am. **70**, 1473 (1981); Geophysics **47**, 898 (1982); Bull. Seismol. Soc. Am.**72**, 809 (1982); J. Geophys. Res. **88**, 4299 (1983).
[3] Ju. M. Berezanskii, AMS Transl. Ser. 2 **35**, 167 (1964).
[4] D. C. Stickler and P. A. Deift, J. Acoust. Soc. Am. **70**, 1723 (1981).
[5] M. Cheney, J. Math. Phys. **25**, 94 (1984); **25**, 94 (1984).
[6] C. Van Winter, Trans. Am. Math. Soc. **162**, 103 (1971).
[7] V. K. Mel'nikov, Am. Math. Soc. Transl. **25**, 271 (1963).
[8] A. B. Weglein, W. E.Boyse, and J. E. Anderson, Geophysics **48**, 1116 (1981).

[9]A. B. Weglein and M. T. Silvia, "Scattering theory approach to the identification of the Helmholtz equation: a near-field solution," J. Acoust. Soc. Am. **69**, 483–488 (1981).
[10]R. H. Stolt and B. Jacobs, "Inversion of seismic data in a laterally heterogeneous media," paper presented at the 1980 S.E.G. meeting, Houston, Texas.
[11]R. G. Newton, J. Math. Phys. **21**, 1968 (1980); **22**, 2191 (1981); **23**, 594 (1982).
[12]H. E. Moses, Phys. Rev. **102**, 559 (1956).
[13]R. Prosser, J. Math. Phys. **21**, 2648 (1980); J. Math. Phys. **23**, 2127 (1982).
[14]M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, New York, 1975).
[15]*Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (National Bureau of Standards, Washington, D.C., 1964).

The quantum scattering amplitude in $\mathbb{R}^n$ for potentials decreasing faster than $|x|^{-(n+1)/2}$

J. M. Combes[a] and J. M. Ghez[a]
Centre de Physique Theorique, Section 2, C.N.R.S. Luminy, Case 907, 13288 Marseille, Cedex 9, France

(Received 9 June 1983; accepted for publication 22 July 1983)

Within the framework of the two-body problem in quantum mechanics, we prove an "analytic limiting absorption principle" for the Schrödinger operator $H = H_0 + V$, on $L^2(\mathbb{R}^n)$, where $H_0 = -\Delta$ and V is a real short-range potential, that is, decreasing like $|x|^{-1-\epsilon}$, $\epsilon > 0$, as $|x| \to \infty$. Applying it in the case where V decreases like $|x|^{-(n+1)/2}(\log|x|)^{-1-\epsilon}$, $\epsilon > 0$, as $|x| \to \infty$, we obtain two classes of results for the non-Born part of the scattering amplitude: finiteness, continuity and high energy properties on the positive real axis, including correction to the Born approximation of the cross section; analyticity and high energy properties in the upper half-plane for the forward amplitude.

PACS numbers: 03.65.Nk

I. INTRODUCTION

Some great progress in stationary scattering theory has been achieved by the beautiful works of Agmon[1] and Agmon and Hörmander.[2] For an optimal class of two-body short-range potentials, they prove the existence of boundary values of Green's operators and some very general results on the asymptotic behavior of the stationary scattering wave functions.

It turns out that such methods can also provide enough information on the non-Born part of the forward scattering amplitude to prove its finiteness and meromorphy in the case of potentials V decaying faster than $|x|^{-(n+1)/2}$ in $\mathbb{R}^n$. For $n = 3$, analyticity properties have been investigated in Refs. 3–6 under the condition $V(x) = O(|x|^{-3-\epsilon})$, $\epsilon > 0$, as $|x| \to \infty$. There is a gap between this type of condition and those imposed in the very recent works of Amrein and Pearson,[7] Martin,[8] Enss and Simon,[9] and Saito[10] on the finiteness of the total cross section and of Jensen[11] on its Born approximation at high energies. All these authors assume essentially that the potential decreases faster than $|x|^{-2}$ as $|x| \to \infty$.

This paper fills this gap. Indeed, for a (non-necessarily spherically symmetric) real potential V such that $V(x) = O(|x|^{-(n+1)/2}(\log|x|)^{-1-\epsilon})$, $\epsilon > 0$, as $|x| \to \infty$, we prove the finiteness of the total cross section $\sigma(k; \omega_-)$ and of its Born approximation $\sigma_B(k; \omega_-)$ for any fixed energy k^2 and incident direction ω_-, their continuity in k, their high energy behavior as k^{-2}, and the high energy behavior as k^{-3} of the correction $\sigma(k; \omega_-) - \sigma_B(k; \omega_-)$. Up to now, finiteness, continuity and Born approximation of the total cross section under conditions similar to ours have been obtained only for spherically symmetric potentials,[8] the average over all incident directions[7,8,11] or the averages over small energy intervals[9,10]. Moreover, our hypotheses are almost optimal, concerning the decay at infinity [see the remark (i) following Lemma 3.2]. Our results on the cross section are derived from the corresponding ones for the non-Born part $t(k; \omega_-, \omega_+)$ of the amplitude by means of the optical theorem (for related results, see the works of Agmon[12] and Saito[10]). We establish the existence of a meromorphic continuation of the forward scattering amplitude $t(k; \omega_-)$ to the upper half-plane having a $|k|^{-1}$ high energy behavior. In order to get dispersion relations, we would still need some information about the low energy behavior; however, the discussion of this problem seems to require stronger assumptions than ours on the potential (see e.g., Ref. 13) and will not be developed here.

To avoid unessential technicalities, we derive first our results under the assumption that the potential is pointwise bounded. We will show that local singularities of type $1/|x - x_0|^\alpha$, $\alpha < 2$, do not modify essentially our conclusions. They influence mostly the high energy behavior of $\sigma(k; \omega_-) - \sigma_B(k; \omega_-)$, which is essentially of order $k^{-\inf(3,4-\alpha)}$; this improves partially the results of Jensen.[11]

Our main tools are Agmon's method of elliptic *a priori* estimates and the theory of complex canonical transformations.[14] Indeed, the key role is played by a limiting absorption principle for the Hamiltonians transformed under the action of complex boost transformations.

In Sec. II, we will collect the definitions on the various spaces and classes of potentials to be used later; we also recall the properties of the complex boosts. In Sec. III, we study the Born approximation to the scattering amplitude and cross section. In Sec. IV, we prove an "analytic limiting absorption principle" and use it to analyze the non-Born part of the amplitude, first on the positive real axis in Sec. V and then in the upper half-plane in Sec. VI. We study in Sec. VII the high energy behaviors in case V has local singularities. Finally, in Sec. VIII, we present some remarks concerning dilation analytic potentials.

II. MATHEMATICAL PRELIMINARIES

A. Weighted Sobolev spaces and Besov spaces

Let $H = -\Delta + V$ denote the quantum Hamiltonian with potential V. The definition of the scattering amplitude involves the resolvent $R(z) = (H - z)^{-1}$ and its boundary value as z reaches the continuous spectrum of H, which is $\mathbb{R}^+$ in the situations investigated here. It is well known that

[a] Phymat, Département de Mathématiques, Université de Toulon et du Var and Centre de Physique Théorique, CNRS, Marseille, France.

 0022-2488/84/061862-13$02.50

this limit does not exist as a bounded operator on $L^2(\mathbb{R}^n)$. The limiting absorption principle is the property that, for a suitable auxiliary Banach space $\mathfrak{X}$, the limits

$$R^{\pm}(k^2) = \lim_{\epsilon \downarrow 0^{\pm}} R(k^2 + i\epsilon)$$

exist in a suitable topology of $\mathscr{L}(\mathfrak{X}, \mathfrak{X}^*)$, the space of bounded linear operators from $\mathfrak{X}$ to $\mathfrak{X}^*$. In case $V = 0$, $H = H_0 = -\Delta$, Agmon[1] showed that such limits exist in the norm operator topology of $\mathscr{L}(L^2_s, L^2_{-s})$, $s > \frac{1}{2}$, where

$$L^2_s(\mathbb{R}^n) = \{u \in L^2_{\rm loc}(\mathbb{R}^n), (1 + |x|^2)^{s/2} u \in L^2(\mathbb{R}^n)\}$$

is a Banach space for the norm

$$\|u\|_{0,s} = \|(1 + |x|^2)^{s/2} u\|_{L^2(\mathbb{R}^n)}.$$

This can be improved[1] to show existence in $\mathscr{L}(L^2_s, \mathscr{H}^2_{-s})$, where

$$\mathscr{H}^m_s(\mathbb{R}^n) = \{u \in L^2_{\rm loc}(R^n), D^\alpha u \in L^2_s(\mathbb{R}^n), |\alpha| \leqslant m\}, \quad m \in \mathbb{Z}, s \in \mathbb{R},$$

is a Banach space for the norm

$$\|u\|_{m,s} = \Big(\sum_{|\alpha| \leqslant m} \|D^\alpha u\|^2_{0,s} \Big)^{1/2}.$$

The restriction $s > \frac{1}{2}$ is easily understood from the trace properties of functions in $L^2_s(\mathbb{R}^n)$.[15] If one looks for an optimal auxiliary space for which the limiting principle would hold, one can also use, following Agmon and Hörmander,[2] the Besov space

$$\mathfrak{B}(\mathbb{R}^n) = \Big\{ u \in L^2_{\rm loc}(\mathbb{R}^n), \ \|u\|_{\mathfrak{B}(\mathbb{R}^n)} = \sum_{j=1}^{\infty} R_j^{1/2} \|u\|_{L^2(\Omega_j)} < \infty \Big\},$$

where $R_0 = 0$, $R_j = 2^{j-1}$ when $j > 0$ and $\Omega_j = \{x \in \mathbb{R}^n, R_{j-1} < |x| < R_j\}$. Notice that $L^2_s(\mathbb{R}^n) \subset \mathfrak{B}(\mathbb{R}^n) \subset L^2_{\frac{1}{2}}(\mathbb{R}^n)$ for all $s > \frac{1}{2}$; in terms of decay at infinity, the elements of $\mathfrak{B}(\mathbb{R}^n)$ essentially decrease faster than $C|x|^{-(n+1)/2}(\log|x|)^{-1-\epsilon}$ for some $C, \epsilon > 0$. The dual space $\mathfrak{B}^*(\mathbb{R}^n)$ can be shown[2] to be identified by the rather suggestive property (from a scattering point of view)

$$\mathfrak{B}^*(\mathbb{R}^n) = \Big\{ u \in L^2_{\rm loc}(\mathbb{R}^n), \ \sup_{R > 1} \frac{1}{R} \int_{|x| < R} |u(x)|^2 dx < \infty \Big\}.$$

Further properties of $\mathfrak{B}^*(\mathbb{R}^n)$ will be described in Appendix A.

B. Short-range potentials

By perturbative arguments, the limiting absorption principle can be proved when the potential V belongs to the so-called "short-range class"; a real function $V \in L^2_{\rm loc}(\mathbb{R}^n)$ is called a short-range potential if, for some $\epsilon > 0$, the mapping $u(x) \mapsto (1 + |x|^2)^{1/2 + \epsilon} V(x) u(x)$ defines a compact operator from $\mathscr{H}^2(\mathbb{R}^n)$ to $L^2(\mathbb{R}^n)$. In view of our discussion of strong singularities in Sec. VII, it is worth noticing that one could weaken the assumption and require compactness from $\mathscr{H}^1(\mathbb{R}^n)$ to $\mathscr{H}^{-1}(\mathbb{R}^n)$,[15] allowing in this way to form bounded perturbations.

For what follows, it is important to make the following remarks:

(i) The short-range class includes obviously bounded potentials such that

$$|V(x)| \leqslant C(1 + |x|)^{-(n+1)/2} \log(2 + |x|)^{-1-\epsilon}, \quad \epsilon > 0.$$

(ii) A short-range potential V is compact from $\mathscr{H}^2(\mathbb{R}^n)$ to $L^2(\mathbb{R}^n)$ and also from $\mathscr{H}^2_{-s}(\mathbb{R}^n)$ to $L^2_s(\mathbb{R}^n)$ for $s < s_0$, where $s_0 > \frac{1}{2}$ depends on the rate of decay of V. Then, in particular, V is compact from $\mathscr{H}^2_{-s}(\mathbb{R}^n)$ to $\mathfrak{B}(\mathbb{R}^n)$ for all $s < s_0$.

(iii) For a short range potential V, the Hamiltonian $H = H_0 + V$ is a self-adjoint operator on $L^2(\mathbb{R}^n)$, with domain $\mathscr{H}^2(\mathbb{R}^n)$. Its essential spectrum is $\mathbb{R}^+$ and nonzero eigenvalues are of finite multiplicity; they form a discrete set with 0 and $+\infty$ as only possible limit points.[1] We will discuss some situations where it can be asserted that $+\infty$ is not a limit point (or equivalently the positive point spectrum is finite). Actually, one can exclude positive point spectrum under rather general conditions (see, e.g., Refs. 16 and 17).

(iv) If V is short range, H has no continuous singular spectrum, and wave operators exist and are complete.

C. The boost group[14]

The boost group is the linear representation in $L^2(\mathbb{R}^n)$ of the group of translations in momentum space. Its action on $u \in L^2(\mathbb{R}^n)$ is given by

$$B(\tau) u(x) = e^{i\tau \cdot x} u(x), \quad \tau \in \mathbb{R}^n.$$

Define $H_0(\tau) = B^{-1}(\tau) H_0 B(\tau) = (p + \tau)^2$ and (if V is local) $H(\tau) = B^{-1}(\tau) H B(\tau) = H_0(\tau) + V$; then, if V is H_0-compact, the family $H(\tau)$ can be extended in $\mathbb{C}^n$ to an analytic family of

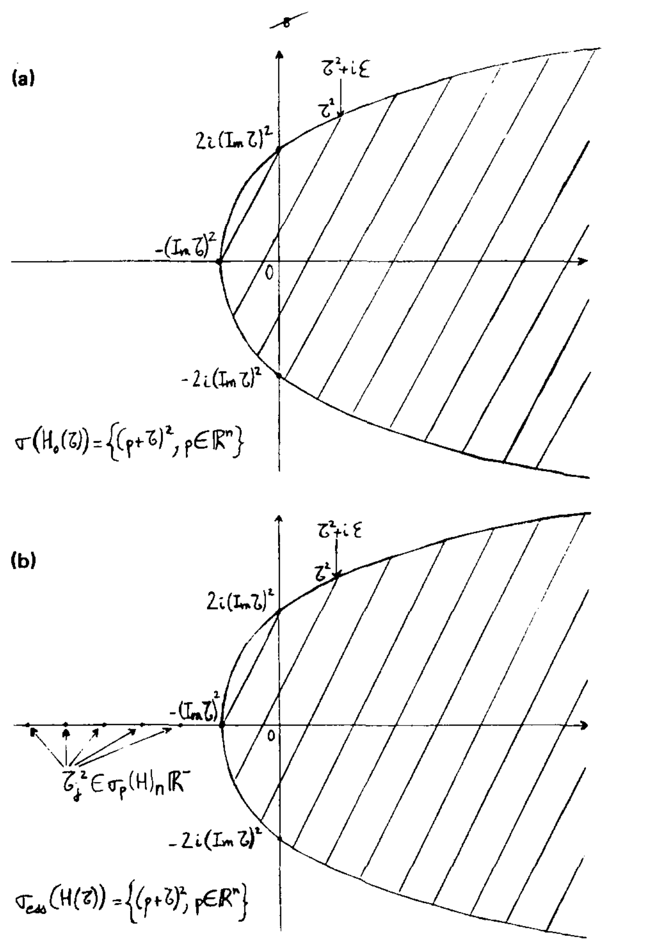


FIG. 1.

type A.[18] The spectrum of $H(\tau)$ depends only on Im τ; it is represented on Fig. 1. We refer to Ref. 14 for further details.

III. THE SCATTERING AMPLITUDE AND THE CROSS SECTION FOR POTENTIALS IN THE BESOV CLASS

The scattering amplitude is defined as

$$F(k;\omega_-,\omega_+) = C_n e^{-i(n-3)\pi/4} k^{(n-3)/2} T(k;\omega_-,\omega_+),$$

where $C_n = -\pi^{-(n-1)/2} 2^{-(n+1)/2} (2\pi)^n$ and

$$T(k;\omega_-,\omega_+) = \frac{1}{(2\pi)^n}\Big[\langle Ve^{ik\omega_+\cdot x}, e^{ik\omega_-\cdot x}\rangle - \lim_{\epsilon\downarrow 0^+}\langle Ve^{ik\omega_+\cdot x}, R(k^2+i\epsilon)Ve^{ik\omega_-\cdot x}\rangle\Big].$$

Here $\omega_\pm \in S^{n-1}$, the unit sphere in $\mathbb{R}^n$, and $k\in\mathbb{R}^+\setminus\{0\}$.

Some general results on continuity properties of T can be found in Refs. 10 and 12. Here we will concentrate mostly on its L^2 properties. Consider first the Born term:

$$t_B(k;\omega_-,\omega_+) = [1/(2\pi)^n]\langle Ve^{ik\omega_+\cdot x}, e^{ik\omega_-\cdot x}\rangle.$$

For potentials in $\mathfrak{B}(\mathbb{R}^n)$, one has

$$t_B(k;\omega_-,\omega_+) = (2\pi)^{-n/2}\hat{V}(k(\omega_+-\omega_-)),$$

where $\hat{V}$ denotes the Fourier transform of V; we will see later that this function exists for $k\in\mathbb{R}^+\setminus\{0\}$ as a square-integrable function of ω_+ for each fixed ω_-; it is worth mentioning that this is an aspect of a trace property for Fourier transforms of elements of Besov space (see Ref. 2, Theorem 2.3). From this point of view, it appears that such a property would not hold for a strictly larger class of potentials than $\mathfrak{B}(\mathbb{R}^n)$ [see remark (i) below]. In general, t_B is not defined pointwise; in particular, the forward Born amplitude

$$t_B(k;\omega_-,\omega_-) = (2\pi)^{-n/2}\hat{V}(0)$$

is not finite in general unless $V\in L^1(\mathbb{R}^n)$.

Concerning the non-Born term

$$t(k;\omega_-,\omega_+) = \frac{1}{(2\pi)^n}\lim_{\epsilon\downarrow 0^+}\langle Ve^{ik\omega_+\cdot x}, R(k^2+i\epsilon)e^{ik\omega_-\cdot x}\rangle, \tag{3.1}$$

we will see that it is finite for all k in $\mathbb{R}^+\setminus\{0\}$ and all $\omega_-,\omega_+\in S^{n-1}$ as a consequence of a limiting absorption principle for H. A generalization of this principle will further lead to an analytic continuation of $t(k;\omega_-,\omega_-)$ for k complex.

The cross section at energy k^2 and incident momentum ω_- is defined as

$$\sigma(k;\omega_-) = \int_{S^{n-1}} |F(k;\omega_-,\omega_+)|^2\, d\omega_+. \tag{3.2}$$

By the properties stated above, it is finite for all $k\in\mathbb{R}^+\setminus\{0\}$ when $V\in\mathfrak{B}(\mathbb{R}^n)$. To motivate the choice of this class, let us consider the Born approximation to $\sigma(k;\omega_-)$:

$$\sigma_B(k;\omega_-) = \int_{S^{n-1}} |F_B(k;\omega_-,\omega_+)|^2\, d\omega_+,$$

where

$$F_B(k;\omega_-,\omega_+) = C_n e^{-i(n-3)\pi/4}k^{(n-3)/2} t_B(k;\omega_-,\omega_+)$$

so that

$$\sigma_B(k;\omega_-) = \pi^{-(n-1)}2^{-(n+1)}(2\pi)^n k^{n-3} \times\int_{S^{n-1}} |\hat{V}(k(\omega_+-\omega_-))|^2\, d\omega_+.$$

One has:

Theorem 3.1: Let $V\in\mathfrak{B}(\mathbb{R}^n)$. Then

(i) $\sigma_B(k;\omega_-)$ is finite for all $k\in\mathbb{R}^+\setminus\{0\}$, $\omega_-\in S^{n-1}$.
(ii) Let $\delta>0$. For $k\geqslant\delta$, one has

$$\sigma_B(k;\omega_-)\leqslant Ck^{-2}\|V\|^2_{\mathfrak{B}(\mathbb{R}^n)},$$

where C is a constant depending only on δ. This immediately follows from

Lemma 3.2: Let $V\in\mathfrak{B}(\mathbb{R}^n)$ and $t_B(k;e,\omega) = (2\pi)^{-n/2}\hat{V}(k(\omega-e))$ with $k\in\mathbb{R}^+\setminus\{0\}$, $e\in\mathbb{R}^n$, $\omega\in S^{n-1}$. Then

(i) $t_B(k;e,\cdot)\in L^2(S^{n-1})$.

(ii) Let $\delta>0$. For $k\geqslant\delta$, one has

$$\|t_B(k;e,.)\|_{L^2(S^{n-1})}\leqslant C_0 k^{-(n-1)/2}\|V\|_{\mathfrak{B}(\mathbb{R}^n)},$$

where C_0 is a constant depending only on δ.

Proof: Consider $H_0(ke)$ as defined in (2.3); by Theorem A.1, one has, for $k\geqslant\delta$,

$$|\langle V,(H_0(ke)-k^2-i\epsilon)^{-1}V\rangle|\leqslant C|k|^{-1}\|V\|^2_{\mathfrak{B}(\mathbb{R}^n)}.$$

On the other hand,

$$\begin{aligned}
\pi^{-1}&\lim_{\epsilon\downarrow 0^+}\operatorname{Im}\langle Ve^{ike\cdot x},(H_0-k^2-i\epsilon)^{-1}Ve^{ike\cdot x}\rangle\\
&=\pi^{-1}\lim_{\epsilon\downarrow 0^+}\operatorname{Im}\langle V,(H_0(ke)-k^2-i\epsilon)^{-1}V\rangle\\
&=\pi^{-1}\lim_{\epsilon\downarrow 0^+}\int_{\mathbb{R}^n}\frac{\epsilon}{[(p+ke)^2-k^2]^2+\epsilon^2}|\hat{V}(p)|^2\,dp\\
&=\frac{1}{2}k^{n-2}\int_{S^{n-1}}|\hat{V}(k(\omega-e))|^2\,d\omega\\
&=\frac{(2\pi)^n}{2}k^{n-2}\int_{S^{n-1}}|t_B(k;e,\omega)|^2\,d\omega,
\end{aligned}$$

which gives the required estimate.

Remarks: (i) Our claim that Besov class is optimal for finiteness of $\sigma_B(k;\omega_-)$ does not seem to be true if one imposes the condition of spherical symmetry on the potential. In fact, it is shown by Martin,[8] for $n=3$, that $\sigma_B(k;\omega-)$ is proportional in this case to

$$I = 4\pi k^{-2}\int\frac{\sin^2[k|x-x'|]}{|x-x'|^2}|V(x)|\,|V(x')|\,d^3x\,d^3x',$$

which is finite if V belongs to the Rollnick class.[6] This allows a decay of V at infinity like $|x|^{-2}(\log|x|)^{-1/2-\epsilon}$, $\epsilon>0$. If V is not spherically symmetric, the cross section averaged over all incident directions $\omega_-\in S^2$ is also finite if V is a Rollnick potential with $\|V\|_{\mathscr{R}}<1$, i.e., in the case of small coupling constant. Furthermore, Martin is able to get explicit bounds on the cross section in terms of the Rollnick norm. In Ref. 9, Enss and Simon noticed that, in the spherically symmetric case, the correct borderline for finiteness of the cross section is $\frac{1}{2}$ for the logarithmic part of the decay.

(ii) The proof of Lemma 3.2 yields the "optical theorem" for $\sigma_B(k;\omega_-)$, namely

$$\sigma_B(k;\omega_-)=(2C_n^2/\pi k)\operatorname{Im} t_0(k;\omega_-), \tag{3.3}$$

where $t_0(k;\omega_-)$ is obtained from (3.1) by replacing R by R_0:

$$t_0(k;\omega_-)=t_0(k;\omega_-,\omega_-)$$
$$=\frac{1}{(2\pi)^n}\lim_{\epsilon\downarrow 0^+}\langle Ve^{ik\omega_-\cdot x},R_0(k^2+i\epsilon)Ve^{ik\omega_-\cdot x}\rangle.$$

In order to investigate the properties of the other component of T, let us rewrite (3.1) in a slightly different form which will be very convenient for analytic continuation. Assuming that $\omega_-\in S^{n-1}$ is fixed, we introduce the definitions:

$$B(k)=B(k\omega_-),\quad H(k)=H(k\omega_-),$$

and

$$R(k,z)=(H(k)-z)^{-1},\quad k\in\mathbb{R},\ \operatorname{Im} z\neq 0$$

and notice that

$$\langle Ve^{ik\omega_-\cdot x},R(k^2+i\epsilon)Ve^{ik\omega_-\cdot x}\rangle=\langle V,R(k,k^2+i\epsilon)V\rangle$$

so that

$$t(k;\omega_-)=t(k;\omega_-,\omega_-)=\frac{1}{(2\pi)^n}\lim_{\epsilon\downarrow 0^+}\langle V,R(k,k^2+i\epsilon)V\rangle. \tag{3.4}$$

We will prove that the limit in the rhs of (3.4) exists for $k\in\mathbb{C}^{++}\cup\mathbb{C}^{+-}$, where

$$\mathbb{C}^{++}(\text{resp.}\,\mathbb{C}^{+-})$$
$$=\{z\in\mathbb{C},\ \operatorname{Im} z\geqslant 0,\ +\operatorname{Re} z>0(\text{resp.}-\operatorname{Re} z>0)\}$$

by using an analytic limiting absorption principle for H. Furthermore, this limit provides with an analytic continuation of $t(k;\omega_-)$ in the upper half-plane, with the exception of poles corresponding to $k^2\in[\sigma_p(H)\cap\mathbb{R}^-]\setminus\{0\}$ [where $\sigma_p(H)$ denotes the point spectrum of H]. Other properties of t, in particular its asymptotic behavior for large k, will also follow from the mathematical background of the next chapter.

IV. THE ANALYTIC LIMITING ABSORPTION PRINCIPLE FOR *H*

The main problem to be solved is, according to (3.4), to give a meaning to the limit of $R(k,k^2+i\epsilon)$ as $\epsilon\downarrow 0^+$. Recall that, for $k\in\mathbb{R}^+$, $k^2+i\epsilon$ falls on the essential spectrum of H, hence of $H(k)$, as $\epsilon\downarrow 0^+$, so that this limit does not exist in any of the usual topologies of $\mathscr{L}(L^2,L^2)$. For $k\in\mathbb{C}^{++}$, the same difficulty occurs as Fig. 1 shows. However, by (3.4) and the fact that V is in $\mathfrak{B}(\mathbb{R}^n)$, it is enough that the limits of these resolvents exist in the weak topology of $\mathscr{L}(\mathfrak{B},\mathfrak{B}^*)$.

For the convenience of the reader and further purposes, in particular the generalization to complex k, we will give here a self-contained version of Agmon's *a priori* estimate method. The perturbation theoretic arguments used here are also direct adaptations of those of Ref. 1 to which we will often refer.

Define $\sigma_0(k)=\sigma(H_0(k))$ and $\rho_0(k)=\mathbb{C}\setminus\sigma_0(k)$. The main result of this chapter is the following:

Theorem 4.1: Let V a short-range potential and $H=-\Delta+V$. Then

(i) The limits

$$R(k,k^2)=\lim_{\epsilon\downarrow 0^+}R(k,k^2+i\epsilon)$$

exist in the weak operator topology of $\mathscr{L}(\mathfrak{B},\mathfrak{B}^*)$ for all $k\in\mathbb{C}^+\setminus\Sigma$, where $\mathbb{C}^+=\{z\in\mathbb{C},\ \operatorname{Im} z\geqslant 0\}$ and $\Sigma=\{k\in\mathbb{C}^+,\ k^2\in\sigma_p(H)\cup\{0\}\}$.

(ii) $R(k,k^2)$ is meromorphic in $\mathbb{C}^+\setminus\mathbb{R}$ with simple poles on $(\Sigma\cap i\mathbb{R}^+)\setminus\{0\}$ and has weakly continuous boundary values on $\mathbb{R}\setminus\Sigma$.

(iii) If $k_0\in(\mathbb{R}^+\cap\Sigma)\setminus\{0\}$, $P\{k_0^2\}$ is the corresponding spectral projection operator and

$$P(k,k_0^2)=B^{-1}(k)P(k_0^2)B(k),\quad k\in\mathbb{R},$$

then there exists a real neighborhood $\nu(k_0)$ of k_0 such that

$$\hat{R}(k,k^2)=(1-P(k,k_0^2))R(k,k^2)$$

is continuous in the weak operator topology of $\mathscr{L}(\mathfrak{B},\mathfrak{B}^*)$.

(iv) Any pole k_0 of $R(k,k^2)$ on $i\mathbb{R}^+$ is simple with residue $-P(k_0,k_0^2)$, where, in the weak operator topology of $\mathscr{L}(\mathfrak{B},\mathfrak{B}^*)$,

$$P(k_0,k_0^2)=\lim_{\substack{k\to k_0\\ |\operatorname{Im} k|<|k_0|}}P(k,k_0^2)$$

and $P(k,k_0^2)$ is the analytic continuation to $\{k\in\mathbb{C}\,|\operatorname{Im} k|<|k_0|\}$ of $B^{-1}(k)P\{k_0^2\}B(k)$.

(v) $R(k,k^2)=R(-\bar{k},\bar{k}^2)^*\quad\forall k\in\mathbb{C}^{++}$.

(vi) If in addition V is bounded, then $\sigma_p(H)$ is a bounded set and

$$\|R(k,k^2)\|_{\mathscr{L}(\mathfrak{B},\mathfrak{B}^*)}=O(|k|^{-1})\quad\text{as }|k|\to\infty.$$

Remarks: (i) We will see in Sec. V that positive eigenvalues decouple from the scattering amplitude and do not give singularities of the cross section. This is a well-known result (see, e.g., Ref. 4); it is worth noticing that, in the time-dependent approach of Enss and Simon,[9] possible positive energy eigenvalues do not enter at any place.

(ii) Property (iii) is not as trivial as it looks; the main reason is that, although $1-P(k,k_0^2)$ projects out of the eigenvalue k_0^2 of $H(k)$, it is not an orthogonal projection in the intermediate space $\mathfrak{B}(\mathbb{R}^n)$, so that it does not necessarily have the same decoupling effect as in $L^2(\mathbb{R}^n)$.

The proof of Theorem 4.1 is perturbative and involves the following sequence of intermediate lemmas.

Lemma 4.2: (i) The limits

$$R_0(k,k^2)=\lim_{\epsilon\downarrow 0^+}R_0(k,k^2+i\epsilon)$$

exist in the weak operator topology of $\mathscr{L}(\mathfrak{B},\mathfrak{B}^*)$ for all $k\in\mathbb{C}^+\setminus\{0\}$ and are analytic in $k\in\mathbb{C}^+\setminus\mathbb{R}$ with weakly continuous boundary values on $\mathbb{R}\setminus\{0\}$.

(ii) $R_0(k,k^2)=R_0(-\bar{k},\bar{k}^2)^*\ \forall k\in\mathbb{C}^{++}$.

(iii) Let $\delta>0$. For $k\in\mathbb{C}^+$, $|k|\geqslant\delta$,

$$\|R_0(k,k^2)\|_{\mathscr{L}(\mathfrak{B},\mathfrak{B}^*)}\leqslant C|k|^{-1},$$

where C is a constant depending only on δ.

Proof: Existence of $R_0(k,k^2)$ for $k\in\mathbb{C}^+\setminus\{0\}$ in the weak operator topology of $\mathscr{L}(\mathfrak{B},\mathfrak{B}^*)$ follows as in Ref. 1 from an *a priori* estimate

$$|k|\ \|u\|_{\mathfrak{B}^*(\mathbb{R}^n)}\leqslant C\,\|[H_0(k)-k^2-i\epsilon]u\|_{\mathfrak{B}(\mathbb{R}^n)},$$
$$u\in\mathscr{H}^2(\mathbb{R}^n)\quad k\in\mathbb{C},\ |k|\geqslant\delta,\quad \epsilon\in\mathbb{R}, \tag{4.1}$$

which is proved in Appendix A.

Actually, (4.1) implies uniform boundedness in

$\mathscr{L}(\mathfrak{B}, \mathfrak{B}^*)$ of $R_0(k, k^2 + i\epsilon)$ for k in any compact subset of $\mathbb{C}^+ \setminus \{0\}$ and $\epsilon > 0$. Existence of the weak limits as ϵ tends to 0 follows from the consideration of the mean values $\langle u, R_0(k, k^2 + i\epsilon)v\rangle$, where $u,v \in \mathscr{C}_0^\infty(\mathbb{R}^n)$. We have, for $\epsilon > 0$,

$$\langle u, R_0(k, k^2 + i\epsilon)v\rangle$$
$$= \int_{\mathbb{R}^n} \frac{\bar{\hat{u}}(\xi)\hat{v}(\xi)}{(\xi + k\omega_-)^2 - k^2 - i\epsilon} d^n\xi$$
$$= \int_{\mathbb{R}^n} \frac{\bar{\hat{u}}(\xi - \bar{k}\omega_-)\hat{v}(\xi - k\omega_-)}{\xi^2 - k^2 - i\epsilon} d^n\xi,$$

where we have used analyticity of the Fourier transforms $\hat{u}$ and $\hat{v}$ to perform the complex change of variables. In the last integral, another deformation of contour allows to avoid the singularity at $\xi^2 = k^2$ when $k \in \mathbb{R} \setminus \{0\}$; then existence of the limits for $\epsilon = 0$ is obvious as well as analyticity of these limits in $k \in \mathbb{C}^+ \setminus \{0\}$. Now previously established uniform boundedness of $R_0(k, k^2 + i\epsilon)$ and density of $\mathscr{C}_0^\infty(\mathbb{R}^n)$ in $\mathfrak{B}(\mathbb{R}^n)$ implies existence of the weak limits $R_0(k, k^2)$ in $\mathscr{L}(\mathfrak{B}, \mathfrak{B}^*)$ for $k \in \mathbb{C}^+ \setminus \{0\}$ and continuity in k of these limits.

To prove analyticity in $\mathbb{C}^+ \setminus \mathbb{R}$, we can use again the density of $\mathscr{C}_0^\infty(\mathbb{R}^n)$ in $\mathfrak{B}(\mathbb{R}^n)$ and the analyticity of $\langle u_n, R_0(k, k^2)v_n\rangle$, for $u_n, v_n \in \mathscr{C}_0^\infty(\mathbb{R}^n)$, in $k \in \mathbb{C}^+ \setminus \mathbb{R}$. We have shown that such functions are uniformly bounded by $\|R_0(k, k^2)\|_{\mathscr{L}(\mathfrak{B}, \mathfrak{B}^*)} \|u_n\|_{\mathfrak{B}} \|v_n\|_{\mathfrak{B}}$. So if $u_n \to u$, $v_n \to v$ in $\mathfrak{B}(\mathbb{R}^n)$, then the limit $\langle u, R_0(k, k^2)v\rangle$ also is analytic in $\mathbb{C}^+ \setminus \mathbb{R}$ by a well-known result on limits of uniformly bounded sequences of analytic functions. From this follows that $R_0(k, k^2)$ is weakly analytic in $\mathbb{C}^+ \setminus \mathbb{R}$, and hence analytic (Ref. 18, p. 152–3).

The symmetry (ii) is obtained by taking the limit $\epsilon = 0$ in

$$R_0(k, k^2 + i\epsilon) = R_0(-\bar{k}, \bar{k}^2 - i\epsilon)^*, \quad \epsilon > 0.$$

Hence $R_0(k, k^2)$ obeys Schwarz reflection around $i\mathbb{R}^+$.

Finally (iii) follows from the *a priori* estimate (4.1) giving $\|R_0(k, k^2 + i\epsilon)\|_{\mathscr{L}(\mathfrak{B}, \mathfrak{B}^*)} \leqslant C|k|^{-1}$ for $|k| \geqslant \delta$ and the required estimate in the limit $\epsilon = 0$.

Let $K(k, z) = VR_0(k, z)$, $z \in \rho_0(k)$. Recall [see remark (ii) in Sec. IIB] that, if V is a short-range potential, then $\{K(k, z)\}$ is an analytic family of compact operators in $\mathfrak{B}(\mathbb{R}^n)$ as long as $z \in \rho_0(k)$.

Lemma 4.3: Let V be a short-range potential. Then the limits

$$K(k, k^2) = \lim_{\epsilon \downarrow 0^+} K(k, k^2 + i\epsilon), \quad k \in \mathbb{C}^+ \setminus \{0\},$$

exist in the norm operator topology of $\mathscr{L}(\mathfrak{B}, \mathfrak{B})$ and are analytic in $\mathbb{C}^+ \setminus \mathbb{R}$ with norm continuous boundary values on $\mathbb{R} \setminus \{0\}$. Furthermore, $K(k, k^2)$ is compact for $k \in \mathbb{C}^+ \setminus \{0\}$.

Proof: The standard compactness arguments of Agmon[1,6] imply with some simple modifications that $K(k, k^2)$ is compact on $L_s^2(\mathbb{R}^n)$ for $\frac{1}{2} < s < s_0$ and also on $\mathfrak{B}(\mathbb{R}^n)$.

Analyticity properties are straightforward consequences of analyticity results of Lemma 4.2.

Lemma 4.4: Let V be a short-range potential, $k \in \mathbb{C}^+ \setminus \mathbb{R}$. Then

$$K(k, k^2) = \lim_{\substack{k' \to k \\ |\mathrm{Im}\, k'| < \mathrm{Im}\, k}} K(k', k^2) \tag{4.2}$$

in the norm operator topology of $\mathscr{L}(\mathfrak{B}, \mathfrak{B})$. Furthermore, the eigenvalues of $K(k', k^2)$ are independent of k' for $|\mathrm{Im}\, k'| \leqslant \mathrm{Im}\, k$.

Proof: By the same type of arguments as in Lemma 4.2, one can show that

$$R_0(k, k^2) = \lim_{\substack{k' \to k \\ |\mathrm{Im}\, k'| < \mathrm{Im}\, k}} R_0(k', k^2)$$

in the weak topology of $\mathscr{L}(\mathfrak{B}, \mathfrak{B}^*)$. By the compactness properties of V, this implies, as in Lemma 4.3, the norm convergence property (4.2). Now let $\mathscr{O} = \{k' \in \mathbb{C}, |\mathrm{Im}\, k'| < \mathrm{Im}\, k\}$; the family $\{K(k', k^2), k' \in \mathscr{O}\}$ is an analytic family of compact operators in $\mathscr{L}(\mathfrak{B}, \mathfrak{B})$ since $k^2 \in \rho_0(k')$ for $k' \in \mathscr{O}$. Accordingly, the eigenvalues of $K(k', k^2)$ are branches of analytic functions.[18] But, since $B(\tau)$, $\tau \in \mathbb{R}$, is unitary on $\mathfrak{B}(\mathbb{R}^n)$, the operators $K(k' + \tau, k^2)$ and $K(k', k^2)$ are equivalent for all real τ. So eigenvalues of $K(k', k^2)$ only depend on $\mathrm{Im}\, k'$, which implies that they are constant. This concludes the proof.

Lemma 4.5: Let V a short-range potential. Then:

(i) $-1 \in \sigma_p(K(k, k^2))$, $k \in \mathbb{C}^+ \setminus \{0\}$, iff $k^2 \in \sigma_p(H)$.

(ii) Nonzero eigenvalues of H are isolated and have finite multiplicities. The corresponding eigenfunctions are in $L_s^2(\mathbb{R}^n)$, for any real s.

Proof: Assume first that $k \in \mathbb{R} \setminus \{0\}$. Since $B(k)$ is unitary on $\mathfrak{B}(\mathbb{R}^n)$, it follows that $K(k, k^2)$ has the same point spectrum as $K(0, k^2)$. By Lemma 4.4, this is also true for $k \in \mathbb{C}^+ \setminus \{0\}$. Then the statements of the lemma follow from the well-known Agmon bootstrap argument.[1,6]

The next result is a version of Klein–Zemach theorem[4]:

Lemma 4.6: Let V a bounded short-range potential. Then

(i) Let $\delta > 0$. For $k \in \mathbb{C}^+$, $|k| \geqslant \delta$, one has

$$\|K(k, k^2)\|_{\mathscr{L}(\mathfrak{B}, \mathfrak{B})} \leqslant C|k|^{-1},$$

where C is a constant depending only on δ.

(ii) $\sigma_p(H)$ is a bounded set.

Proof: (i) This is an immediate consequence of Lemma 4.2 (iii).

(ii) This follows directly from (i) and Lemma 4.5 (i).

Now we come to the following:

Proof of Theorem 4.1: Since $H(k) = H_0(k) + V$, one has by the second resolvent equation

$$R(k, k^2 + i\epsilon)(1 + K(k, k^2 + i\epsilon)) = R_0(k, k^2 + i\epsilon) \tag{4.3}$$

for $\epsilon > 0$. Assume that $k \in \mathbb{C}^+ \setminus \Sigma$; then, by Lemma 4.5 and the Fredholm alternative, $1 + K(k, k^2)$ is invertible and, by Lemma 4.3, its inverse $(1 + K(k, k^2))^{-1}$ is the norm limit of $(1 + K(k, k^2 + i\epsilon))^{-1}$ as $\epsilon \downarrow 0^+$. Hence, by Lemma 4.2(i),

$$R(k, k^2 + i\epsilon) = R_0(k, k^2 + i\epsilon)(1 + K(k, k^2 + i\epsilon))^{-1}$$

converges in the weak operator topology of $\mathscr{L}(\mathfrak{B}, \mathfrak{B}^*)$ as $\epsilon \downarrow 0^+$ and

$$R(k, k^2) = R_0(k, k^2)(1 + K(k, k^2))^{-1}. \tag{4.4}$$

Then, by the analytic Fredholm theorem, Lemmas 4.2(i), 4.3, and 4.5, $R(k, k^2)$ is meromorphic in $\mathbb{C}^+\setminus\mathbb{R}$ with poles on $(\Sigma\cap i\mathbb{R}^+)\setminus\{0\}$ and has continuous boundary values on $\mathbb{R}\setminus\Sigma$.

This shows (i) and (ii) of Theorem 4.1.

Property (v) follows as for $R_0(k, k^2)$ [Lemma 4.2(ii)] from taking the limit $\epsilon = 0$ in $R(k, k^2 + i\epsilon) = R(-\bar{k}, \bar{k}^2 - i\epsilon)^*$, $\epsilon > 0$.

The estimate in (vi) follows from (4.4), Lemma 4.2(iii) and Lemma 4.6(i), while the boundedness of $\sigma_p(H)$ is just Lemma 4.6(ii).

The proof of (iii) is given in Appendix B.

Finally, we want to investigate the poles of $R(k, k^2)$ on $i\mathbb{R}^+$. Let $k_0^2\in[\sigma_p(H)\cap\mathbb{R}^-]\setminus\{0\}$. Consider first $P(k, k_0^2)$ for $k\in\mathbb{R}$; by the O'Connor–Combes–Thomas theorem,[14] $P(k, k_0^2)$ is analytic in $\{k\in\mathbb{C}, |\mathrm{Im}\, k| < |k_0|\}$ as a compact operator-valued family in $\mathscr{L}(L^2, L^2)$, hence also in $\mathscr{L}(\mathfrak{B}, \mathfrak{B}^*)$. Consider now the Laurent expansion of $R(k, k^2)$ around k_0, which we can write in the form

$$R(k, k^2) = \sum_{l=1}^{N} \frac{D_l}{(k^2 - k_0^2)^l} + \hat{R}(k, k^2),$$

where $D_l \in \mathscr{L}(\mathfrak{B}, \mathfrak{B}^*)$ and $\hat{R}(k, k^2)$ is the regular part of $R(k, k^2)$ in the neighborhood of k_0. Let $u, v\in\mathscr{C}_0^\infty(\mathbb{R}^n)$ and $u(k) = B(k)u$, $v(k) = B(k)v$, $k\in\mathbb{R}$. These $\mathfrak{B}(\mathbb{R}^n)$-valued functions obviously have analytic continuations to $\mathbb{C}$. Consider, for $k\in\mathbb{C}^+\setminus\Sigma$:

$$\begin{aligned} g(k) &= \langle u, R(k, k^2)v\rangle \\ &= \sum_{l=1}^{N} \frac{\langle u, D_l v\rangle}{(k^2 - k_0^2)^l} + \langle u, \hat{R}(k, k^2)v\rangle. \end{aligned} \tag{4.5}$$

For k real, one has also, by unitarity of $B(k)$:

$$g(k) = \langle u(\bar{k}), R(0, k^2)v(k)\rangle, \tag{4.6}$$

where $R(0, k^2)$ denotes the weak limit in $\mathscr{L}(\mathfrak{B}, \mathfrak{B}^*)$ of $(H - k^2 - i\epsilon)^{-1}$ as $\epsilon\downarrow 0^+$. The rhs of (4.6) has a meromorphic continuation to $\mathbb{C}^+$ since $u(\bar{k})$, $v(k)$ and $R(0, k^2)$ do; around the pole k_0, one has, from the spectral theorem, an expansion

$$\begin{aligned} g(k) &= \frac{\langle u(\bar{k}), P\{k_0^2\}v(k)\rangle}{k_0^2 - k^2} \\ &\quad + \langle u(\bar{k}), \hat{R}(0, k^2)v(k)\rangle, \end{aligned}$$

where $\hat{R}(0, k^2)$ is the reduced resolvent of H. Comparing with (4.5), one gets $D_l = 0$ for $l > 1$ and

$$\langle u, D_1 v\rangle = -\langle u(\bar{k}_0), P\{k_0^2\}v(k_0)\rangle. \tag{4.7}$$

Now, for k real,

$$\langle u(\bar{k}), P\{k_0^2\}v(k)\rangle = \langle u, P(k, k_0^2)v\rangle.$$

Since both terms in this equality have analytic continuations to $\{k\in\mathbb{C}, |\mathrm{Im}\, k| < |k_0|\}$, equality also holds in this strip, and therefore (4.7) gives

$$\langle u, D_1 v\rangle = -\lim_{\substack{k\to k_0 \\ |\mathrm{Im}\, k| < |k_0|}} \langle u, P(k, k_0^2)v\rangle$$

showing in particular, since $D_1\in\mathscr{L}(\mathfrak{B}, \mathfrak{B}^*)$, that the weak limit of $P(k, k_0^2)$ as $k\to k_0$, $|\mathrm{Im}\, k| < |k_0|$, exists in $\mathscr{L}(\mathfrak{B}, \mathfrak{B}^*)$.

V. PROPERTIES OF THE CROSS SECTION

In this section and in the following, we will consider potentials V such that

$$\text{(A)}\quad |V(x)| \leqslant C(1 + |x|)^{-(n+1)/2}(\log(2 + |x|))^{-1-\epsilon},\quad C,\ \epsilon > 0.$$

We will investigate mostly the non-Born part $t(k; \omega_-)$ of the forward scattering amplitude and the total scattering cross section $\sigma(k; \omega_-)$ at fixed incident direction $\omega_-\in S^{n-1}$.

For $k\in\mathbb{R}^+\setminus\{0\}$,

$$t(k; \omega_-) = \frac{1}{(2\pi)^n} \lim_{\epsilon\downarrow 0^+} \langle Ve^{ik\omega_-\cdot x}, R(0, k^2 + i\epsilon)Ve^{ik\omega_-\cdot x}\rangle.$$

By (3.4), one has

$$t(k; \omega_-) = \frac{1}{(2\pi)^n} \langle V, R(k, k^2)V\rangle. \tag{5.1}$$

The cross section $\sigma(k; \omega_-)$ is related to the imaginary part of $t(k; \omega_-)$ by the optical theorem which will be stated below. So let us show first:

Theorem 5.1: Let V satisfy assumption (A). Then
(i) $t(k; \omega_-)$ is finite and continuous on $\mathbb{R}^+\setminus\{0\}$.
(ii) $|t(k; \omega_-)| = O(k^{-1})$ as $k\to\infty$.

Proof: Since $V\in\mathfrak{B}(\mathbb{R}^n)$, continuity of $t(k; \omega_-)$ on $\mathbb{R}^+\setminus\Sigma$ follows from Theorem 4.1(i). So we only need to investigate $t(k; \omega_-)$ for k in the neighborhood of any point $k_0\in\mathbb{R}^+\setminus\{0\}$ such that $k_0^2\in\sigma_p(H)$. For k near k_0, $k\neq k_0$, we make in (5.1) the splitting

$$\begin{aligned} t(k; \omega_-) &= \frac{1}{(2\pi)^n} [(k_0^2 - k^2)^{-1}\langle V, P(k, k_0^2)V\rangle \\ &\quad + \langle V, \hat{R}(k, k^2)V\rangle]. \end{aligned}$$

The second term on the right-hand side is continuous at k_0 by Theorem 4.1(iii). Let us show that the first one, which can be rewritten as

$$\langle Ve^{ik\omega_-\cdot x}, P\{k_0^2\}Ve^{ik\omega_-\cdot x}\rangle$$

vanishes when k approaches k_0.

For this, since $P\{k_0^2\}$ is finite dimensional, it is enough to verify that, if $\varphi_0\in L^2(\mathbb{R}^n)$ satisfies $(H - k_0^2)\varphi_0 = 0$, then

$$\lim_{k\to k_0} (k_0^2 - k^2)^{-1}|\langle V, \varphi_0 e^{-ik\omega_-\cdot x}\rangle|^2 = 0 \quad \text{when } \omega_-\in S^{n-1}.$$

Since $\mathscr{D}(H) = \mathscr{D}(H_0)$, one has $(H_0 - k_0^2)\varphi_0 = -V\varphi_0$; as stated in Lemma 4.5, the Fourier transform $\hat{\varphi}_0$ of φ_0 is in $\mathscr{C}_\infty(\mathbb{R}^n)$ and, accordingly, $\widehat{(V\varphi_0)}(k\omega_-) = (k_0^2 - k^2)\hat{\varphi}_0(k\omega_-)$. So

$$(k_0^2 - k^2)^{-1}|\langle V, \varphi_0 e^{-ik\omega_-\cdot x}\rangle|^2 = (2\pi)^n(k_0^2 - k^2)|\hat{\varphi}_0(k\omega_-)|^2,$$

which has a zero limit as $k\to k_0$. This concludes the proof of (i).

The asymptotic estimate (ii) follows immediately from (5.1) and theorem 4.1(vi).

The method used in the proof of Theorem 5.1 works as well for the analysis of the non-Born term $t(k; \omega_-, \omega_+)$ given by (3.1). One obtains in this way the following:

Theorem 5.2: Let V satisfy assumption (A). Then:
(i) $t(k; \omega_-, \omega_+)$ is finite and continuous on $(\mathbb{R}^+\setminus\{0\})\times S^{n-1}\times S^{n-1}$.

(ii) $|t(k;\omega_-,\omega_+)| = O(k^{-1})$ as $k\to\infty$, uniformly in $\omega_\pm \in S^{n-1}$.

Proof: With our previous notational convention (3.4), one can rewrite $t(k;\omega_-,\omega_+)$ as

$$t(k;\omega_-,\omega_+) = [1/(2\pi)^n]\langle V(k;\omega_-,\omega_+), R(k,k^2)V\rangle \tag{5.2}$$

with $V(k;\omega_-,\omega_+)(x) = V(x)e^{ik(\omega_+-\omega_-)\cdot x}$.

Now $V(k;\omega_-,\omega_+)$ is strongly continuous in $\mathfrak{B}(\mathbb{R}^n)$ for $(k;\omega_-,\omega_+)\in(\mathbb{R}^+\setminus\{0\})\times S^{n-1}\times S^{n-1}$; the verification is elementary with the definition of $\mathfrak{B}(\mathbb{R}^n)$. This gives the continuity of $t(k;\omega_-,\omega_+)$ for $(k;\omega_-,\omega_+)\in(\mathbb{R}^+\setminus\Sigma)\times S^{n-1}\times S^{n-1}$. Near $k_0\in\mathbb{R}^+\setminus\{0\}$ such that $k_0^2\in\sigma_p(H)$, one can decompose as before

$$\begin{aligned} t(k;\omega_-,\omega_+) &= [1/(2\pi)^n][(k_0^2-k^2)^{-1}\langle V(k;\omega_-,\omega_+),P(k,k_0^2)V\rangle \\ &\quad + \langle V(k;\omega_-,\omega_+),\hat{R}(k,k^2)V\rangle].\end{aligned}$$

The second term is continuous at k_0 by Theorem 4.1(iii). The first one can be rewritten as

$$(k_0^2-k^2)^{-1}\langle V(k;0,\omega_+), P\{k_0^2\}V(k;-\omega_-,0)\rangle;$$

one shows that its limit at $k=k_0$ vanishes as in the proof of Theorem 5.1(i).

Finally (ii) follows from (5.2) and theorem 4.1(vi).

We now turn to the analysis of the cross section $\sigma(k;\omega_-)$ given by (3.2). However, contrarily to the case of $\sigma_B(k;\omega_-)$, the integral form of $\sigma(k;\omega_-)$ in terms of F is not very convenient for our purpose. A standard consequence of the unitarity of the S matrix is the "optical theorem" stating that

$$\sigma(k;\omega_-) = (2C_n^2/\pi k)\,\mathrm{Im}\,t(k;\omega_-). \tag{5.3}$$

It is a rather elementary exercise to justify (5.3) under our assumption on V, using the connection between the kernel of the S matrix and the function $T(k;\omega_-,\omega_+)$, namely

$$(S-I)(k\omega_+,k'\omega_-) = -2i\pi\delta(k^2-k'^2)T(k;\omega_-,\omega_+).$$

From (5.3) and Theorem 5.1 follows our main result on the cross section:

Theorem 5.3: Let V satisfy assumption (A). Then:
(i) $\sigma(k;\omega_-)$ is finite and continuous on $\mathbb{R}^+\setminus\{0\}$.
(ii) $\sigma(k;\omega_-) = O(k^{-2})$ as $k\to+\infty$.

Remark: By (3.3) and Lemma 4.2(i), $\sigma_B(k;\omega_-)$ is also continuous on $\mathbb{R}^+\setminus\{0\}$.

To conclude this section, let us give an estimate of the correction $\sigma(k;\omega_-)-\sigma_B(k;\omega_-)$ to the Born approximation of the cross section.

Theorem 5.4: Let V satisfy assumption (A). Then

$$|\sigma(k;\omega_-)-\sigma_B(k;\omega_-)| = O(k^{-3}) \quad \text{as } k\to\infty. \tag{5.4}$$

Proof: By the optical theorems (5.3) and (3.3), one has

$$\begin{aligned}\sigma(k;\omega_-)-\sigma_B(k;\omega_-) &= \frac{2C_n^2}{\pi k}\frac{1}{(2\pi)^n}\mathrm{Im}\langle V[R(k,k^2)-R_0(k,k^2)]V\rangle \\ &= -\frac{2C_n^2}{\pi k}\frac{1}{(2\pi)^n}\mathrm{Im}\langle V,R(k,k^2)K(k,k^2)V\rangle.\end{aligned}$$

In the last equality, we have used the resolvent equation (4.4). Now, by Theorem 4.1(vi) and Lemma 4.6(i),

$$\langle V,R(k,k^2)K(k,k^2)V\rangle = O(k^{-2}) \quad \text{as } k\to+\infty.$$

This completes the proof.

Remark: As noticed in Sec. VII, the estimate (5.4) is weakened if V has local singularities.

VI. ANALYTICITY PROPERTIES OF THE FORWARD SCATTERING AMPLITUDE

Continuation of the scattering amplitude in the complex plane is a fundamental problem in scattering theory, leading, if one has a suitable low energy behavior (see the Introduction), to the powerful dispersion relations

$$\mathrm{Re}\,t(k;\omega_-) = \frac{2}{\pi}\int_0^{+\infty}\frac{k'\,\mathrm{Im}\,t(k';\omega_-)}{k'^2-k^2}dk' + \sum_j\frac{r_j}{k^2-k_j^2}. \tag{6.1}$$

The sum in the rhs extends over $k_j^2\in[\sigma_p(H)\cap\mathbb{R}^-]\setminus\{0\}$ and r_j is the "residue" given by (6.2) below. For $n\geqslant 3$, the sum is finite under our assumption (A) on V. We refer to Ref. 5 for a derivation of (6.1). For potential scattering, $n=3$, the analyticity problem is reviewed in Simon's monograph.[4] Our approach allows a unified treatment in any dimension; since we consider only the non-Born part of the scattering amplitude, our assumption (A) on the potential weakens the traditional decay assumption

$$V(x) = O(|x|^{-3-\epsilon}), \quad \epsilon>0, \quad \text{as } |x|\to\infty,$$

in the case $n=3$. Indeed, since the Born part is real, it does not enter into the dispersion relations which involve only the imaginary part of the forward amplitude, whose quotient by k is proportional to the scattering cross section by (5.3). In this way, our decay assumption (A) on the potential appears as optimal as long as physical dispersion relations are concerned.

The main statement of this section is:

Theorem 6.1: Let V satisfy assumption (A). Then:

(i) $t(k;\omega_-)$ has a meromorphic continuation to $\mathbb{C}^+$ with simple poles on $(\Sigma\cap i\mathbb{R}^+)\setminus\{0\}$.

(ii) If $k_j^2\in[\sigma_p(H)\cap\mathbb{R}^-]\setminus\{0\}$, then the residue at $k=k_j$ is given by

$$(2k_j)^{-1}r_j = -[1/(2\pi)^n]\langle V,P(k_j,k_j^2)V\rangle, \tag{6.2}$$

where $P(k_j,k_j^2)\in\mathscr{L}(\mathfrak{B},\mathfrak{B}^*)$ is defined in Theorem 4.1(iv).

(iii) $t(k;\omega_-) = \overline{t(-\bar{k};\omega_-)}, \quad k\in\mathbb{C}^+\setminus\Sigma.$

(iv) $|t(k;\omega_-)| = O(|k|^{-1})$ as $|k|\to\infty$, $k\in\mathbb{C}^+$.

Proof: Since $V\in\mathfrak{B}(\mathbb{R}^n)$ (see Sec. IIB), Theorem 6.1 immediately follows from (3.4) and Theorem 4.1, the meromorphic continuation of $t(k;\omega_-)$ being provided by $[1/(2\pi)^n]\langle V,R(k,k^2)V\rangle$.

VII. HIGH-ENERGY BEHAVIORS FOR SINGULAR POTENTIALS

In the previous sections, we have considered only bounded potentials. Here, we allow local singularities of the type $1/|x-x_0|^\alpha$, $\alpha<2$. For simplicity, we will consider only one singularity at $x_0=0$; with more work it is possible to extend the results to a finite number of singular centers. Notice first that the restriction $\alpha<2$ allows us to define self-adjoint realizations of $-\Delta+V$ on $L^2(\mathbb{R}^n)$; we refer to Ref. 15 for the various ways to define a self-adjoint H in this case.

For $s\in\mathbb{R}$, we define $\rho_s(x)=1/|x|^s$ and $\tilde\rho_s(x)=1/(1+|x|)^s$; one has

Theorem 7.1: Assume that $V(x)=v(x)/|x|^\alpha$ with $v\in L^\infty(\mathbb{R}^n)$ and $1<\alpha<2$. Then, for any s, $\frac12<s<1$,

(i) $\|\tilde\rho_s R(k,k^2)\tilde\rho_s\|=O(|k|^{-1})$ as $|k|\to\infty$, $k\in\mathbb{C}^+$,

(ii) $\|\rho_s R(k,k^2)\rho_s\|=O(|k|^{-2+2s})$

as $|k|\to\infty$, $k\in\mathbb{C}^+$.

Remark: Obviously, this result applies as well to short-range potentials having a weaker singularity $1/|x|^\alpha$, $\alpha\leqslant 1$. It shows that the conditions on the potential imposed in Theorem 4.1(vi) can certainly be drastically weakened without altering the asymptotic behavior of $R(k,k^2)$.

Proof of Theorem 7.1: Write $V=UW$, with $U=v\rho_{\alpha/2}$ and $W=\rho_{\alpha/2}$. Let $\eta>0$. (i) By the resolvent Eq. (4.4), for $k\in\mathbb{C}^+$ and $s\in\mathbb{R}$,

$$\tilde\rho_s R(k,k^2)\tilde\rho_s=\tilde\rho_s R_0(k,k^2)\tilde\rho_s-\tilde\rho_s R(k,k^2)UWR_0(k,k^2)\tilde\rho_s. \quad (7.1)$$

By Lemma 4.2(iii),

$$\|\tilde\rho_s R_0(k,k^2)\tilde\rho_s\|\leqslant C|k|^{-1},\quad k\in\mathbb{C}^+,\quad |k|\geqslant\eta. \quad (7.2)$$

Applying Theorem C.1 to $\tilde\rho_s\leqslant\rho_s$, $\frac12<s<1$, we get

$$\|WR_0(k,k^2)\tilde\rho_s\|\leqslant C'|k|^{-2+\alpha/2+s},\quad k\in C^+,\quad |k|\geqslant\eta,$$

and

$$\|\tilde\rho_s R(k,k^2)U\|=O(|k|^{-2+\alpha/2+s})\quad \text{as } |k|\to\infty,\quad k\in\mathbb{C}^+.$$

Therefore,

$$\|\tilde\rho_s R(k,k^2)UWR_0(k,k^2)\tilde\rho_s\|=O(|k|^{-4+\alpha+2s})\quad \text{as } |k|\to\infty,\quad k\in\mathbb{C}^+. \quad (7.3)$$

So (i) holds if $1<2s<3-\alpha$ and, since $\|\tilde\rho_s R(k,k^2)\tilde\rho_s\|$ is decreasing in s, it holds for all $\frac12<s<1$.

(ii) Replacing $\tilde\rho_s$ by ρ_s in (7.1), we have, for $s\in\mathbb{R}$,

$$\rho_s R(k,k^2)\rho_s=\rho_s R_0(k,k^2)\rho_s-\rho_s R(k,k^2)UWR_0(k,k^2)\rho_s. \quad (7.1')$$

By Theorem C.1(i), for $\frac12<s<1$,

$$\|\rho_s R_0(k,k^2)\rho_s\|\leqslant C|k|^{-2+2s},\quad k\in\mathbb{C}^+,\quad |k|\geqslant\eta. \quad (7.2')$$

Applying now Theorem C.1, we get

$$\|\rho_s R(k,k^2)UWR_0(k,k^2)\rho_s\|=O(|k|^{-4+\alpha+2s})\quad \text{as } |k|\to\infty,\quad k\in\mathbb{C}^+. \quad (7.3')$$

Since $\alpha<2$, applying (7.2') and (7.3') in (7.1'), we obtain the desired result.

Applying Theorem 7.1 to $t(k;\omega_-)$ and $t(k;\omega_-,\omega_+)$, we obtain

Theorem 7.2: Assume that V satisfies the following condition: $V(x)=v(x)/|x|^\alpha$, with $v\in L^\infty(\mathbb{R}^n)$, $n\geqslant 3$, and $1<\alpha<2$, and, for some $\epsilon>0$, $V(x)=O(|x|^{-(n+1)/2-\epsilon})$ as $|x|\to\infty$.

Then:

(i) If $1<\alpha<\frac32$, $|t(k;\omega_-)|=O(|k|^{-1})$, $|k|\to\infty$, $k\in\mathbb{C}^+$.

(ii) If $\frac32\leqslant\alpha<2$, $|t(k;\omega_-)|=O(|k|^{-1+\delta})$ for any $\delta>O$, as $|k|\to\infty$, $k\in\mathbb{C}^+$.

The same result holds for $t(k;\omega_-,\omega_+)$, as $k\to+\infty$, uniformly in $\omega_\pm\in S^{n-1}$.

Before proving Theorem 7.2, we can state its immediate consequence, by means of the optical theorem (5.3):

Theorem 7.3: Assume that V satisfies the condition of Theorem 7.2. Then:

(i) If $1<\alpha<\frac32$, $\sigma(k;\omega_-)=O(k^{-2})$ as $k\to+\infty$.

(ii) If $\frac32\leqslant\alpha<2$, $\sigma(k;\omega_-)=O(k^{-2+\delta})$, for any $\delta>0$, as $k\to+\infty$.

Now we come to the proof:

Proof of Theorem 7.2: (i) By (3.4), this is a direct consequence of Theorem 7.1(i), since, in this case, $V\in L_s^2(\mathbb{R}^n)$ for any s, $\frac12<s<\frac12+\epsilon$.

(ii) By (3.4), this is a direct consequence of Theorem 7.1(ii). Indeed, in this case, $\rho_{-s}V\in L^2(\mathbb{R}^n)$ for $s=\frac12+\delta/2$, $\delta/2<\epsilon$, and arbitrarily small, and we can write

$$\langle V,R(k,k^2)V\rangle=\langle\rho_{-s}V,\rho_s R(k,k^2)\rho_s(\rho_{-s}V)\rangle.$$

Remark: For $n\geqslant4$, $|t(k;\omega_-)|=O(|k|^{-1})$ as $|k|\to\infty$, $k\in\mathbb{C}^+$, for any $1<\alpha<2$ since, in this case, $V\in L_s^2(\mathbb{R}^n)$ for any s, $\frac12<s<\frac12+\epsilon$.

Finally, we prove the asymptotic estimate of the correction to the Born approximation of $\sigma(k;\omega_-)$ which, for any $\alpha<2$, is weaker than in the case of bounded potentials:

Theorem 7.4: Assume that V satisfies the condition of Theorem 7.2. Then $|\sigma(k;\omega_-)-\sigma_B(k;\omega_-)|=O(k^{-4+\alpha+\delta})$ for any $\delta>0$ as $k\to+\infty$.

Remarks: (i) Comparing Theorems 7.3 and 7.4 and also 3.1, we see that $\sigma(k;\omega_-)-\sigma_B(k;\omega_-)$ is really a correction to $\sigma_B(k;\omega_-)$ since, in any case, the former decreases more rapidly than the latter as $k\to+\infty$.

(ii) In general, it is not clear whether the estimate for $\alpha>1$ can be improved by another method. In fact, it seems reasonable to expect that, the stronger the singularities of V are, the slower is the decay of $\sigma(k;\omega_-)-\sigma_B(k;\omega_-)$ for large k; this is suggested in particular by the analysis of Jensen who obtained in Ref. 11 another estimate in the case of the cross sections averaged over all incident directions in $\mathbb{R}^3$. The potentials considered by Jensen are of the type $V(x)=(1+|x|^2)^{-\beta}v(x)$ with $\beta>1$ and $v\in L^p(\mathbb{R}^3)$, $\frac32\leqslant p\leqslant+\infty$, which locally corresponds exactly to our condition $\alpha<2$; so (for $n=3$), he covers a larger class than ours in the sense that the local singularities do not have to be in finite number. However, his result for the averaged cross sections $\bar\sigma(k)$ and $\bar\sigma_B(k)$,

$$|\bar\sigma(k)-\bar\sigma_B(k)|=O(k^{-3+3/2p})\quad \text{as } k\to+\infty, \quad (7.4)$$

coincides with ours only for $p=+\infty$, that is $\alpha=0$ (see our Theorem 5.4) and is weaker for $\frac32\leqslant p<+\infty$, that is $0<\alpha<2$. For example, for $\alpha=1$ (Coulomb and Yukawa potentials), (7.4) becomes

$|\bar{\sigma}(k) - \bar{\sigma}_B(k)| = O(k^{-5/2+\delta})$, for any $\delta > 0$, as $k \to \infty$, while our Theorem 7.4 gives

$$|\sigma(k;\omega_-) - \sigma_B(k;\omega_-)| = O(k^{-3+\delta}) \quad \text{for any } \delta > 0, \text{ as } k \to \infty.$$

Proof of Theorem 7.4: As in the proof of Theorem 5.4, one has

$$\sigma(k;\omega_-) - \sigma_B(k;\omega_-) = -\frac{2C_n^2}{\pi k}\frac{1}{(2\pi)^n}\,\mathrm{Im}\langle V, R(k,k^2)UWR_0(k,k^2)V\rangle.$$

Using the same trick as in the proof of Theorem 7.2, this implies that, for $s = \frac{1}{2} + \delta/2$, $\delta/2 < \epsilon$ and arbitrarily small,

$$|\sigma(k;\omega_-) - \sigma_B(k;\omega_-)| \leqslant \frac{2C_n^2}{\pi k}\frac{1}{(2\pi)^n}\,\|\rho_{-s}V\|^2\|\rho_s R(k,k^2)UWR_0(k,k^2)V\|.$$

So, by (7.3′)

$$|\sigma(k;\omega_-) - \sigma_B(k;\omega_-)| = O(k^{-4+\alpha+\delta}) \quad \text{for any } \delta > 0 \text{ as } k \to +\infty.$$

VIII. DILATION ANALYTIC POTENTIALS: ANALYTIC CONTINUATION ON THE UNPHYSICAL SHEET

We recall that a potential is said to be dilation analytic if the family $\{V_\lambda, \lambda \in \mathbb{R}^+\}$, where $V_\lambda(x) = V(\lambda^{-1}x)$, has an analytic continuation as a family of compact operators from $\mathscr{H}^2(\mathbb{R}^n)$ to $L^2(\mathbb{R}^n)$. We refer to Ref. 19 for the investigation of the family $\{H(\lambda) = \lambda^2 H_0 + V_\lambda\}$ and in particular a proof that it is analytic of type A with essential spectrum $\lambda^2\mathbb{R}^+$, its point spectrum containing, in addition to the point spectrum of H, some complex eigenvalues in the sector $\{z \in \mathbb{C}, 2\arg\lambda < \arg z < 0\}$, in case $\arg\lambda < 0$. Such complex eigenvalues are commonly interpreted as resonances.[20] Actually, we could now repeat the arguments of Sec. IV, using the dilation group instead of the boost group, provided the potential is dilation analytic. We refer to Ref. 14(b) for details. The important point is that this extra assumption allows to reach regions $\{k \in \mathbb{C}, \theta_0 < \arg k < 0\}$ and to prove meromorphy of the scattering amplitude in these regions, with possible poles only at resonance eigenvalues. We will not develop the detailed proofs; they are developed and used in Ref. 21 in the special case of electron-atom scattering.

ACKNOWLEDGMENTS

It is a pleasure for us to thank A. Jensen for several discussions.

APPENDIX A: *A PRIORI* ESTIMATE FOR $H_0(k)$

Following Agmon,[1] we establish the *a priori* estimate used in the proof of Lemma 4.2(iii). In fact, we prove it for any $k = (k_1, ..., k_n) \in \mathbb{C}^n$ and not only for $k\omega_-$, where $k \in \mathbb{C}$ and ω_- is a unit vector in $\mathbb{R}^n$.

Before this, we need to list some properties of Besov norms.[2] First, recall the following two equivalent norms on $\mathfrak{B}(\mathbb{R}^n)$:

$$\|u\|_{\mathfrak{B}^*} = \sup_{j>0}\left[\frac{1}{R_j}\int_{\Omega_j}|u(x)|^2\,dx\right]^{1/2}$$

(see Sec. II for the definitions of R_j and Ω_j) and

$$\|u\|'_{\mathfrak{B}^*} = \sup_{R>1}\left[\frac{1}{R}\int_{|x|<R}|u(x)|^2\,dx\right]^{1/2}.$$

Let now $\varphi \in \mathscr{C}_0^\infty(\mathbb{R}^n)$, $\varphi \geqslant 0$, $\varphi(x) = 1$ for $|x| \leqslant 1$ and $\operatorname{supp}\varphi \subset \{x \in \mathbb{R}^n, |x| < 2^N\}$ for some $N \geqslant 1$.

Define $\varphi^\gamma(x) = \varphi(\gamma x)$ for $0 < \gamma \leqslant 1$,

$$\varphi_R^\gamma(x) = \frac{1}{R^{1/2}}\varphi^\gamma\left(\frac{x}{R}\right) \quad \text{for } R > 1$$

and

$$\|u\|_{\mathfrak{B}^*}^{\varphi^\gamma} = \sup_{R>1}\|\varphi_R^\gamma u\|_{L^2}.$$

If χ is the characteristic function of $\operatorname{supp}\varphi$, define similarly χ^γ, χ_R^γ, and $\|u\|_{\mathfrak{B}^*}^{\chi^\gamma}$.

One can easily prove that

$$\tfrac{1}{2}\|u\|'_{\mathfrak{B}^*} \leqslant \|u\|_{\mathfrak{B}^*} \leqslant \|u\|'_{\mathfrak{B}^*} \tag{A1}$$

and

$$\|u\|'_{\mathfrak{B}^*} \leqslant \|u\|_{\mathfrak{B}^*}^{\chi^\gamma} \leqslant \|u\|_{\mathfrak{B}^*}^{\varphi^\gamma}, \tag{A2}$$

and it is not much more difficult to show that we have also

$$\|u\|_{\mathfrak{B}^*}^{\varphi^\gamma} \leqslant (2^{N/2+1}/\gamma^{1/2})\|u\|'_{\mathfrak{B}^*} \tag{A3}$$

and

$$\|u\|_{\mathfrak{B}^*}^{\chi^\gamma} \leqslant (2^{N/2}/\gamma^{1/2})\|u\|'_{\mathfrak{B}^*}. \tag{A3′}$$

Therefore, all the norms $\|\cdot\|_{\mathfrak{B}^*}$, $\|\cdot\|'_{\mathfrak{B}^*}$, $\|\cdot\|_{\mathfrak{B}^*}^{\varphi^\gamma}$, and $\|\cdot\|_{\mathfrak{B}^*}^{\chi^\gamma}$ are equivalent.

Now we can state:

Theorem A.1: Let $\delta > 0$. Then, for any $u \in \mathscr{H}^2(\mathbb{R}^n)$, $k \in \mathbb{C}^n$, $|k| \geqslant \delta$, and $\epsilon \in \mathbb{R}$,

$$|k|\,\|u\|_{\mathfrak{B}^*(\mathbb{R}^n)} \leqslant C\,\|(H_0(k) - k^2 - i\epsilon)u\|_{\mathfrak{B}^*(\mathbb{R}^n)}, \tag{A4}$$

where C is a constant depending only on δ.

For the proof of this theorem, we need four lemmas.

Lemma A.2: For any $u \in \mathscr{H}^1(\mathbb{R})$ and $\lambda \in \mathbb{C}$,

$$\|u\|_{\mathfrak{B}^*(\mathbb{R})} \leqslant \left\|\left(\frac{d}{dx} - \lambda\right)u\right\|_{L^1(\mathbb{R})}. \tag{A5}$$

Proof: We can assume $\mathrm{Re}\,\lambda \leqslant 0$ without loss of generality. Let $f = (d/dx - \lambda)u$.

If $f \notin L^1(\mathbb{R})$, then the right-hand side of (A5) is infinite and there is nothing to prove. Suppose then $f \in L^1(\mathbb{R})$. We have

$$u(x) = \int_{-\infty}^{x} f(t)e^{\lambda(x-t)}\,dt.$$

Thus

$$|u(x)| \leqslant \int_{-\infty}^{+\infty}|f(t)|\,dt = \|f\|_{L^1(\mathbb{R})}.$$

So

$$\int_{\Omega_j}|u(x)|^2\,dx \leqslant R_j\|f\|_{L^1(\mathbb{R})}^2 \quad \forall j > 0$$

and finally

$$\sup_{j>0} \frac{1}{R_j} \int_{\Omega_j} |u(x)|^2 \, dx \leqslant \|f\|^2_{L^1(\mathbb{R})}.$$

Lemma A.3: Let $x = (x_1, x')$, where $x' = (x_2, ..., x_n)$. Then, for any $u \in \mathfrak{B}(\mathbb{R}^n)$,

$$\int_{-\infty}^{+\infty} \|u(x_1, \cdot)\|_{L^2(\mathbb{R}^{n-1})} \, dx_1 \leqslant \sqrt{2} \|u\|_{\mathfrak{B}(\mathbb{R}^n)}.$$

Proof: Set $u = \Sigma_{j=1}^{\infty} u_j$, where $u_j = \chi_j u$ and χ_j is the characteristic function of Ω_j.

Then

$$\int_{-\infty}^{+\infty} \|u(x_1, \cdot)\|_{L^2(\mathbb{R}^{n-1})} \, dx_1 \leqslant \sum_{j=1}^{\infty} \int_{-\infty}^{+\infty} \|u_j(x_1, \cdot)\|_{L^2(\mathbb{R}^{n-1})} \, dx_1.$$

For any j, since $u_j(x_1, x') = 0$ for any $x' \in \mathbb{R}^{n-1}$ if $|x_1| > R_j$,

$$\int_{-\infty}^{+\infty} \|u_j(x_1, \cdot)\|_{L^2(\mathbb{R}^{n-1})} \, dx_1 = \int_{-R_j}^{R_j} \|u_j(x_1, \cdot)\|_{L^2(\mathbb{R}^{n-1})} \, dx_1.$$

By Schwarz inequality, this gives

$$\int_{-\infty}^{+\infty} \|u_j(x_1, \cdot)\|_{L^2(\mathbb{R}^{n-1})} \, dx_1 \leqslant \sqrt{2R_j} \left(\int_{\mathbb{R}} \|u_j(x_1, \cdot)\|^2_{L^2(\mathbb{R}^{n-1})} \, dx_1 \right)^{1/2} = \sqrt{2R_j} \left(\int_{\mathbb{R}^n} |u_j(x)|^2 \, dx \right)^{1/2} = \sqrt{2} \left(R_j \int_{\Omega_j} |u(x)|^2 \, dx \right)^{1/2}.$$

So finally

$$\int_{-\infty}^{+\infty} \|u(x_1, \cdot)\|_{L^2(\mathbb{R}^{n-1})} \leqslant \sqrt{2} \|u\|_{\mathfrak{B}(\mathbb{R}^n)}.$$

Lemma A.4: Let $D_j = -i\partial/\partial x_j$. For any $u \in \mathscr{H}^2(\mathbb{R}^n)$, $k \in \mathbb{C}^n$, $z \in \mathbb{C}$, and $j = 1, ..., n$,

$$\|(D_j + k_j)u\|_{\mathfrak{B}^*(\mathbb{R}^n)} \leqslant 2\sqrt{2} \|(H_0(k) - z)u\|_{\mathfrak{B}(\mathbb{R}^n)}.$$

Proof: It is sufficient to prove the estimate for $u \in \mathscr{C}_0^\infty(\mathbb{R}^n)$. Consider first the case $n = 1$. Let $D = -id/dx$, $k \in \mathbb{C}$, $z \in \mathbb{C}$, and $\lambda \in \mathbb{C}$ such that $z = -\lambda^2$. By Lemma A.2, we have

$$\begin{aligned}
\|(D + k)u\|_{\mathfrak{B}^*(\mathbb{R})} &\leqslant \tfrac{1}{2} [\| [i(D + k) + \lambda] u \|_{\mathfrak{B}^*(\mathbb{R})} + \| [i(D + k) - \lambda] u \|_{\mathfrak{B}^*(\mathbb{R})}] \\
&\leqslant \tfrac{1}{2} [\| [i(D + k) - \lambda] [i(D + k) + \lambda] u \|_{L^1(\mathbb{R})} + \| [i(D + k) + \lambda] [i(D + k) - \lambda] u \|_{L^1(\mathbb{R})}] \\
&= \| [-(D + k)^2 - \lambda^2] u \|_{L^1(\mathbb{R})}.
\end{aligned}$$

So

$$\|(D + k)u\|_{\mathfrak{B}^*(\mathbb{R})} \leqslant \| [(D + k)^2 - z] u \|_{L^1(\mathbb{R})}.$$

Let now n be arbitrary and choose $j = 1$. Write $k = (k_1, ..., k_n) \in \mathbb{C}^n$, $u(x) = u(x_1, x')$, where $x' = (x_2, ..., x_n)$ and denote by $\hat{u}(x_1, p')$ the Fourier transform of u with respect to x', with $p' = (p_2, ..., p_n)$.

The result for $n = 1$ gives us, for any fixed p',

$$\|(D_1 + k_1)\hat{u}(\cdot, p')\|_{\mathfrak{B}^*(\mathbb{R})} \leqslant \| [(D_1 + k_1)^2 + (p_2 + k_2)^2 + \cdots + (p_n + k_n)^2 - z] \hat{u}(\cdot, p') \|_{L^1(\mathbb{R})}.$$

Therefore, by (A1), we have, for any $R > 1$,

$$\frac{1}{R} \int_{|x_1| < R} |(D_1 + k_1)\hat{u}(x_1, p')|^2 \, dx_1 \leqslant 4 \left(\int_{-\infty}^{+\infty} [(D_1 + k_1)^2 + (p_2 + k_2)^2 + \cdots + (p_n + k_n)^2 - z] \hat{u}(x_1, p') \, dx_1 \right)^2.$$

Integrating now both terms of this inequality with respect to p' and applying, on one hand, Schwarz inequality to the integral of the right-hand side, and, on the other hand, Parseval's formula, we obtain

$$\frac{1}{R} \int_{|x_1| < R} |(D_1 + k_1)u(x)|^2 \, dx \leqslant 4 \left(\int_{-\infty}^{+\infty} \| [H_0(k) - z] u(x_1, \cdot) \|_{L^2(\mathbb{R}^{n-1})} dx_1 \right)^2.$$

By Lemma A.3, the right-hand side of this inequality is smaller than

$$8 \|(H_0(k) - z)u\|^2_{\mathfrak{B}(\mathbb{R}^n)}.$$

By (A1),

$$\|(D_1 + k_1)u\|^2_{\mathfrak{B}^*(\mathbb{R}^n)} \leqslant \sup_{R>1} \frac{1}{R} \int_{|x| < R} |(D_1 + k_1)u(x)|^2 \, dx \leqslant \sup_{R>1} \frac{1}{R} \int_{|x_1| < R} |(D_1 + k_1)u(x)|^2 \, dx.$$

So, finally,

$$\|(D_1 + k_1)u\|^2_{\mathfrak{B}^*(\mathbb{R}^n)} \leqslant 8 \|(H_0(k) - z)u\|^2_{\mathfrak{B}(\mathbb{R}^n)}$$

for any $u \in \mathscr{C}_0^\infty(\mathbb{R}^n)$ and in fact for any $u \in \mathscr{H}^2(\mathbb{R}^n)$ by a density argument (see Ref. 1).

Lemma A.5: Let $\delta > 0$. Then, for any $u \in \mathscr{H}^2(\mathbb{R}^n)$, $k \in \mathbb{C}^n$, $|k| \geqslant \delta$, and $\epsilon \in \mathbb{R}$,

$$|k| \, \|u\|_{\mathfrak{B}^*(\mathbb{R}^n)} \leqslant C_\delta \left(\|(H_0(k) - k^2 - i\epsilon)u\|_{\mathfrak{B}^*(\mathbb{R}^n)} + 2 \sum_{j=1}^{n} \|(D_j + k_j)u\|_{\mathfrak{B}^*(\mathbb{R}^n)} \right), \tag{A6}$$

where C_δ is a constant depending only on δ.

Proof: Let $k = a + ib \in \mathbb{C}^n$ and $\epsilon \in \mathbb{R}$.

We prove first (A6) with L^2-norms. Let

$$F(p) = |(p + k)^2 - k^2 - i\epsilon|^2 + 4 \sum_{j=1}^{n} |p_j + k_j|^2.$$

An easy estimate shows that, for $|k| \geqslant \delta$,

$$F(p) \geqslant \mathscr{C}_\delta^2 |k|^2, \quad \text{where } \mathscr{C}_\delta^2 = \min(2, \delta^2/4). \tag{A7}$$

Let now $u \in \mathscr{C}_0^\infty(\mathbb{R}^n)$. Multiplying both sides of (A7) by $|\hat{u}(p)|^2$, integrating on $\mathbb{R}^n$, and applying Parseval's formula, we get, for any $k \in \mathbb{C}^n$, $|k| \geqslant \delta$, and $\epsilon \in \mathbb{R}$

$$\mathscr{C}_\delta^2 \, |k|^2 \|u\|^2 \leqslant \|(H_0(k) - k^2 - i\epsilon)u\|^2 + 4 \sum_{j=1}^{n} \|(D_j + k_j)u\|^2$$

and then

$$\mathscr{C}_\delta |k| \, \|u\| \leqslant \|(H_0(k) - k^2 - i\epsilon)u\| + 2\sum_{j=1}^{n} \|(D_j + k_j)u\|. \tag{A8}$$

Let now φ_R^γ, $0 < \gamma \leqslant 1$, and $R > 1$, as defined in the beginning of this appendix, and apply (A8) to $\varphi_R^\gamma u \in \mathscr{C}_0^\infty(\mathbb{R}^n)$:

$$\mathscr{C}_\delta |k| \, \|\varphi_R^\gamma u\| \leqslant \|(H_0(k) - k^2 - i\epsilon)(\varphi_R^\gamma u)\| + 2\sum_{j=1}^{n} \|(D_j + k_j)(\varphi_R^\gamma u)\|. \tag{A9}$$

Remark now that, for any $\alpha = (\alpha_1, \ldots, \alpha_n)$, $|\alpha| = \alpha_1 + \cdots + \alpha_n$, $x \in \mathbb{R}^n$, and $R > 1$,

$$|D^\alpha \varphi_R^\gamma(x)| \leqslant \gamma C_\alpha(\varphi) \chi_R(x), \tag{A10}$$

where $C_\alpha(\varphi) = C_\alpha \sup_{1 \leqslant |\alpha| \leqslant 2} \|D^\alpha \varphi\|_{L^\infty(\mathbb{R}^n)}$, C_α depending only on α.

By means of (A10), we can evaluate the right-hand side of (A9):

$$\begin{aligned} &\|(H_0(k) - k^2 - i\epsilon)(\varphi_\gamma^R u)\| \\ &\quad \leqslant \|\varphi_R^\gamma (H_0(k) - k^2 - i\epsilon)u\| \\ &\qquad + 2\sum_{j=1}^{n} \|(D_j \varphi_R^\gamma)(D_j + k_j)u\| + \sum_{j=1}^{n} \|(D_j^2 \varphi_R^\gamma)u\|. \end{aligned}$$

So

$$\begin{aligned} &\|(H_0(k) - k^2 - i\epsilon)(\varphi_\gamma^R u)\| \\ &\quad \leqslant \|\varphi_R^\gamma (H_0(k) - k^2 - i\epsilon)u\| \\ &\qquad + \gamma 2C_1(\varphi) \sum_{j=1}^{n} \|\chi_R (D_j + k_j)u\| + \gamma n C_2(\varphi) \|\chi_R u\| \end{aligned} \tag{A11}$$

and

$$\|(D_j + k_j)(\varphi_R^\gamma u)\| \leqslant \|\varphi_R^\gamma (D_j + k_j)u\| + \|(D_j \varphi_R^\gamma)u\|.$$

So

$$\|(D_j + k_j)(\varphi_R^\gamma u)\| \leqslant \|\varphi_R^\gamma (D_j + k_j)u\| + \gamma C_1(\varphi) \|\chi_R u\|. \tag{A12}$$

Applying (A11) and (A12) in (A9), we have

$$\begin{aligned} &\mathscr{C}_\delta |k| \, \|\varphi_R^\gamma u\| \\ &\quad \leqslant \|\varphi_R^\gamma (H_0(k) - k^2 - i\epsilon)u\| + 2\sum_{j=1}^{n} \|\varphi_R^\gamma (D_j + k_j)u\| \\ &\qquad + \gamma \Big[2C_1(\varphi) \sum_{j=1}^{n} \|\chi_R (D_j + k_j)u\| \\ &\qquad + n(C_1(\varphi) + C_2(\varphi)) \|\chi_R u\| \Big]. \end{aligned}$$

Take the supremum over all $R > 1$. This gives, using (A2),

$$\begin{aligned} &\mathscr{C}_\delta |k| \, \|u\|_{\mathfrak{B}^*} \\ &\quad \leqslant \|(H_0(k) - k^2 - i\epsilon)u\|_{\mathfrak{B}^*}^{\varphi^\gamma} + 2\sum_{j=1}^{n} \|(D_j + k_j)u\|_{\mathfrak{B}^*}^{\varphi^\gamma} \\ &\qquad + \gamma \big[2C_1(\varphi) \|(D_j + k_j)u\|_{\mathfrak{B}^*}^{\chi^\gamma} \\ &\qquad + n(C_1(\varphi) + C_2(\varphi)) \|u\|_{\mathfrak{B}^*}^{\chi^\gamma} \big]. \end{aligned} \tag{A13}$$

Applying (A3) and (A3′) in (A13), we get

$$\begin{aligned} &\mathscr{C}_\delta |k| \, \|u\|_{\mathfrak{B}^*} \leqslant 2^{N/2+1} \\ &\quad \times \Big[\|(H_0(k) - k^2 - i\epsilon)u\|_{\mathfrak{B}^*} + 2\sum_{j=1}^{n} \|(D_j + k_j)u\|_{\mathfrak{B}^*} \Big] \\ &\quad + \gamma^{1/2} 2^{N/2} \Big[2C_1(\varphi) \sum_{j=1}^{n} \|(D_j + k_j)u\|_{\mathfrak{B}^*} \\ &\quad + n(C_1(\varphi) + C_2(\varphi)) \|u\|_{\mathfrak{B}^*} \Big]. \end{aligned} \tag{A14}$$

Since $0 < \gamma \leqslant 1$, (A14) gives, setting $C_\delta = 2^{N/2+1}(2 + C_1(\varphi))/\mathscr{C}_\delta$:

$$\begin{aligned} |k| \, \|u\|_{\mathfrak{B}^*} &\leqslant \tfrac{1}{2} C_\delta \Big[\|(H_0(k) - k^2 - i\epsilon)u\|_{\mathfrak{B}^*} \\ &\quad + 2\sum_{j=1}^{n} \|(D_j + k_j)u\|_{\mathfrak{B}^*} \Big] \\ &\quad + |k| \gamma^{1/2} \frac{2^{N/2} n (C_1(\varphi) + C_2(\varphi))}{\delta \mathscr{C}_\delta} \|u\|_{\mathfrak{B}^*}. \end{aligned} \tag{A15}$$

Finally, choosing γ such that

$$\gamma^{1/2} \, 2^{N/2} n (C_1(\varphi) + C_2(\varphi))/\delta \mathscr{C}_\delta \leqslant \tfrac{1}{2}$$

(A15) becomes

$$\begin{aligned} |k| \, \|u\|_{\mathfrak{B}^*} &\leqslant C_\delta \Big(\|(H_0(k) - k^2 - i\epsilon)u\|_{\mathfrak{B}^*} \\ &\quad + 2\sum_{j=1}^{n} \|(D_j + k_j)u\|_{\mathfrak{B}^*} \Big), \end{aligned}$$

which is the desired result for $u \in \mathscr{C}_0^\infty(\mathbb{R}^n)$. The result for $u \in \mathscr{H}^2(\mathbb{R}^n)$ follows again by a density argument.

Now, we can give:

Proof of Theorem A.1: Let $u \in \mathscr{H}^2(\mathbb{R}^n)$, $k \in \mathbb{C}^n$, $|k| \geqslant \delta$, and $\epsilon \in \mathbb{R}$. By Lemma A.5,

$$\begin{aligned} |k| \, \|u\|_{\mathfrak{B}^*} &\leqslant C_\delta \Big[\|(H_0(k) - k^2 - i\epsilon)u\|_{\mathfrak{B}^*} \\ &\quad + 2\sum_{j=1}^{n} \|(D_j + k_j)u\|_{\mathfrak{B}^*} \Big]. \end{aligned}$$

By Lemma A.4,

$$\|(D_j + k_j)u\|_{\mathfrak{B}^*} \leqslant 2\sqrt{2} \|(H_0(k) - k^2 - i\epsilon)u\|_{\mathfrak{B}}.$$

Thus, finally,

$$|k| \, \|u\|_{\mathfrak{B}^*} \leqslant C_\delta (1 + 4\sqrt{2} n) \|(H_0(k) - k^2 - i\epsilon)u\|_{\mathfrak{B}}$$

since

$$\|\cdot\|_{\mathfrak{B}^*} \leqslant \|\cdot\|_{\mathfrak{B}}.$$

APPENDIX B

We provide here the tools for the proof of (iii) in Theorem 4.1.

Lemma B.1: Let V be a short-range potential. Then, if $k_0^2 \in \sigma_p(H) \cap \mathbb{R}^+$, $k_0^2 \neq 0$, there exists a real neighborhood $\nu(k_0)$ of k_0 and a norm continuous family $\{\Gamma(k), k \in \nu(k_0)\}$ of bounded linear operators on $\mathfrak{B}(\mathbb{R}^n)$ such that

$$(1 - P(k, k_0^2))(1 + K(k, k^2))\Gamma(k) = 1 - P(k, k_0^2)$$

and, for $k \in \nu(k_0) \setminus \{k_0\}$,

$$(1 - P(k, k_0^2))R(k, k^2) = (1 - P(k, k_0^2))R_0(k, k^2)\Gamma(k). \tag{B1}$$

Proof: We refer to Refs. 1, 6, and 22 for a proof that the null space $N(1 + VR(0, k_0^2)^*)$ is $P\{k_0^2\}\mathscr{H}$. Equivalently,

$$N(1 + K(k_0, k_0^2)^*) = P(k_0, k_0^2)\mathscr{H}. \tag{B2}$$

Actually, since $P\{k_0^2\}$ is finite-dimensional and $P\{k_0^2\}L^2(\mathbb{R}^n) \subset L_s^2(\mathbb{R}^n)$ for all $s \in \mathbb{R}$, $P(k_0, k_0^2)$ is a bounded operator on $\mathfrak{B}(\mathbb{R}^n)$ and $\mathfrak{B}^*(\mathbb{R}^n)$. This also applies to $\{P(k, k_0^2), k \in \mathbb{R}\}$, which is readily seen to be norm continuous in $k \in \mathbb{R}$ on both $\mathfrak{B}(\mathbb{R}^n)$ and $\mathfrak{B}^*(\mathbb{R}^n)$. Now, by (B2), the range of $1 + K(k_0, k_0^2)$ is equal to $(1 - P(k_0, k_0^2))\mathfrak{B}(\mathbb{R}^n)$; in fact, both subspaces are closed, since $K(k_0, k_0^2)$ and $P(k_0, k_0^2)$ are compact on $\mathfrak{B}(\mathbb{R}^n)$, and they have the same annihilator in $\mathfrak{B}^*(\mathbb{R}^n)$, namely, $P(k_0, k_0^2)L^2(\mathbb{R}^n)$. By the general theory of Fredholm operators,[23] there exists a bounded linear operator $\Gamma(k_0)$ on $\mathfrak{B}(\mathbb{R}^n)$ such that, with $Q(k_0, k_0^2) = 1 - P(k, k_0^2)$, one has

$$(1 + K(k_0, k_0^2))\Gamma(k_0) = Q(k_0, k_0^2).$$

Consider now $A(k) = Q(k, k_0^2)(1 + K(k, k^2))$; then $\{A(k), k \in \mathbb{R}\setminus\{0\}\}$ is norm continuous in k on $\mathfrak{B}(\mathbb{R}^n)$ since both factors are. Since, furthermore, the kernels and cokernels of $A(k)$ and $A(k_0)$ have the same dimensions for k in a neighborhood $\nu(k_0)$ of k_0, there exists a continuous family $\{\Gamma(k), k \in \nu(k_0)\}$ of right pseudo-inverses of $A(k)$ such that

$$A(k)\Gamma(k) = Q(k, k_0^2).$$

Now, applying $R(k, k^2)$ to the left of this equality, $k \in \nu(k_0)\setminus\{k_0\}$, one gets

$$\begin{aligned} R(k, k^2)Q(k, k_0^2) &= R(k, k^2)Q(k, k_0^2)(1 + K(k, k^2))\Gamma(k) \\ &= Q(k, k_0^2)R_0(k, k^2)\Gamma(k). \end{aligned}$$

In the last equality, we have used the easily justified $R(k, k^2)Q(k, k_0^2) = Q(k, k_0^2)R(k, k^2)$ and (4.4).

Lemma B.2: $(1 - P(k, k_0^2))R(k, k^2)$ satisfies (iii) of Theorem 4.1.

Proof: This follows immediately from (B1) since $Q(k, k_0^2)$ and $\Gamma(k)$ are norm continuous on $\mathfrak{B}^*(\mathbb{R}^n)$ and $\mathfrak{B}(\mathbb{R}^n)$, respectively, for $k \in \nu(k_0)$ and $R_0(k, k^2)$ is weakly continuous in $\mathscr{L}(\mathfrak{B}, \mathfrak{B}^*)$ by Lemma 4.2(i).

APPENDIX C

We give a technical estimate, based on scaling arguments, which is needed in the proof of Theorem 7.1.

Theorem C.1: Assume that V satisfies the condition of Theorem 7.1. Then, for $\frac{1}{2} < s_1, s_2 < 1$, one has

$$\begin{aligned} &\text{(i) } \| \rho_{s_1} R_0(k, k^2)\rho_{s_2} \| \\ &\quad = O(|k|^{-2+s_1+s_2}) \quad \text{as } |k| \to \infty, \ k \in \mathbb{C}^+, \\ &\text{(ii) } \| \rho_{\alpha/2} R(k, k^2)\rho_{s_2} \| \\ &\quad = O(|k|^{-2+\alpha/2+s_2}) \quad \text{as } |k| \to \infty, \ k \in \mathbb{C}^+. \end{aligned}$$

Proof: (i) We show this result first for $R_0(k, k^2)$ and for $k \in \mathbb{R}^+$. The proof extends easily to $k \in \mathbb{C}^+$. An essential role is played by the fact that, for $0 \leqslant s < 1$, ρ_s is a bounded map from $\mathscr{H}^{+1}(\mathbb{R}^n)$ to $L^2(\mathbb{R}^n)$ and from $L^2(\mathbb{R}^n)$ to $\mathscr{H}^{-1}(\mathbb{R}^n)$ (see, e.g., Ref. 15). Accordingly, $\rho_{s_1} R_0(k, k^2 + i\epsilon)\rho_{s_2}$ makes sense as a bounded operator on $L^2(\mathbb{R}^n)$ for $\epsilon > 0$. By unitarity of the scaling transformation $D(k)$: $u(x) \mapsto k^{n/2}u(kx)$ and since $D^{-1}(k)H_0(k)D(k) = k^2H_0(1)$, one has

$$\begin{aligned} \| \rho_{s_1}(H_0(k) - k^2 - i\epsilon)^{-1}\rho_{s_2} \| &= k^{-2+s_1+s_2} \| \rho_{s_1}(H_0(1) - 1 - i\epsilon k^{-2})^{-1}\rho_{s_2} \| \\ &= k^{-2+s_1+s_2} \| \rho_{s_1}(H_0 - 1 - i\epsilon k^{-2})\rho_{s_2} \|. \end{aligned}$$

So it remains to show that the last factor on the rhs of this equality is uniformly bounded on $L^2(\mathbb{R}^n)$ as $k \to +\infty$, $\epsilon \downarrow 0^+$. Using the first iteration of the resolvent equation and the fact that $\tilde{\rho}_{s_1}(H_0 - 1 - i\epsilon k^2)^{-1}\tilde{\rho}_{s_2}$ is uniformly bounded for $s_1, s_2 > \frac{1}{2}$ by Agmon's theory,[1] it is enough to show that $\rho_{s_1}(H_0 + 1)^{-1}\rho_{s_2}$, $\rho_{s_1}(H_0 + 1)^{-2}\rho_{s_2}$, and $\rho_{s_1}(H_0 + 1)^{-1}\tilde{\rho}_{s_1}$ are bounded. For the first two terms, this follows from the previously mentioned properties of the mappings ρ_s. For the last one, it is enough to use the commutation property

$$\begin{aligned} \rho_{s_1}(H_0 + 1)^{-1}\tilde{\rho}_{s_1} &= \tilde{\rho}_{s_1}\rho_{s_1}(H_0 + 1)^{-1} \\ &\quad + \rho_{s_1}(H_0 + 1)^{-1}[H_0, \tilde{\rho}_{s_1}](H_0 + 1)^{-1}. \end{aligned}$$

It is easy to show that, since $s_1 < 1$, the commutator $[H_0, \tilde{\rho}_{s_1}^{-1}]$ is H_0-bounded, so that $[H_0, \tilde{\rho}_{s_1}](H_0 + 1)^{-1}$ is bounded on $L^2(\mathbb{R}^n)$. The other terms $\tilde{\rho}_{s_1}\rho_{s_1}(H_0 + 1)^{-1}$ and $\rho_{s_1}(H_0 + 1)^{-1}$ also are by relative boundedness arguments.

(ii) We now apply this result to $R(k, k^2)$; by the second resolvent equation

$$\begin{aligned} \rho_{\alpha/2}R(k, k^2)\rho_{s_2} &= \rho_{\alpha/2}R_0(k, k^2)\rho_{s_2} \\ &\quad - \rho_{\alpha/2}R_0(k, k^2)U\rho_{\alpha/2}R(k, k^2)\rho_{s_2}, \end{aligned}$$

where $U = v\rho_{\alpha/2}$. By (i), $\| \rho_{\alpha/2}R_0(k, k^2)U \| = O(k^{-2+\alpha})$ as $k \to \infty$, so that, for k large enough, $1 + \rho_{\alpha/2}R_0(k, k^2)U$ is invertible and

$$\begin{aligned} \rho_{\alpha/2}R(k, k^2)\rho_{s_2} &= (1 + \rho_{\alpha/2}R_0(k, k^2)U)^{-1}\rho_{\alpha/2}R_0(k, k^2)\rho_{s_2}, \end{aligned}$$

which concludes the proof by another application of (i).

[1] S. Agmon, Ann. Scuola Norm. Sup. Pisa, Ser. IV **2**, 151 (1975).
[2] S. Agmon and L. Hörmander, J. Anal. Math. **30**, 1 (1976).
[3] N. Khuri, Phys. Rev. **107**, 1148 (1957).
[4] B. Simon, *Quantum Mechanics for Hamiltonians Defined as Quadratic Forms*, Princeton Series in Physics (Princeton U. P., Princeton, NJ, 1971).
[5] J. Taylor, *Scattering Theory* (Wiley, New York, 1972).
[6] M. Reed and B. Simon, *Methods of Modern Mathematical Physics. Vol. III. Scattering Theory* (Academic, New York, 1979).
[7] W. Amrein and D. Pearson, J. Phys. A **12**, 1469 (1979).
[8] A. Martin, Comm. Math. Phys. **69**, 89 (1979); **73**, 79 (1980); Acta Phys. Austr., lectures given at the XX "Universitätwochen für Kernphysik," Schladming (1981).
[9] V. Enss and B. Simon, Phys. Rev. Lett. **44**, 319, 764 (1980); Comm. Math. Phys. **76**, 177 (1980).
[10] Y. Saito, preprint, Heidelberg University, 1980.
[11] A. Jensen, Helv. Phys. Acta **53**, 398 (1980).
[12] S. Agmon, in "Séminaire Goulaouic-Schwartz" Lecture 2 (1978–1979) (unpublished).
[13] (a) A. Jensen and T. Kato, Duke Math. J. **46**, 583 (1979); (b) A. Jensen,

Duke Math. J. **47**, 57 (1980).
[14](a) J. M. Combes and L. Thomas, Commun. Math. **34**, 251 (1973); (b) J. M. Combes, in *Scattering Theory in Mathematical Physics*, edited by J. A. La Vita and J. P. Marchand (Reidel, Dordrecht, 1974).
[15]M. Reed and B. Simon, *Methods of Modern Mathematical Physics. Vol. II. Fourier Analysis—Self-Adjointness* (Academic, New York, 1975).
[16]M. Reed and B. Simon, *Methods of Modern Mathematical Physics. Vol. IV. Analysis of Operators* (Academic, New York, 1978).
[17](a) R. Froese, I. Herbst, M. Hoffmann-Ostenhof, and T. Hoffmann-Ostenhof, J. Anal. Math. **41**, 272 (1982); (b) R. Froese and I. Herbst, Comm. Math. Phys. **87**, 429 (1982).
[18]T. Kato, *Perturbation Theory for Linear Operators* (Springer-Verlag, Berlin, 1976), 2nd ed.
[19]J. Aguilar and J. M. Combes, Comm. Math. Phys. **22**, 269 (1971).
[20]W. P. Reinhardt, Ann. Rev. Phys. Chem. **33**, 223 (1982).
[21]J. M. Combes and A. Tip, Ann. Inst. Henri Poincaré **40**, 117 (1984).
[22]J. M. Combes, in *Rigorous Atomic and Molecular Physics*, edited by G. Velo and A. Wightman (Plenum, New York, 1981).
[23]M. Schechter, *Principles of Functional Analysis* (Academic, New York, 1971).

Finite rank approximation of wave operators for short-range and Coulomb interactions

Helmut Kröger
Département de Physique, Université Laval, Québec, P.Q., Canada G1K 7P4

(Received 16 November 1983; accepted for publication 3 February 1984)

For the two-body system wave operators are shown to be strongly approximatable by finite rank operators which are obtained from finite rank approximations of the full and asymptotic Hamiltonian in the sense of strong resolvent convergence. This is demonstrated for short-range potentials plus the Coulomb potential with expansion functions chosen in momentum space as step functions. A generalization to the N-body system is indicated.

PACS numbers: 03.65.Nk, 02.60. + y

I. INTRODUCTION

In a previous paper[1] a new approach has been suggested by the author for the calculation of quantum mechanical S-matrix elements between wave packets. It is based on strong approximations of the wave operators by exponentials of bounded or even finite-rank operators. The approximation using finite-rank operators is practically very convenient because the calculation of the approximate wave operator essentially reduces to the diagonalization of two finite, real, symmetric matrices (corresponding to the full and asymptotic Hamiltonian). In Ref. 1 the finite-rank approximation has been derived as a second step from an approximation of the wave operators by exponentials of bounded (but not finite-rank) operators. In the first step the approximation of wave operators by exponentials of bounded operators is achieved by some "deformation" of the unbounded Hamiltonians. For example, the kinetic energy Hamiltonian H^0 is substituted by a bounded Hamiltonian H^0_S

$$H^0 \to H^0_S = E/S \arctan(SH^0/E), \tag{1.1}$$

where $E > 0$ is a scaling factor of energy dimension and $S > 0$ is an approximation parameter. If, in the simplest case, the potential V is a bounded operator (which is not valid for the Coulomb potential), one can substitute the full Hamiltonian H by a bounded Hamiltonian H_S via

$$H \to H_S = H^0_S + V. \tag{1.2}$$

If the Moller wave operator $\Omega^{(\pm)}$ is substituted by $\Omega_{S,T}$ via

$$\begin{aligned}\Omega^{(\pm)} &= s-\lim_{t\to\mp\infty} \exp(iHt)\exp(-iH^0t)\\ &\to \Omega_{S,T} = \exp(iH_S T)\exp(-iH^0_S T),\end{aligned} \tag{1.3}$$

it can be shown[1] that $\Omega_{S,T}$ tends strongly to $\Omega^{(\pm)}$ for suitably chosen $S\to 0, T\to\mp\infty$.

This approach can be generalized to include a large class of short-range unbounded potentials, like, e.g., the Yukawa potential, if the wave operator exists and the potential can be "deformed" to a bounded one. Also the Coulomb potential can be included which requires "deformations" of H^0, of Dollard's anomalous term in the wave operator[2] and of the Coulomb potential.

In a second step an approximation of the wave operators by exponentials of finite-rank operators was introduced,[1] which, in the simplest case of a bounded short-range potential, substitutes

$$H^0_S \to H^0_{S,N} = P_N H^0_S P_N, \tag{1.4a}$$

$$H_S \to H_{S,N} = P_N H_S P_N, \tag{1.4b}$$

where P_N is an orthogonal projector on the first N elements of a complete orthogonal function system. Substituting

$$\Omega_{S,T} \to \Omega_{S,N,T} = \exp(iH_{S,N}T)\exp(-iH^0_{S,N}T), \tag{1.5}$$

one can show[1] that $\Omega_{S,N,T}$ tends strongly to $\Omega^{(\pm)}$ for suitably chosen $S\to 0$, $N\to\infty$, $T\to\mp\infty$.

For practical calculations the approximation by exponentials of finite-rank operators seems very convenient because its evaluation basically means the diagonalization of the finite, real, symmetric matrices corresponding to $H_{S,N}$, $H^0_{S,N}$.

This approach has been tested in practical calculations of S-matrix elements in the two-body system with a short-range potential, with a Coulomb potential and with a sum of both, and was found to give results converging to the reference solution.[3] Also it has been applied in the three-body system for $p + d$ scattering.[4,5] The first results demonstrate the practical applicability of the method.

In practical calculations it has turned out, however, that the approximation by exponentials of finite-rank operators can be obtained directly from projections, i.e., by avoiding the "deformations" which were intended for the boundedness of the approximate Hamiltonians. For a bounded short-range potential that means

$$H^0 \to H^0_N = P_N H^0 P_N, \tag{1.6a}$$

$$H \to H_N = P_N H P_N, \tag{1.6b}$$

$$\Omega^{(\pm)} \to \Omega_{N,T} = \exp(iH_N T)\exp(-iH^0_N T). \tag{1.6c}$$

The validity of the direct approximation of wave operators by exponentials of finite-rank operators, Eq. (1.6), compared with the indirect one, Eqs. (1.4), (1.5) would improve and strengthen the method because one source of numerical error could be avoided. It is the aim of this paper to demonstrate that the direct approximation is actually valid. In particular we will prove it for bounded or H^0-relatively bounded short-range potentials and also the Coulomb potential, using as expansion functions the step functions in momentum space.

 0022-2488/84/061875-06$02.50

In Sec. II, we discuss the two-body case. In Sec. III a generalization to the N-body case is indicated.

II. TWO-BODY SYSTEM

In the two-body system we assume that the center of mass motion can be split off and consider only the relative motion. For simplicity we drop spin dependence, which can be included also. (The notation is taken from Ref. 1.) I quote a result given by Weidmann,[6] which will be used in the following.

Definition of strong resolvent convergence (src). Let $T_n\,(n\in\mathbb{N})$, T be self-adjoint operators with domains $D(T_n)$, $D(T)$, respectively, in a Hilbert space $\mathscr{H}$. T_n is said to converge to T in the sense of strong resolvent convergence if $(z-T_n)^{-1}\xrightarrow{s}(z-T)^{-1}$ for some $z\in\mathbb{C}\setminus\mathbb{R}$.

Theorem on src: Let $T_n\,(n\in\mathbb{N})$, T be self-adjoint operators with domains in a Hilbert space $\mathscr{H}$. Assume there is a core C of T, i.e., a subset of $D(T)$ with the property $\overline{T_{|C}}=T$. For every $f\in C$, let $f\in D(T_n)$ for n larger than some n_0 and $T_n f\rightarrow Tf$. Then T_n converges to T in the sense of src. Moreover $u(T_n)\xrightarrow{s}u(T)$ for every continuous bounded function u defined on $\mathbb{R}$, in particular $\exp(iT_n)\xrightarrow{s}\exp(iT)$.

In the following we want to use this theorem. Thus we have to find cores C,C^{as} for the Hamiltonian H and the asymptotic Hamiltonian H^{as}, respectively, and finite-rank Hamiltonians H_n,H_n^{as} such that for $f\in C$ $H_n f\rightarrow Hf$, for $g\in C^{\text{as}}$ $H_n^{\text{as}}g\rightarrow H^{\text{as}}g$.

In the following we use the momentum space, which means in this section the relative momentum between the two particles. We denote by $\mathscr{H}=\mathscr{L}_2(\mathbb{R}^3)$ the usual Hilbert space and $\mathscr{H}_0$ by

$$\mathscr{H}_0=\left\{h(q)\,\middle|\,\int_0^\infty dq\,q^2|h(q)|^2\text{ exists}\right\}\tag{2.1}$$

with the usual scalar product.

We want to deal with the expansion functions above referred to as step functions.

Definition 1: For each $N\in\mathbb{N}$ let $\{q_i^{(N)}|i=1,\ldots,N+1\}$ denote a monotonous discretization of the interval $[0,q_{\text{cut}}^{(N)}]$, with $q_1^{(N)}=0$, $q_{N+1}^{(N)}=q_{\text{cut}}^{(N)}$, such that with increasing N

$$q_{\text{cut}}^{(N)}\rightarrow\infty,\quad \Delta^{(N)}=\max_{i\in1,\ldots,N}|q_{i+1}^{(N)}-q_i^{(N)}|\rightarrow0\,.$$

Let

$$h_i^{(N)}(q)=\alpha_i^{(N)}\Theta(q-q_i^{(N)})\Theta(q_{i+1}^{(N)}-q)\,,$$

where $\alpha_i^{(N)}$ normalizes $h_i^{(N)}$ to 1, i.e.,

$$\alpha_i^{(N)}=(\tfrac{1}{3}\cdot(q_{i+1}^{(N)\,3}-q_i^{(N)3}))^{-1/2}\,.$$

The functions $h_i^{(N)}$, $i=1,\ldots,N$; $N=1,2,\ldots$ form a complete basis of $\mathscr{H}_0$, with

$$\langle h_i^{(N)}|h_j^{(N)}\rangle=\delta_{ij}\,.$$

Hence,

$$P_N=\sum_{i=1}^N|h_i^{(N)}\rangle\langle h_i^{(N)}|$$

is an orthogonal projector on $\mathscr{H}_0$ with

$$P_N\xrightarrow{s}1\,.$$

From this set of expansion functions on $\mathscr{H}_0$, one can construct a set of expansion functions on $\mathscr{H}$, e.g., by using the spherical harmonics Y_{lm}

$$h_{i,l,m}^{(N)}(\hat{q})=h_i^{(N)}(q)\,Y_{l,m}(\hat{q})\,.$$

Hence,

$$P_N=\sum_{i=1}^N\sum_{l=0}^N\sum_{m=-l}^l\left|h_{i,l,m}^{(N)}\rangle\langle h_{i,l,m}^{(N)}\right|$$

is an orthogonal projector on $\mathscr{H}$ with

$$P_N\xrightarrow{s}1\,.$$

Now we claim

Proposition 1: Let $D=D(H^0)=\{\phi\,|\phi\in\mathscr{H},H^0\phi\in\mathscr{H}\}$. For each $\psi\in D$ and using the projectors given in Definition 1 one has

$$\text{(i)}\quad H^0P_N\psi\rightarrow H^0\psi\,,\tag{2.2a}$$

$$\text{(ii)}\quad P_NH^0P_N\psi\rightarrow H^0\psi\,.\tag{2.2b}$$

Proof: The proof would be trivial, if H^0 would be bounded, which is not the case. But H^0 is rotationally symmetric and acts in the space of the angular variables $\hat{q}$ like a constant, hence it is sufficient to consider $H^0|_{\mathscr{H}_0}$, with $D_0=\{\phi\,|\phi\in\mathscr{H}_0,\,H^0\phi\in\mathscr{H}_0\}$. One has for $\psi\in D_0$

$$\|H^0P_N\psi\|^2=\sum_{i=1}^N\langle\psi|h_i^{(N)}\rangle\langle h_i^{(N)}|H^{02}|h_i^{(N)}\rangle\langle h_i^{(N)}|\psi\rangle\,,\tag{2.3}$$

where the H^{02} matrix elements vanish off the diagonal which is a property of the step functions. One can estimate (omitting the upper index N)

$$|\langle\psi|h_i\rangle|^2=\left|\int_{q_i}^{q_{i+1}}dq\,q^2\alpha_i\psi(q)\right|^2$$
$$\leqslant\alpha_i^2\bar{\psi}_i^2\left(\int_{q_i}^{q_{i+1}}dq\,q^2\right)^2=\bar{\psi}_i^2/\alpha_i^2\,,\tag{2.4a}$$

where $\bar{\psi}_i^2=\sup_{q\in[q_i,q_{i+1}]}|\psi(q)|^2$ and

$$\langle h_i|H^{02}|h_i\rangle=\int_{q_i}^{q_{i+1}}dq\,q^2\,\alpha_i^2\left(\frac{q^2}{2m}\right)^2$$
$$\leqslant\alpha_i^2\,\overline{H_i^{02}}\int_{q_i}^{q_{i+1}}dq\,q^2=\overline{H_i^{02}}\,,\tag{2.4b}$$

where

$$\overline{H_i^{02}}=\sup_{q\in[q_i,q_{i+1}]}\left(\frac{q^2}{2m}\right)^2=\left(\frac{q_{i+1}^2}{2m}\right)^2.$$

Hence,

$$\|H^0P_N\psi\|^2\leqslant\sum_{i=1}^N\bar{\psi}_i^2\alpha_i^{-2}\,\overline{H_i^{02}}$$
$$=\sum_{i=1}^N\int_{q_i}^{q_{i+1}}dq\,q^2\,\bar{\psi}_i^2\,\overline{H_i^{02}}\,.\tag{2.5}$$

The expression on the rhs is an upper sum of

$$\int_0^{q_{\text{cut}}}dq\,q^2\,|\psi(q)|^2\left(\frac{q^2}{2m}\right)^2.$$

Because of $\psi \in D_0$ that integral exists in the limit $q_{\text{cut}} \to \infty$, yielding $\|H^0\psi\|^2$. The upper sum

$$\sum_{i=1}^{N} \int_{q_i}^{q_{i+1}} dq\, q^2\, \overline{\psi}_i^{\,2}\, \overline{H_i^0}^2$$

tends with $N \to \infty$ to the same limit, because by Definition 1 $q_{\text{cut}}^{(N)} \to \infty$, $\Delta^{(N)} \to 0$, hence $\|H^0 P_N \psi\|$, $N = 1,2...$ has no accumulation point larger than $\|H^0\psi\|$ and there is an upper bound $M > 0$,

$$\|H^0 P_N \psi\| < M, \quad N = 1,2,\ldots. \tag{2.6}$$

Because $P_N \xrightarrow{s} 1$, one has for every ψ', $\psi \in D_0$

$$\langle \psi' | H^0 P_N | \psi \rangle = \langle H^0 \psi' | P_N \psi \rangle$$
$$\to \langle H^0 \psi' | \psi \rangle = \langle \psi' | H^0 | \psi \rangle. \tag{2.7}$$

Because $D(H^0)$ is dense in $\mathscr{H}$ (see, e.g., Ref. 7), D_0 is dense in $\mathscr{H}_0$, and $\|H^0 P_N \psi\|$ is bounded [Eq. (2.6)], one has for every $\phi \in \mathscr{H}_0$, $\psi \in D_0$

$$\langle \phi | H^0 P_N | \psi \rangle \to \langle \phi | H^0 | \psi \rangle, \tag{2.8a}$$

i.e.,

$$H^0 P_N \psi \xrightarrow{w} H^0 \psi. \tag{2.8b}$$

Now we claim for $\psi \in D_0$

$$\|H^0 P_N \psi\| \to \|H^0 \psi\| \tag{2.9}$$

$\|H_0 P_N \psi\|$, $N = 1,2,...$ has no accumulation point larger than $\|H^0\psi\|$. If a subsequence $\|H^0 P_{N'} \psi\|$, $N' \in \mathrm{N}$ would have a smaller accumulation point

$$\|H^0 P_{N'} \psi\| \to \|H^0 \psi\| - \delta, \quad \delta > 0, \tag{2.10}$$

then

$$|\langle H^0 \psi | H^0 P_{N'} \psi \rangle| \leqslant \|H^0 \psi\|\, \|H^0 P_{N'} \psi\|$$
$$\to \|H^0 \psi\|^2 - \delta \|H^0 \psi\|, \tag{2.11}$$

but (2.8b) implies

$$\langle H^0 \psi | H^0 P_{N'} \psi \rangle \to \|H^0 \psi\|^2, \tag{2.12}$$

which contradicts (2.11) and thus establishes (2.9).

Now Eq. (2.8b) and (2.9) together imply for $\psi \in D_0$

$$H^0 P_N \psi \to H^0 \psi. \tag{2.13}$$

Because P_N is an orthogonal projection, $\|P_N\| = 1$, and because of (2.13), one can estimate for $\psi \in D_0$

$$\|P_N (H^0 P_N \psi - H^0 \psi)\| \leqslant \|P_N\|\, \|H^0 P_N \psi - H^0 \psi\| \to 0. \tag{2.14}$$

Because $P_N \xrightarrow{s} 1$, $\psi \in D_0$, $H^0 \psi \in \mathscr{H}_0$ one has

$$P_N H^0 \psi \to H^0 \psi. \tag{2.15}$$

Equations (2.14), (2.15) imply

$$P_N H^0 P_N \psi \to H^0 \psi. \tag{2.16}$$

The generalization of Eqs. (2.13), (2.16) on $\mathscr{H}_0$ to Eq. (2.2) on $\mathscr{H}$ is straightforward due to the rotational symmetry of H^0.

If we want to also treat the case where the Coulomb interaction is involved, Dollard's anomalous term[2] in the asymptotic Hamiltonian has to be approximated. It can be written as

$$A^c(t) = \operatorname{sign}(t) \left(\frac{m}{2H^0} \right)^{1/2} e_1 e_2 \log(4|t| H^0) \tag{2.17}$$

and the time-dependent asymptotic Hamiltonian as

$$H^{0c}(t) = H^0 \cdot t + A^c(t). \tag{2.18}$$

Proposition 2: Let $D = D(H^{0c}(t))$ $= \{\phi \,|\, \phi \in \mathscr{H}, H^{0c}(t) \phi \in \mathscr{H}\}$. For each $\psi \in D$ and using the projections from Definition 1 one has

$$\text{(i)} \quad H^{0c}(t) P_N \psi \to H^{0c}(t) \psi, \tag{2.19a}$$

$$\text{(ii)} \quad P_N H^{0c}(t) P_N \psi \to H^{0c}(t) \psi. \tag{2.19b}$$

Proof: Also $H^{0c}(t)$ is an unbounded but rotationally symmetric operator and one can proceed similarly as in Proposition 1. We consider $H^{0c}(t)|_{\mathscr{H}_0}$ with $D_0 = \{\phi \,|\, \phi \in \mathscr{H}_0,\ H^{0c}(t) \phi \in \mathscr{H}_0\}$. D is dense in $\mathscr{H}$ and D_0 is dense in $\mathscr{H}_0$. For $\psi \in D_0$ one has

$$\|H^{0c}(t) P_N \psi\|^2 = \sum_{i=1}^{N} \langle \psi | h_i^{(N)} \rangle \langle h_i^{(N)} | H^{0c}(t)^2 | h_i^{(N)} \rangle \langle h_i^{(N)} | \psi \rangle \tag{2.20}$$

and

$$\langle h_i | H^{0c}(t)^2 | h_i \rangle = \int_{q_i}^{q_{i+1}} dq\, q^2 \alpha_i^2 \Big[\frac{q^2}{2m} \cdot t + \operatorname{sign}(t) \frac{m}{q} e_1 e_2 \log\left(\frac{2q^2}{m} |t| \right) \Big]^2$$
$$\leqslant \alpha_i^2\, \overline{H^{0c}(t)_i}^2 \int_{q_i}^{q_{i+1}} dq\, q^2 = \overline{H^{0c}(t)_i}^2, \tag{2.21}$$

where

$$\overline{H^{0c}(t)_i}^2 = \sup_{q \in [q_i, q_{i+1}]} \left[\frac{q^2}{2m} \cdot t + \operatorname{sign}(t) \frac{m}{q} e_1 e_2 \log\left(\frac{2q^2}{m} |t| \right) \right]^2.$$

Hence

$$\|H^{0c}(t) P_N \psi\|^2 \leqslant \sum_{i=1}^{N} \overline{\psi}_i^{\,2} \alpha_i^{-2}\, \overline{H^{0c}(t)_i}^2$$
$$= \sum_{i=1}^{N} \int_{q_i}^{q_{i+1}} dq\, q^2\, \overline{\psi}_i^{\,2}\, \overline{H^{0c}(t)_i}^2. \tag{2.22}$$

The rhs is an upper sum of

$$\int_0^{q_{\text{cut}}} dq\, q^2 |\psi(q)|^2 \left[\frac{q^2}{2m} \cdot t + \operatorname{sign}(t) \frac{m}{q} e_1 e_2 \log\left(\frac{2q^2}{m} |t| \right) \right]^2.$$

Because $\psi \in D_0$ this integral exists in the limit $q_{\text{cut}} \to \infty$ yielding $\|H^{0c}(t)\psi\|^2$. The upper sum

$$\sum_{i=1}^{N} \int_{q_i}^{q_{i+1}} dq\, q^2\, \overline{\psi}_i^{\,2}\, \overline{H^{0c}(t)_i}^2$$

tends with $N \to \infty$ to the same limit, because of $q_{\text{cut}}^{(N)} \to \infty$, $\Delta^{(N)} \to 0$, hence $\|H^{0c}(t) P_N \psi\|$, $N = 1,2,...$ has no accumulation point larger than $\|H^{0c}(t)\psi\|$.

The rest of the proof is quite analogous to the steps in the proof of Proposition 1.

Now we have to establish the same kind of result for the full Hamiltonian H as was established in Propositions 1 and 2 for the asymptotic Hamiltonian.

Proposition 3: Let P_N denote the orthogonal projectors from Definition 1.

(i) If V is bounded one has

$$P_N V P_N \xrightarrow{s} V. \tag{2.23a}$$

(ii) If V is relatively bounded with respect to H^0 one has for $\psi \in D(H^0)$

$$P_N V P_N \psi \to V\psi. \tag{2.23b}$$

Proof: (i) is trivial.

(ii) is called relatively bounded with respect to H^0,[7] if $D(H^0) \subseteq D(V)$ and there are $\alpha > 0$, $\beta > 0$ such that for each $\psi \in D(H^0)$

$$\|V\psi\| \leqslant \alpha \|\psi\| + \beta \|H^0\psi\|. \tag{2.24}$$

Because range $(P_N) \subseteq D(H^0)$, $N = 1,2,\ldots$, one implies for $\psi \in D(H^0)$

$$\|V(P_N - 1)\psi\| \leqslant \alpha \|(P_N - 1)\psi\| + \beta \|H^0(P_N - 1)\psi\|. \tag{2.25}$$

The rhs tends to 0 with $N \to \infty$ because $P_N \xrightarrow{s} 1$ and because of Proposition 1.

Moreover

$$P_N V\psi \to V\psi. \tag{2.26}$$

Thus Eqs. (2.25), (2.26) and $\|P_N\| = 1$ imply (2.23b).

Now we are prepared to establish an improved result of Theorem 2 and Theorem 4 of Ref. 1.

Theorem: Assume the potential V is such that it yields a self-adjoint Hamiltonian $H = H_0 + V$ with $D(H) = D(H^0)$, moreover let V be bounded or be relatively bounded with respect to H^0. Finally we assume that V guarantees the existence of the Moller wave operators $\Omega^{(\pm)}$ in the short-range case [given by Eq. (1.3)] or the Dollard[2] modified wave operators $\Omega^{c(\pm)}$ given by

$$\Omega^{c(\pm)} = s - \lim_{t \to \mp \infty} \exp(iHt)\exp(-iH^{0c}(t)) \tag{2.27}$$

in the case where long-range Coulomb forces are involved. Let P_N, $N = 1,2,\ldots$ denote the orthogonal projections from Definition 1 and

$$\begin{aligned} H_N^0 &= P_N H^0 P_N, \\ H_N^{0c}(t) &= P_N H^{0c}(t) P_N, \\ H_N &= P_N H P_N. \end{aligned} \tag{2.28}$$

Then for each $\psi \in \mathscr{H}$ one can find $N \in \mathbb{N}$, $T > 0$ such that in the short-range case

$$\|\Omega^{(+)}\psi - \exp(-iH_N T)\exp(iH_N^0 T)\psi\|$$

and

$$\|\Omega^{(-)}\psi - \exp(iH_N T)\exp(-iH_N^0 T)\psi\|$$

become arbitrarily small. In the case where the Coulomb potential is involved

$$\|\Omega^{c(+)}\psi - \exp(-iH_N T)\exp(iH_N^{0c}(T))\psi\|$$

and

$$\|\Omega^{c(-)}\psi - \exp(iH_N T)\exp(-iH_N^{0c}(T))\psi\|$$

become arbitrarily small. The corresponding approximated S-matrices

$$(\exp(iH_N T)\exp(-iH_N^0 T))^\dagger \exp(-iH_N T)\exp(iH_N^0 T),$$

$$(\exp(iH_N T)\exp(-iH_N^{0c}(T)))^\dagger \exp(-iH_N T)\exp(iH_N^{0c}(T))$$

are unitary.

Proof: The proof relies on the theorem of src. Let us consider firstly the short-range case. Then the asymptotic Hamiltonian is H^0. H^0 is self-adjoint, the domain $D(H^0)$ being dense in $\mathscr{H}$ (see, e.g., Ref. 7). Hence H^0 is closed and $D(H^0)$ is a core of H^0. The expansion functions are such that range $(P_N) \subseteq D(H^0)$. Proposition 1 and the src theorem together imply for each $t \in \mathbb{R}$

$$\exp(iH_N^0 t) \xrightarrow{s} \exp(iH^0 t). \tag{2.29}$$

For the full Hamiltonian one has by assumption that H is self-adjoint with the domain $D(H) = D(H^0)$. Hence H is closed and $D(H^0)$ is a core of H. Proposition 1, Proposition 3 and the src theorem imply for each $t \in \mathbb{R}$

$$\exp(iH_N t) \xrightarrow{s} \exp(iHt). \tag{2.30}$$

Then the existence of the wave operators (strong limit) and Eqs. (2.29), (2.30) imply the claim for the short-range case. Let us consider the case, where the long-range Coulomb force is involved. The (time-dependent) asymptotic Hamiltonian is $H^{0c}(t)$. $H^{0c}(t)$ is self-adjoint (which is a special case of a more general result from Ref. 7) and $D(H^{0c}(t))$ is dense in $\mathscr{H}$. $H^{0c}(t)$ is closed, $D(H^{0c}(t))$ is a core of $H^{0c}(t)$ and the expansion functions obey range $(P_N) \subseteq D(H^{0c}(t))$. Proposition 2 together with the src theorem implies for $t \neq 0$.

$$\exp(iH_N^{0c}(t)) \xrightarrow{s} \exp(iH^{0c}(t)).$$

For the full Hamiltonian H, which now includes a Coulomb potential, one has to verify that the assumptions made above on the potential give no contradiction. The Coulomb potential is relatively bounded with respect to H^0 and $H^0 + V^{\text{coul}}$ is a self-adjoint operator with $D(H) = D(H^0)$. This is shown, e.g., in Ref. 7. The rest of the proof in the long-range case is analogous to that of the short-range case. The finite-rank Hamiltonians $H_N^0, H_N^{0c}(t), H_N$ are self-adjoint and hence the approximate S-matrices are unitary.

We want to remark that some general conditions can be given[2,6,7] such that the potential meets the above assumptions. For example, this holds for the square-well potential, the Yukawa potential, separable potentials if the form factors are square integrable, the Coulomb potential and combinations of those.

III. *N*-BODY SYSTEM

In this section a generalization of the result for the two-body system will be sketched without giving proofs (for notation see Ref. 1). Let us firstly consider only short-range potentials. Let V^{ij} denote the interaction between the elementary particles i, j and assume a total interaction of the form

$$V = \sum_{1 \leqslant i < j \leqslant N} V^{ij}. \tag{3.1}$$

Corresponding to a cluster channel α, which denotes which of the N elementary particles form a composite particle, there is a channel interaction

$$V^{\alpha}=\sum_{\substack{1<i<j<N\\ ij\subset\alpha}} V^{ij}, \tag{3.2}$$

where $ij\subset\alpha$ indicates that the elementary particles i and j are both contained in the same composite particle of the cluster α.

Let H^0 denote the free Hamiltonian, $H^{\alpha}=H^0+V^{\alpha}$ a channel Hamiltonian, $H=H^0+V$ the full Hamiltonian and

$$\Omega^{\alpha(\pm)}=s-\lim_{t\to\mp\infty}\exp(iHt)\exp(-iH^{\alpha}t) \tag{3.3}$$

the wave operators defined on $\mathscr{H}^{\alpha}$, a subspace of the Hilbert space $\mathscr{H}$, which is a projection on all the boundstate wave function for any composite particles in channel α.

As a generalization of the set of expansion functions one could take the set of expansion functions from above for each single particle and form the tensor products. Alternatively, using a set of Jacobi coordinates in momentum space, one could take the set of expansion functions from above for each Jacobi variable and also form the tensor products. Using any of those generalized expansion functions, orthogonal projections P_N can be defined with

$$\begin{aligned}&P_N\xrightarrow{s}1,\quad H^0P_N\psi\to H^0\psi,\\&P_NH^0P_N\psi\to H^0\psi,\quad \psi\in D(H^0)\end{aligned} \tag{3.4}$$

in analogy to Proposition 1. Now we assume that each pair interaction V^{ij} is a bounded or relatively bounded operator with respect to $H^{0\,ij}$, which describes the free relative motion of the particles i,j, i.e., there are $\alpha^{ij}>0$, $\beta^{ij}>0$ such that for each $\psi\in D(H^{0\,ij})$

$$\|V^{ij}\psi\|\leqslant\alpha^{ij}\|\psi\|+\beta^{ij}\|H^{0\,ij}\psi\|. \tag{3.5}$$

For each i,j there are Jacobi coordinates describing the relative motion complementary to the relative motion of particles i,j (plus the center of mass motion) and there is a corresponding free Hamiltonian $H^{0\,\mathrm{comp}\,ij}$ with $H^0=H^{0\,ij}+H^{0\,\mathrm{comp}\,ij}$. $H^{0\,ij}$,$H^{0\,\mathrm{comp}\,ij}$ commute and are nonnegative operators, hence Eq. (3.5) yields for $\psi\in D(H^0)$

$$\|V^{ij}\psi\|\leqslant\alpha^{ij}\|\psi\|+\beta^{ij}\|H^0\psi\|. \tag{3.6}$$

Thus Eqs. (3.4), (3.6), imply for $\psi\in D(H^0)$

$$P_NV^{ij}P_N\psi\to V^{ij}\psi \tag{3.7}$$

in analogy to Proposition 3 and hence

$$P_NHP_N\psi\to H\psi. \tag{3.8}$$

In order to find finite-rank approximations for the asymptotic channel Hamiltonian H^{α} a set of expansion functions more suitable for channel α can be chosen. Thus corresponding to a channel α let $I=1,2,\ldots,I_{\max}$ count the distinct elementary or composite particles, now called generalized particles, in that channel. The expansion functions can be chosen as tensor products of the single generalized particle expansion functions or as tensor products of expansion functions corresponding to a set of Jacobi coordinates of the generalized particles. Let us denote the corresponding orthogonal projections P_N^{α}, with

$$\begin{aligned}&P_N^{\alpha}\psi\to\psi,\quad \psi\in\mathscr{H}^{\alpha},\quad H^{\alpha}P_N^{\alpha}\psi'\to H^{\alpha}\psi',\\&P_N^{\alpha}H^{\alpha}P_N^{\alpha}\psi'\to H^{\alpha}\psi',\quad \psi'\in D(H^{\alpha})\cap\mathscr{H}^{\alpha}.\end{aligned} \tag{3.9}$$

That leads us to the following.

Conjecture: Let V^{ij},V^{α},V be defined as above, such that H,H^{α} are self-adjoint Hamiltonians with $D(H)=D(H^{\alpha})=D(H^0)$, let V^{ij} be bounded or relatively bounded with respect to $H^{0\,ij}$, and assume the existence of the Moller wave operators $\Omega^{\alpha(\pm)}$. Let P_N denote the above projectors and $H_N=P_NHP_N$, $H_N^{\alpha}=P_N^{\alpha}H^{\alpha}P_N^{\alpha}$. Then for each channel α, $\psi\in\mathscr{H}^{\alpha}$,

$$\|\Omega^{\alpha(+)}\psi-\exp(-iH_NT)\exp(iH_N^{\alpha}T)\psi\|$$

and

$$\|\Omega^{\alpha(-)}\psi-\exp(iH_NT)\exp(-iH_N^{\alpha}T)\psi\|$$

can be made arbitrarily small for suitable $N\in\mathbb{N}$, $T>0$.

Finally we want to consider some of the elementary particles as charged and take into account the Coulomb potential. For a channel α let e^I denote the charge of the generalized particle I, m^I its mass, m^{IJ} the reduced mass of the generalized particles I,J and H^{0IJ} the Hamiltonian of the free relative motion of the generalized particles I,J. Let

$$\begin{aligned}&A^{IJc}(t)=\operatorname{sign}(t)e^Ie^J\left(\frac{m^{IJ}}{2H^{0IJ}}\right)^{1/2}\log(4|t|H^{0IJ}),\\&A^{\alpha c}(t)=\sum_{\substack{1<I<J<I_{\max}\\ IJ\subset\alpha}}A^{IJc}(t).\end{aligned} \tag{3.10}$$

$A^{\alpha c}(t)$ denotes Dollard's anomalous term of the asymptotic channel Hamiltonian

$$H^{\alpha c}(t)=H^{\alpha}\cdot t+A^{\alpha c}(t) \tag{3.11}$$

and

$$\Omega^{\alpha c(\pm)}=s-\lim_{t\to\mp\infty}\exp(iHt)\exp(-iH^{\alpha c}(t)) \tag{3.12}$$

is Dollard's modified wave operator defined on $\mathscr{H}^{\alpha}$. Because the Coulomb potential is a relatively bounded operator,[7] the finite-rank approximations of H and e^{iHt} can be used from the short-range case. The anomalous term in the asymptotic channel Hamiltonian has to be treated additionally. Because $A^{IJc}(t)$ commutes with $A^{I'J'c}(t)$ for IJ different from $I'J'$, it is sufficient to find for each IJ a set of expansion functions. Because $A^{IJc}(t)$ effects only the relative motion between the generalized particles I,J one chooses the above expansion functions only in the Jacobi variable corresponding to that relative motion. Let P_N^{IJ} denote the corresponding orthogonal projectors, with

$$P_N^{IJ}A^{IJc}(t)P_N^{IJ}\psi\to A^{IJc}(t)\psi \tag{3.13}$$

for $\psi\in D(A^{IJc}(t))$. Thus we are led in the Coulomb case to the following.

Conjecture: Let V^{ij},V^{α},V be defined as above but may include also Coulomb potentials, such that H,H^{α} are self-adjoint with $D(H)=D(H^{\alpha})=D(H^0)$. Let V^{ij} be bounded or relatively bounded with respect to H^{0ij}, assume the existence of the wave operators $\Omega^{\alpha c(\pm)}$. Let P_N denote orthogonal projectors corresponding to an arbitrary complete set of Jacobi variables. Let P_N^{α}, P_N^{IJ} denote the orthogonal projectors as described above. Let

$$\begin{aligned}
&A_N^{IJc}(t) = P_N^{IJ} A^{IJc}(t) P_N^{IJ}, \\
&A_N^{\alpha c}(t) = \sum_{\substack{1 < I < J < I_{\max} \\ IJ \subset \alpha}} A_N^{IJc}(t), \\
&H_N^{\alpha c}(t) = P_N^{\alpha} H^{\alpha} \cdot t P_N^{\alpha} + A_N^{\alpha c}(t), \\
&H_N = P_N H P_N .
\end{aligned} \tag{3.14}$$

Then for each channel α, $\psi \in \mathscr{H}^{\alpha}$

$$\|\Omega^{\alpha c(+)}\psi - \exp(-iH_N T)\exp(iH_N^{\alpha c}(T))\psi\|$$

and

$$\|\Omega^{\alpha c(-)}\psi - \exp(iH_N T)\exp(-iH_N^{\alpha c}(T))\psi\|$$

can be made arbitrarily small for suitable $N \in \mathrm{N}$, $T > 0$.

IV. CONCLUSION

It has been shown in the two-body system and indicated in the N-body system, how wave operators can be approximated by finite-rank operators which makes its calculation relatively simple. That holds for a wide class of short-range potentials and includes the long-range Coulomb potential. The finite-rank approximations are obtained from a set of expansion functions, using step functions in momentum space. The question can be asked, if that would also hold for other sets of expansion functions, like, e.g., the Hermite functions. The crucial point is to find a core C such that $P_N H^0 P_N \psi \to H^0 \psi$ for $\psi \in C$. One can find D_0, a dense subset of $D(H^0)$ where that convergence property holds, but D_0 could not be verified to be a core.

ACKNOWLEDGMENT

This work has been supported by NSERC—Canada.

[1] H. Kröger, J. Math. Phys. **24**, 1509 (1983).

[2] If the long-range Coulomb force is involved a modified wave operator has to be used, see J. D. Dollard, J. Math Phys. **5**, 729 (1964); Rocky Mountain J. Math. **1**, 5 (1971); **2**, 317 (1972).

[3] H. Kröger, Annual report of the Van de Graaff Laboratory, University Laval, Québec, p. 80 (1982).

[4] H. Kröger, A. M. Nachabe, and R. J. Slobodrian, in *Tenth International Conference on Few Body Problems in Physics*, Karlsruhe (1983), p. 262; A. M. Nachabe *et al.*, *ibid.*, p. 250.

[5] H. Kröger, Phys. Lett. B **135**, 1 (1984).

[6] J. Weidmann, *Linear Operators in Hilbert Spaces* (Springer-Verlag, New York, 1980).

[7] W. O. Amrein, J. M. Jauch, and K. B. Sinha, *Scattering Theory in Quantum Mechanics, Lecture Notes and Supplements in Physics* (Benjamin, Reading, MA, 1977).

Spectral properties of vaguely elliptic pseudodifferential operators with momentum-dependent long-range potentials using time-dependent scattering theory

Pl. Muthuramalingam
Indian Statistical Institute, New Delhi 110016, India

(Received 5 April 1983; accepted for publication 23 September 1983)

(a) Absence of singular continuous spectrum for H and unitary equivalence of $h_0(P)$ with the absolutely continuous part of H is proved for the self-adjoint operator $H = h_0(P) + W_S(Q, P) + W_L(Q, P)$ on $L^2(R^n)$, where (i) h_0 is a smooth, nonnegative, real valued function on R^n of at most polynomial growth with $h_0(\infty) = \infty$, (ii) the critical values of h_0 has a countable closure, (iii) the symbol $W_S(x, \xi)$ is a smooth function of at most polynomial growth in ξ and is of short range in x, and (iv) $W_L(x, \xi)$ is a smooth real-valued function of x, ξ with at most a polynomial growth in ξ and $O(|x|^{-\delta})$ at ∞ for some $\delta > 0$. Further we have (b) essential spectra of H and $h_0(P)$ are equal and (c) eigenvalues of H can accumulate only at the critical values of h_0.

PACS numbers: 03.65.Nk, 02.30.Tb, 03.65.Fd

1. INTRODUCTION

Let the self-adjoint operator $H = h_0(P) + W_S(Q, P) + W(Q, P)$ be as in Sec. 3 satisfying conditions (i)–(xi). Our aim is to prove

Theorem 1: (a)

$$\sigma_{ess}(H) = \sigma_{ess}(H_0),\ H_0 = h_0(P).$$

(b) There exists a unitary evolution $\exp[-iX(t,P)]$ such that the wave operator $\Omega_\pm = \text{s-lim}_{t\to\pm\infty} \exp(itH) \times \exp[-iX(t,P)]$ exists.

(c) Range $\Omega_\pm = \mathscr{H}_c(H)$, the continuous spectral subspace of H.

(d) $\{0\} = \mathscr{H}_{sc}(H)$, the singularly continuous spectral subspace of H.

(e) $\Omega_\pm : L^2(R^n) \rightarrow \mathscr{H}_{ac}(H)$ (the absolutely continuous spectral subspace of H) is unitary and $\Omega_\pm^* H | \mathscr{H}_{ac} \Omega_\pm = H_0$.

(f) The eigenvalues of H can accumulate only at the critical values of h_0.

When $h_0(\xi) = \frac{1}{2}\xi^2$, $W(x,\xi) = W_L(x,\xi) = 0$, and $W_S(x, \xi) = W_S(x)$ is of short range the existence of the wave operators $\Omega_\pm$ as strong limits of $\exp(itH)\exp(-itH_0)$ as $t\rightarrow \pm\infty$ for the pair (H, H_0) is well known,[1–4] where $H_0 = h_0(P)$, $H = h_0(P) + W_S(Q, P)$. The completeness, i.e., Range $\Omega_+ =$ Range $\Omega_- = \mathscr{H}_{ac}(H)$ is also known using time-independent methods.[1,4,5] Using time-dependent methods strong asymptotic completeness, viz., Range $\Omega_+ =$ Range $\Omega_- = \mathscr{H}_c(H)$ was proved by Enss.[6] The method was refined by Davies,[7] Mourre,[8] Perry,[9] and Sinha.[10] For extensions in this direction one can refer Simon[11] and Umeda.[12,13]

When there is also long range present, i.e., when $H = \frac{1}{2}P^2 + W_S(Q) + W(Q)$ with W smooth long range potential of $O(|x|^{-\delta})$ at ∞ for some $\delta > 0$ existence of the wave operators was proved by Alsholm,[14] Buslaev–Matveev,[15] Hormander,[16] and Berthier and Collet.[17] Using time-independent methods, completeness was proved by many people which includes Lavine,[18] Kitada and Ikebe-Isozaki,[19] Agmon,[20] Amrein, Martin, and Misra,[21] and Thomas.[22] By using time-dependent methods, completeness was proved by Enss[23,24] for $\delta > (2n+2)/(2n+3)$, Perry,[25] for $\delta > \frac{1}{2}$ with the further assumption W is dilation analytic, Muthuramalingam and Sinha[26,27] and Muthuramalingam[28] for $\delta > \frac{1}{2}$; Muthuramalingam[29] for $\delta = 1$, and Kitada and Yajima[30] for $\delta > 0$. While Refs. 25 and 27–29 require the theory of asymptotic evolution of observable, Refs. 23, 24, and 30 do not require it. The results of Umeda[13] for short range are extended here to long range. While we use the ideas of Enss,[23] Kitada and Yajima,[30] for estimating the norms we use the techniques of Hormander.[16]

Finally we sketch the contents of the article. In Sec. 2, we introduce a positive operator valued map T which was introduced and used by Davies in Ref. 7. Also it was used successfully in Refs. 27 and 28. This operator T will help us make calculations simultaneously in position and momentum variables. We derive certain simple useful properties of T. We also have some simple theorems on integral kernels.

In Sec. 3 we state the assumptions on the Hamiltonian H and prove Theorem 1(a) using the arguments of Ref. 13. Let us remark that, in Ref. 13, (a) of Theorem 1 was a consequence of (c) and (f).

In Sec. 4, we derive the modified free evolution and a position-momentum-dependent evolution approximating the total evolution. This can be done either by solving the Hamilton–Jacobi equation[16,30] or use iteration,[14,15] and we use iteration. We derive various growth and decay properties of the iterations to be used in Secs. 5 and 6.

In Sec. 5 we sketch a proof of the existence of the wave operators for two reasons: (i) for the sake of fullness and (ii) to bring out the similarity between existence proof and completeness proof. We reproduce[16] almost verbatim.

In Sec. 6 we prove that the total evolution can be approximated in operator norm by an auxiliary position-momentum dependent evolution. Here we closely follow.[16]

In Sec. 7 we prove Theorem (1) (c)–(f). In Sec. 8, we prove smoothness and growth properties of diffeomorphisms of functions used in Sec. 6.

We very freely use the techniques from Refs. 13, 16, 23, and 30.

 0022-2488/84/061881-19$02.50

Throughout this publication K will denote a generic constant and $F(M)$ the indicator (characteristic) function of the set M.

2. A POSITIVE OPERATOR VALUED MEASURE

Choose and fix some c^* in (0, 1), η in $\mathscr{S}(R^n)$—the Schwartz space of rapidly decreasing functions such that $\hat{\eta}$—the Fourier transform of η defined by

$$\hat{\eta}(k) = (2\pi)^{-n/2}\int dx\, e^{-ik\cdot x}\eta(x)$$

has

$$\operatorname{supp}\hat{\eta}\subset\{k\in R^n:\ |k|\leqslant\tfrac{1}{8}c^*\}. \tag{2.1}$$

We further normalize η by

$$\|\eta\|^2 = \int dx\, |\eta(x)|^2 = 1. \tag{2.2}$$

Define for (x, k) in $R^n\times R^n$ the function η_{xk} by

$$\eta_{xk}(y) = \eta(y-x)\exp[ik\cdot(y-x)] \tag{2.3}$$

so that

$$(\eta_{xk})\hat{}\,(p) = \hat{\eta}(p-k)\exp[-ix\cdot p]. \tag{2.4}$$

η_{xk} is called a generalized coherent state.

Define Z_η: $L^2(R^n)\to L^\infty(R^n\times R^n)$ by

$$(Z_\eta f)(x, k) = \langle\eta_{xk}|f\rangle, \tag{2.5}$$

where $\langle\ |\ \rangle$ denotes the inner product on $L^2(R^n)$, linear in the second variable. Then it is known[31] that Z_η takes values in $L^2(R^n\times R^n)$ and Z_η: $L^2(R^n)\to L^2(R^n\times R^n)$ is an isometry but not unitary. For any Borel subset M of $R^n\times R^n$ let $F(M)$ denote the indicator (characteristic) function of M. For every Borel subset M of $R^n\times R^n$ define the operator $T(M): L^2(R^n)\to L^2(R^n)$ by

$$T(M) = Z_\eta^* F(M) Z_\eta. \tag{2.6}$$

Clearly T is a positive operator-valued measure defined on the Borel subsets of $R^n\times R^n$ so that for Borel subsets M_1, M_2 of $R^n\times R^n$

$$0\leqslant T(M_1\cup M_2)\leqslant T(M_1) + T(M_2), \tag{2.7}$$

$$T(M_1\cup M_2) = T(M_1) + T(M_2) \quad \text{if } M_1, M_2 \text{ are disjoint}, \tag{2.8}$$

$$0\leqslant T(M_1)\leqslant T(M_2) \ \text{ if } M_1\subset M_2, \tag{2.9}$$

and

$$0\leqslant T^2(M_1)\leqslant T(M_1)\leqslant T(R^n\times R^n) = 1. \tag{2.10}$$

$T^2(M)$ stands for $(T(M))^2$.

An explicit representation of T is given by

$$T(M) = (2\pi)^{-n/2}\int_M dx\, dk\, \langle\eta_{xk}|\,\rangle\eta_{xk}, \tag{2.11}$$

where the integral is the weak integral.

Let us remark that $Q = (Q_1, Q_2, ..., Q_n)$, $P = (P_1, P_2, ..., P_n)$, $P_j = -i\,\partial/\partial x_j = -iD_j$ will denote the self-adjoint operator families on $L^2(R^n)$ representing the position and momentum observables, respectively.

Of special interest is when $M_1 = B_1\times R^n$ and $M_2 = R^n\times B_2$ with B_1, B_2 Borel subsets of R^n. In such a case $T(B_1\times R^n)$ $[T(R^n\times B_2)]$ is a multiplication operator in the position [momentum] space and is given by

$$T(B_1\times R^n) = (F(B_1)*|\eta|^2)(Q), \tag{2.12}$$

$$T(R^n\times B_2) = (F(B_2)*|\hat{\eta}|^2)(P), \tag{2.13}$$

where $*$ denotes the convolution operation. All the above can be found in Refs. 7, 31, and 32.

Lemma 2.1: (i) Let $0\leqslant X\leqslant Y_1$, $0\leqslant X\leqslant Y_2$ be bounded self-adjoint operators on a Hilbert space with Y_1Y_2 compact. Then X is compact.

(ii) For any two bounded Borel subsets B_1, B_2 of R^n the operator $T(B_1\times B_2)$ is compact.

(iii) For any bounded Borel subset M of $R^n\times R^n$ the operator $T(M)$ is compact.

(iv) Let X, $0\leqslant Y_1 < Y_2$, be bounded operators on a Hilbert space. If XY_2 is compact then so is XY_1.

Proof: (i) Since $0\leqslant Y_1XY_1\leqslant Y_1Y_2Y_1$, we get Y_1XY_1 is compact. Noting $(\sqrt{X}\ Y_1\sqrt{X})^2 = \sqrt{X}\ Y_1XY_1\sqrt{X}$, we have $\sqrt{X}Y_1\sqrt{X}$ is compact. The result follows by noting $0\leqslant X^2\leqslant\sqrt{X}\,Y_1\sqrt{X}$.

(ii) It is well known that if f_1 and f_2 are bounded real-valued functions on R^n vanishing at ∞, then the operator $f_1(Q)f_2(P)$ on $L^2(R^n)$ is compact. The result follows from (i) by taking $Y_1 = T(B_1\times R^n)$, $Y_2 = T(R^n\times B_2)$, and $X = T(B_1\times B_2)$ and using (2.12), (2.13), and (2.9).

(iii) If M is bounded, then there are two bounded Borel subsets of R^n such that $M\subset B_1\times B_2$. The result follows from (i) and (ii) by noting $0\leqslant T(M)\leqslant T(B_1\times B_2)$.

(iv) By assumption $|X|Y_2|X|$ is compact. Since $0\leqslant|X|Y_1|X|\leqslant|X|Y_2|X|$, we have $|X|\sqrt{Y_1}$ is compact and now the result follows. Q.E.D.

Lemma 2.2: (i) For Borel subsets M_1, M_2 of $R^n\times R^n$ and f in $L^2(R^n)$

$$\|T(M_1\cup M_2)f\|^2\leqslant[\|T(M_1)f\| + \|T(M_2)f\|]\|f\|.$$

(ii) For M_1, M_2 as above and Y any bounded operator on $L^2(R^n)$

$$\|T(M_1\cup M_2)Y\|^2\leqslant[\|T(M_1)Y\| + \|T(M_2)Y\|]\|Y\|.$$

(iii) For $M_1\subset M_2$ and f in $L^2(R^n)$

$$\|T(M_1)f\|^2\leqslant\|T(M_2)f\|\ \|f\|.$$

(iv) For $M_1\subset M_2$ and Y any bounded operator on $L^2(R^n)$

$$\|T(M_1)Y\|^2\leqslant\|T(M_2)Y\|\ \|Y\|.$$

(v) Let B_2 be any Borel subset of R^n, M any Borel subset of $R^n\times B_2$ and ϕ any bounded continuous function on R^n such that $\operatorname{dist}(B_2, \operatorname{supp}\phi) > \frac{1}{8}c^*$. Then

$$\phi(P)T(M) = 0 = T(M)\phi(P).$$

Proof: (i) Follows from (2.7), (2.10), and Cauchy–Schwartz inequality. (ii) Follows from (i). (iii) Follows from (2.9), (2.10), and Cauchy–Schwartz inequality. (iv) Follows from (iii). (v) Clearly it is enough to prove when ϕ is real-valued.

Case (i): Let $M = R^n\times B_2$. Then by (2.13) the operator $T(M)$ is a multiplication operator in the momentum space with support in a $\frac{1}{8}c^*$ neighborhood of B_2 and so the result is clear.

Case (ii): Let $M \subset R^n \times B_2$. Then by (iv) we get $\|T(M)\phi(P)\|^2 \leqslant \|T(R^n \times B_2)\phi(P)\| \, \|\phi(P)\|$ so that by case (i) $T(M)\phi(P) = 0$. Taking adjoints, one gets
$\phi(P)T(M) = 0$. Q.E.D.

Let us also note that, for any bounded complex valued function ϕ on R^n and for any Borel subset M of $R^n \times R^n$, we get

$$[\phi(P)T(M)f](q) = \int_M dx\, dk\, \langle \eta_{xk}|f\rangle \times \int d\xi\, \phi(\xi)\hat{\eta}(\xi - k) \times \exp[i(q\cdot\xi - x\cdot\xi)] \tag{2.14}$$

for f in $L^2(R^n)$, q in R^n.

Lemma 2.3: Let I: $R^n \times R^n \to \mathbb{C}$ be a measurable function such that $\int dq\, |I(q,x)| \leqslant C_1^2$ a.e. x and $\int dx\, |I(q,x)| \leqslant C_2^2$ a.e. q. Then the operator $\mathscr{I}: L^2(R^n) \to L^2(R^n)$ defined by $(\mathscr{I}f)(q) = \int dx I(q,x) f(x)$ is bounded and $\|\mathscr{I}\| \leqslant C_1 C_2$.

Proof: The above lemma is the same as corollary of Theorem 6.24 of Ref. 3.

Lemma 2.4: Let I: $R^n \times R^n \times R^n \to \mathbb{C}$ be any measurable function and B any bounded subset of R^n. Further let, for each fixed k in B, $I(k,\cdot,\cdot)$ be the integral kernel of a bounded operator $\mathscr{I}(k)$: $L^2(R^n) \to L^2(R^n)$ and

$$C = \sup\{\|\mathscr{I}(k)\| : k \in B\}.$$

Define an operator on $\mathscr{S}(R^n)$ by

$$(Lf)(q) = \int_{R^n \times B} dx\, dk\, \langle \eta_{xk}|f\rangle I(k,q,x)$$

for f in $\mathscr{S}(R^n)$ and q in R^n. Then L is a bounded operator on $L^2(R^n)$ and

$$\|L\| \leqslant [\mathrm{meas}(B)]C.$$

Proof: Define $F(k,x) = \langle \eta_{xk}|f\rangle$ for f in $\mathscr{S}(R^n)$. Then

$$\int dq\, |(Lf)(q)|^2 = \int dq \left| \int_B dk\, [\mathscr{I}(k)F(k,\cdot)](q) \right|^2 \leqslant [\mathrm{meas}(B)]^2 \int dk \int dq\, |[\mathscr{I}(k)F(k,\cdot)](q)|^2$$

by Cauchy–Schwartz inequality and Fubini's Theorem

$$\leqslant [\mathrm{meas}(B)]^2 \int_B dk\, \|\mathscr{I}(k)F(k,\cdot)\|^2 \leqslant [\mathrm{meas}(B)]^2 C^2 \int dk\, dx\, |F(k,x)|^2 = [\mathrm{meas}(B)]^2 C^2 \|f\|^2.$$

The result is now clear. Q.E.D.

We now state a result similar in spirit to that of Calderon–Vaillan court theorem.[33] Unlike Ref. 33, we do not use the notion of oscillatory integrals, and we give a simple proof for our case.

Theorem 2.5: Let X: $R^n \times R^n \to R$, a: $R^n \times R^n \to \mathbb{C}$ be C^∞ functions with all the derivatives bounded. Assume that there is a compact set C such that $a(x,\xi) = 0$ if ξ is not in C, x in R^n. On X we impose the following conditions:

$$\sup\{|\nabla_\xi \nabla_x X(x,\xi) - I| : \xi \in C_1,\ x \in R^n\} \leqslant \tfrac{1}{2}, \tag{2.15}$$

and

$$\sup\{|D_\xi^\alpha \nabla_x X(x,\xi)| : \xi \in C_1, x \in R^n\} \leqslant K_\alpha \quad \text{for } |\alpha| \geqslant 1, \tag{2.16}$$

where C_1 is a compact neighborhood of C.

Now define A: $R^n \times R^n \to \mathbb{C}$ by

$$A(q,x) = \int_C d\xi\, a(x,\xi) \exp(i[q\cdot\xi - X(x,\xi)]).$$

Then $A(\cdot,\cdot)$ is the integral kernel of a bounded operator $\mathscr{A}$ on $L^2(R^n)$ and

$$\|\mathscr{A}\| \leqslant K_M(C_1) \sum_{|\alpha| \leqslant M} \sup_{x,\xi} |D_\xi^\alpha(a(x,\xi))|, \quad M \geqslant n+1,$$

where $K_M(C_1)$ depends on C and a bound on

$$\sup\{|D_\xi^\alpha \nabla_x X(x,\xi)| : \xi \in C_1, x, 1 \leqslant |\alpha| \leqslant M+1\}.$$

Proof: Let $\mathscr{B} = \mathscr{A}^*\mathscr{A}$. Then an easy calculation shows that $B(q,x)$, the integral kernel of $\mathscr{B}$, is given by

$$B(q,x) = \int_C d\xi\, \bar{a}(q,\xi) a(x,\xi) \exp(i[X(q,\xi) - X(x,\xi)]). \tag{2.17}$$

Let $M(n,R)$, $\mathrm{GL}(n,R)$ be the set of all $n \times n$, $n \times n$ invertible matrices over R and $S = \{E \in M(n,R) : |E - I| \leqslant \frac{1}{2}\}$. Then S is a compact convex subset of $\mathrm{GL}(n,R)$. Now

$$\nabla_\xi[X(q,\xi) - X(x,\xi)] = \int_0^1 d\rho\, \nabla_x \nabla_\xi X(\rho q + (1-\rho)x, \xi)\cdot(q - x) = \phi(q,x,\xi)\cdot(q-x), \tag{2.18}$$

where, by (2.15), the matrix $\phi(q,x,\xi)$ is in S for all ξ in C_1, q, x. From (2.18)

$$|q - x| = |[\phi(q,x,\xi)]^{-1} \nabla_\xi\{X(q,\xi) - X(x,\xi)\}| \leqslant K_S |\nabla_\xi\{X(q,\xi) - X(x,\xi)\}|, \tag{2.19}$$

where $0 < K_S < \infty$ depends only on S. By (2.16) and the mean value theorem we get

$$|D_\xi^\alpha\{X(q,\xi) - X(x,\xi)\}| \leqslant K_\alpha |q - x| \quad \text{for } |\alpha| \geqslant 1. \tag{2.20}$$

Now apply the stationary phase Lemma A.1 of Ref. 16 or Lemma I of the Appendix of Ref. 12 to (2.17), using (2.19) and (2.20), to get

$$|B(q,x)| \leqslant K_M(1 + |q-x|)^{-M} \times \sum_{|\alpha| \leqslant M} \sup_{q,x,\xi} |D_\xi^\alpha\{\bar{a}(q,\xi) a(x,\xi)\}|$$

so that, by Lemma 2.3,

$$\|\mathscr{A}\|^2 \leqslant K_M \sum_{|\alpha| \leqslant M} \sup_{q,x,\xi} |D_\xi^\alpha\{\bar{a}(q,\xi) a(x,\xi)\}| \quad \text{if } M \geqslant n+1.$$

Now the result is immediate. Q.E.D.

Remark 2.6: Suppose X is as in Theorem 2.5 and $a: R^n \times R^n \times R^n \to \mathbb{C}$ is C^∞ with all the derivatives bounded and for some compact set C we have $a(q, x, \xi) = 0$ for ξ not in C. Now define the kernel

$$A(q, x) = \int d\xi\, a(q, x, \xi) \exp(i[q\cdot\xi - X(x, \xi)]) \quad (2.21)$$

and $\mathscr{A}$ be the operator on $L^2(R^n)$ induced by A.

For the simple case when $X(x, \xi) = x\cdot\xi$, $\|\mathscr{A}\|$ has a bound in terms of the derivatives of the symbol a by the Calderon–Vaillan Court theorem.[33]

For the general X, the answer is not known to the author. If one can prove a result similar to Theorem 2.5 for (2.21), then the proof of Lemma 6.1 (vii) can be made short and simple.

The next lemma will be useful in the proof of Lemma 6.2 (ii). It is adapted from Ref. 34.

Lemma 2.7: Let B be any bounded subset of R^n. For each $m = 1, 2, \ldots$ and k in B, let $\mathscr{I}_m(k)$ be an integral operator on $L^2(R^n)$ with kernel $I_m(k, q, x)$ so that $[\mathscr{I}_m(k) f](q) = \int dx\, I_m(k, q, x) f(x)$. Further assume that

(i) $\sup\{\|\mathscr{I}_m(k)\|: k \text{ in } B,\ m = 1, 2, \ldots\} < \infty$,

(ii) $\text{s-lim}_{m\to\infty} \mathscr{I}_m(k)$ exists for each k in B.

Define $(L_m f)(q) = \int_{R^n\times B} dx\, dk \langle \eta_{xk} | f\rangle I_m(k, q\, x)$.

Then $\text{s-lim}_{m\to\infty} L_m$ exists.

Proof: *Step i*: For g in $L^2(R^n)$ define $G: R^n \to L^2(R^n)$ by $[G(k)](x) = \langle \eta_{xk} | g\rangle$. We show that $G(B)$ is a bounded subset of $L^2(R^n)$.

Let θ_k denote the multiplication operator (in the position variable) by $\bar{\hat{\eta}}(\cdot - k)$ and $\mathscr{F}$ the Fourier transform. Then, using Parseval's relations, we see that $G(k) = C\mathscr{F}^{-1}\theta_k \mathscr{F} g$ for some constant C. Since $\hat{\eta} \in C_0^\infty(R^n)$, G is continuous. Since B is bounded, it is a subset of a compact subset and so $G(B)$ is bounded in $L^2(R^n)$.

Step ii: For g in $L^2(R^n)$ it suffices to show that $L_m g$ is a Cauchy sequence. With G as in step (i), a careful look at the proof of Lemma 2.4 yields

$$\|L_m g - L_p g\|^2 \leqslant [\text{meas}(B)]^2 \int_B dk\, \|[\mathscr{I}_m(k) - \mathscr{I}_p(k)] G(k)\|^2.$$

Now the result is a consequence of Lebesgue dominated convergence Theorem, the assumptions on $\mathscr{I}_m(k)$ and step (i). Q.E.D.

3. THE HAMILTONIAN

In this section we state the assumptions on the free and total Hamiltonians $h_0(P)$, H and prove Theorem 1(a).

Let $h_0: R^n \to R$ be any function such that

(i) h_0 is C^∞,
(ii) $\lim_{|\xi|\to\infty} |h_0(\xi)| = \infty$,
(iii) $|h_0(\xi)| \leqslant K(1 + |\xi|)^k\ \forall \xi$ for suitable constants K, k,
(iv) $\{\xi: h_0'(\xi) = 0 \text{ or } \det h_0''(\xi) = 0\}$ is a (closed) subset of R^n of (Lebesgue) measure 0.

We shall denote by H_0 the closure in $L^2(R^n)$ of $h_0(P)$ with domain $\mathscr{S}(R^n)$.

For the part $W_S(Q, P)$ with the symbol $W_S(x, \xi)$, we make the same assumptions as in Ref. 13. For this we introduce, as in Ref. 13, the following notation; let V be a measurable function on R^n. If there exist constants $\epsilon > 0$, θ, γ in R, such that, for some constant K and measurable set $\Gamma \subset R^n$ with

$$\text{meas}(\Gamma \cap \{x \in R^n: |x| \geqslant r\}) = O(r^{-2\theta}) \quad \text{as } r \to \infty, \quad (3.1)$$

the estimates

$$|V(x)| \leqslant K(1 + |x|)^{-1-\epsilon} \quad \text{for } x \text{ in } R^n \setminus \Gamma \quad (3.2)$$

$$|V(x)| \leqslant K(1 + |x|)^\gamma \quad \text{for } x \text{ in } \Gamma \quad (3.3)$$

are valid for a measurable function V then we say V is in $V(\epsilon, \gamma, \theta)$. Our assumption on $W_S(x, \xi)$ is

(v) $W_S: R^n \times R^n \to \mathbb{C}$ is measurable and for each fixed x the map $W_S(x, \cdot)$ is C^∞ function of ξ,

(vi) there are constants $\epsilon > 0$, $\gamma \geqslant -1$, $\theta > \gamma + 1$, and m^* in R such that for some V in $V(\epsilon, \gamma, \theta)$ the estimate

$$|D_\xi^\mu W_S(x, \xi)| \leqslant |V(x)|(1 + |\xi|)^{m^*}$$

is valid for every multi-index μ with $|\mu| \leqslant 2(2 + [\frac{1}{2}(n + 2|\gamma|)])$. Here $[a]$ denotes the integral part of the real number a.

For the long range part $W(Q, P)$ we assume on the symbol $W(x, \xi)$ the following:

(vii) $W: R^n \times R^n \to R$ is a C^∞ function and there is some δ in (0,1] such that for any compact subset B of R^n

$$\sup_{\xi\in B} |D_\xi^\alpha D_x^\beta W(x, \xi)| \leqslant K(B, \alpha, \beta)(1 + |x|)^{-|\beta| - \delta}$$

for all α, β where $K(B, \alpha, \beta)$ are constants,

(viii) for the same m^* as in (vi) we have

$$|D_\xi^\mu W(x, \xi)| \leqslant K_\mu (1 + |x|)^{-\delta}(1 + |\xi|)^{m^*}$$

for suitable constants K_μ for all μ with $|\mu| \leqslant 2([\frac{1}{2}n] + 2)$.

Now let us recall that[13] the operators $W_S(Q, P)$, $W(Q, P)$ are given for f in $\mathscr{S}(R^n)$ by

$$\{W_S(Q, P) f\}(q) = \int d\xi\, \hat{f}(\xi) W_S(q, \xi) \exp(iq\cdot\xi),$$
$$\{W(Q, P) f\}(q) = \int d\xi\, \hat{f}(\xi) W(q, \xi) \exp(iq\cdot\xi). \quad (3.4)$$

When the above assumptions are satisfied we can prove

Lemma 3.1: Let N be any positive integer such that $N \geqslant k$ and $N > m^* + n$. Then

(i) for some $\epsilon_0 > 0$ the operator $W_S(Q, P) (1 + P^2)^{-N/2}(1 + |Q|)^{1+\epsilon_0}$ is bounded,

(ii) $W(Q, P)(1 + P^2)^{-N/2}(1 + |Q|)^\delta$ is bounded.

Proof: (i) is the same as Corollary 3.2 of Ref. 13 and proof of (ii) is similar to the proof of Corollary 3.2 of Ref. 13 by using the condition (viii). Q.E.D.

By Lemma 3.1 we see that the operator $W_S + W = W_S(Q, P) + W(Q, P)$ is defined on $\text{Dom}((1 + |P|)^N)$ taking values in $L^2(R^n)$. We make, as in Ref. 13, the final assumptions:

(ix) The operator $W_S + W$ is symmetric and

(x) $H_0 + (W_S + W)$ on $\text{Dom}(1 + |P|^N)$ has a self-adjoint extension H on $L^2(R^n)$ such that

Dom $H \subset \mathrm{Dom}((1+|P|)^s)$ for some $s>0$.

Also we may impose the following condition on critical values, viz.,

(xi) Closure of $\{h_0(\xi): h_0'(\xi)=0 \text{ or } \det h_0''(\xi)=0\}$ is a countable set.

Now we give some examples.

Example 3.2: (Hamiltonian for the static electromagnetic field): Take $h_0(\xi)=\frac{1}{2}\xi^2$, $W_S(Q,P)=W_{S,0}(Q) + \Sigma_{j=1}^n W_{S,j}(Q)P_j$ and $W(Q,P)=W_0(Q)+\Sigma_{j=1}^n W_j(Q)P_j$, where (i) the functions $W_{S,j}$, W_j are all real-valued and W_j are C^∞ for $j=0,1,2,...,n$, (ii) for some $\epsilon>0$ the operators $W_{S,0}(H_0+1)^{-1}(1+|Q|)^{1+\epsilon}$, $W_{S,j}(Q)(H_0+1)^{-1/2}(1+|Q|)^{1+\epsilon}$ are all bounded for $j=1,2,...,n$, (iii) for some $\delta>0$ $\Sigma_{j=0}^n |D^\alpha W_j(x)| \leqslant K_\alpha(1+|x|)^{-|\alpha|-\delta}$ for suitable constants K_α for all multi-indices α, and (iv) $0=\Sigma_{j=1}^n \partial W_{S,j}(x)/\partial x_j + \Sigma_{j=1}^n \partial W_j(x)/\partial x_j$ in the sense of distributions. The condition (iv) is only to ensure symmetry of H.

In this case note that (xi) is satisfied for h_0. Further, the symbols W_S, W for the operators $W_S(Q,P)$, $W(Q,P)$ are given by $W_S(x,\xi)=W_{S,0}(x)+\Sigma_{j=1}^n W_{S,j}(x)\xi_j$ and $W(x,\xi) = W_0(x)+\Sigma_{j=1}^n W_j(x)\xi_j$, respectively.

Though this example does not satisfy the condition (v) and (vi), nevertheless, Lemma 3.1 is valid. A careful look at the proofs will show that all the results hereafter are valid under the statement of Lemma 3.1. So Theorem 1 is valid for Example 3.2.

Example 3.3[13]: Let $n=3$, $H_0=(1+P^2)^{1/2}$, $W_S(Q,P)=(1+Q^2)^{-\delta}(1+P^2)^{3/2}(1+Q^2)^{-\delta}$ with $\delta>\frac{1}{4}$ and $W(Q,P)=0$. Clearly, h_0 satisfies condition (xi). Let us note that the symbols W, W_S for the operators $W(Q,P)$, $W_S(Q\,P)$ are given by $W(x,\xi)=0$ and $W_S(x,\xi)=\mathrm{Osc}$ $\int\int dy\, d\lambda\, (1+x^2)^{-\delta}(1+(\xi+\lambda)^2)^{3/2}(1+(x+y)^2)^{-\delta} \times \exp[-i(y\cdot\lambda)]$. For details see Refs. 12 and 13.

Now let the conditions (i)–(x) be satisfied. Let $\zeta \in C_0^\infty(R^n)$ be such that $\zeta(x)=1$ for $|x|\leqslant 1$, 0 for $|x|\geqslant 2$, and $0\leqslant\zeta(x)\leqslant 1$ for all x. Then

Lemma 3.4: (i) The operator $\zeta(Q/r)(H+i)^{-1}$ is compact for every $r>0$.

(ii) The operator $\zeta(Q/r)(H_0+i)^{-1}$ is compact for every $r>0$.

(iii) For ψ in $C_0^\infty(R)$ the operator $(1+P^2)^{s/2}\psi(H)$ is bounded.

(iv) For ψ in $C_0^\infty(R)$ the operator $(1+P^2)^{s/2}\psi(H_0)$ is bounded.

(v) for ϕ, ψ in $C_0^\infty(R)$ we have

$$\lim_{r\to\infty} \|\{\phi(H)-\phi(H_0)\}[1-\zeta(Q/r)]\psi(H)\|=0.$$

(vi) For ϕ, ψ in $C_0^\infty(R)$ we have

$$\lim_{r\to\infty} \|\{\phi(H)-\phi(H_0)\}[1-\zeta(Q/r)]\psi(H_0)\|=0.$$

(vii) For ϕ: $R\to\mathbb{C}$ continuous and vanishing at $\pm\infty$ the operator $\phi(H)-\phi(H_0)$ is compact.

Proof: (i) Same as Lemma 4.1 of Ref. 13.

(ii) Follows from the assumptions (i) and (ii) on h_0.

(iii) Same as Lemma 4.2 of Ref. 13.

(iv) Obvious.

(v) Same as Lemma 4.3 of Ref. 13.

(vi) Similar to the proof of Lemma 4.3 of Ref. 13.

(vii) *Step i*: Assume that $\phi\in C_0^\infty(R)$ and ϕ is real valued. By (i) and (v) the operator $\{\phi(H)-\phi(H_0)\}\phi(H)$ is compact. Similarly, using (ii) and (vi) we have the compactness of $[\phi(H)-\phi(H_0)]\phi(H_0)$. So $[\phi(H)-\phi(H_0)]^2$ is compact yielding the result.

Step ii: By step (i) the result is true for ϕ in $C_0^\infty(R)$. The result follows by noting that any ϕ as in hypothesis can be approximated uniformly by functions in $C_0^\infty(R)$. Q.E.D.

Proof of Theorem 1(a): By Lemma 3.4 (vii) the operator $(H+i)^{-1}-(H_0+i)^{-1}$ is compact. The result follows from Weyl's theorem on essential spectrum (given in, for example, Ref. 5 as Theorem XIII. 14). Q.E.D.

4. ITERATIONS FOR THE SOLUTION OF A HAMILTON–JACOBI EQUATION

In this section all we need to assume on h_0 is that h_0 is a C^∞ real-valued function on R^n and

$$G=\{\xi: h_0'(\xi)\neq 0, \det h_0''(\xi)\neq 0\} \tag{4.1}$$

is a nonempty (open) subset of R^n. On $W=W(x,\xi)$ we assume condition (vii) of Sec. 3.

Motivated by Ref. 30, we solve the Hamilton–Jacobi equation $\partial X(t,x,\xi)/\partial t = h_0(\xi)+W(\nabla_\xi X(t,x,\xi),\xi)$ with the initial condition $X(0,x,\xi)=x\cdot\xi$. What we do is simply use Picard's method of iteration.

Let δ be as in condition (vii) of Sec. 3. By decreasing δ, if necessary we assume that $\delta>0$ and 1 is not in $\{0,\delta,2\delta,...,j\delta,...\}$. Choose an integer m_0 such that

$$m_0\delta<1<(m_0+1)\delta. \tag{4.2}$$

Let C be a fixed compact subset of G such that for some $c>0$ we have

$$C_{3c}=\{p\in R^n: \operatorname{dist}(p,C)\leqslant 3c\}\subset G. \tag{4.3}$$

Let x, ξ in R^n be such that

$$x\cdot h_0'(k)\geqslant 0 \quad \text{for some } k \text{ in } C \tag{4.4}$$

and

$$|\xi-k|\leqslant\tfrac{1}{2}c \quad \text{with } k \text{ as in (4.4)}. \tag{4.5}$$

Let

$$3a=\inf\{|h_0'(k)|: k\in C_{3c}\}. \tag{4.6}$$

We can further assume that C_{3c} is chosen so that $\xi\to h_0'(\xi)$ is diffeomorphism on an open neighborhood of C_{3c} and

$$\sup\{|h_0'(\xi_1)-h_0'(\xi_2)|: |\xi_1-\xi_2|\leqslant 2c,\ \xi_1,\xi_2\in C_{3c}\}\leqslant a2^{-1/2}. \tag{4.7}$$

Then for x, k, ξ as in (4.4), (4.5) and $s\geqslant 0$ we get

$$\begin{aligned} &|x+sh_0'(\xi)| \\ &\quad\geqslant |x+sh_0'(k)|-sa2^{-1/2} \quad \text{by (4.7)}, \\ &\quad\geqslant 2^{-1/2}(|x|+3as)-sa2^{-1/2} \quad \text{by (4.4), (4.6)}, \\ &\quad\geqslant 2^{-1/2}(|x|+2as) \quad \text{with } a>0. \end{aligned} \tag{4.8}$$

For all the "position" x and "momentum" ξ [not necessarily satisfying (4.4) and (4.5)] define $Y(m, t_0, t, x, \xi)$ for integral $m = 0, 1, ..., m_0$, time $t \geqslant t_0 \geqslant 0$, by

$$Y(0, t_0, t, x, \xi) = 0, \tag{4.9}$$

$$Y(m, t_0, t, x, \xi) = \int_{t_0}^{t} ds\, W(x + s\nabla h_0(\xi) + \nabla_\xi Y(m-1, t_0, s, x, \xi), \xi) \tag{4.10}$$

so that if we define $X(m,t_0,t,x,\xi) = x\xi + th_0(\xi) + Y(m,t_0,t,x,\xi)$, then $\partial X(m,t_0,t,x,\xi)/\partial t = h_0(\xi) + W(\nabla_\xi X(m-1,t_0,t,x,\xi),\xi)$ and $X(m,t_0,t_0,x,\xi) = x\cdot\xi + t_0 h_0(\xi)$.

We study the decay and growth properties of the $Y(m)$ in Lemma 4.1 which can be compared with Propositions 2.3 and 2.8 of Ref. 30.

Lemma 4.1: Let x, ξ be as in (4.4) and (4.5), $0 \leqslant m \leqslant m_0$, $0 \leqslant M$. Then there exists a $t_{-1} = t_{-1}(m, M, C) \geqslant 0$ such that for $t \geqslant t_0 \geqslant t_{-1}$ the following hold:

(i) $|D_\xi^\alpha Y(m, t_0, t, x, \xi)| \leqslant K_\alpha (1+t)^{1-\delta}$ for $|\alpha| \leqslant M$.

(ii) $|D_\xi^\alpha D_x^\beta Y(m, t_0, t, x, \xi)|$

$\leqslant K_{\alpha\beta}(1+|x|)^{-\delta}$ for $\beta \neq 0$, $|\alpha+\beta| \leqslant M$.

(iii) There exists $r_0 = r_0(t_{-1})$ such that, for all $r \geqslant r_0$ and $|x| \geqslant r$,

$$|D_\xi^\alpha D_x^\beta\{W(x+(t-t_0)\nabla h_0(\xi) + \nabla_\xi Y(m_0, t_0, t, x, \xi), \xi)\}| \leqslant K_{\alpha\beta}(|x|+t)^{-\delta} \quad \text{for } |\alpha+\beta| \leqslant m,$$

(iv) There exists $r_0 = r_0(t_{-1})$ such that, for all $r \geqslant r_0$ and $|x| \geqslant r$,

$$|D_\xi^\alpha D_x^\beta\{W(x+(t-t_0)\nabla h_0(\xi) + \nabla_\xi Y(m_0, t_0, t, x, \xi), \xi) - W(x + t\nabla h_0(\xi) + \nabla_\xi Y(m_0, t_0, t, x, \xi), \xi)\}| \leqslant K_{\alpha\beta} t_0 (1+|x|+t)^{-1-\delta} \quad \text{for } |\alpha+\beta| \leqslant M.$$

(v) $|D_\xi^\alpha D_x^\beta\{Y(m, t_0, t, x, \xi) - Y(m-1, t_0, t, x, \xi)\}| \leqslant K_{\alpha\beta}(1+t)^{1-m\delta}$

for $1 \leqslant m \leqslant m_0$, $|\alpha+\beta| \leqslant M$.

(vi) $$|D_\xi^\alpha D_x^\beta\{W(x+t\nabla h_0(\xi) + \nabla_\xi Y(m_0, t_0, t, x, \xi), \xi) - W(x+t\nabla h_0(\xi) + \nabla_\xi Y(m_0-1, t_0, t, x, \xi), \xi)\}| \leqslant K_{\alpha\beta}(1+t)^{1-m_0\delta}(1+t+|x|)^{-1-\delta}$$

for $|\alpha+\beta| \leqslant M$,

(vii) $$\sup\{|W(t\nabla h_0(\xi) + \nabla_\xi Y(m_0, t_0, t, 0, \xi), \xi) - W(t\nabla h_0(\xi) + \nabla_\xi Y(m_0-1, t_0, t, 0, \xi), \xi)| : \xi \in C_{2c}\} \leqslant K(1+t)^{-(m_0+1)\delta}.$$

(viii) $\lim_{t\to\infty} Y(m_0, t_0, t, x, \xi) - Y(m_0, t_0, t, 0, \xi)$ exists.

(ix) $\lim_{t\to\infty} Y(m_0, t_0, t+s, 0, \xi) - Y(m_0, t_0, t, 0, \xi) = 0 \quad \forall s.$

Proof: Let us remark that the proof of (v)–(ix) will not depend on (iii) or (iv).

(i) and (ii): We prove simultaneously (i) and (ii) by induction on m. For the sake of simplicity we give the proof when n (the dimension of the space R^n) is 1. The results are easily true when $m = 0$. Assume the result to be true for $m \leqslant m_0 - 1$. Then

$$\begin{aligned} D_\xi^\alpha D_x^\beta Y(m+1, t_0, t, x, \xi) &= \int_{t_0}^{t} ds\, D_\xi^\alpha D_x^\beta\{W(x + sh_0'(\xi) + \partial_\xi Y(m, t_0, s, x, \xi), \xi)\} \\ &= \int_{t_0}^{t} ds\, D_\xi^\alpha D_x^\beta\{W(Z, \xi)\}, \end{aligned} \tag{4.11}$$

where $Z = Z(m, t_0, s, x, \xi) = x + sh_0'(\xi) + \partial_\xi Y(m, t_0, s, x, \xi)$. By the induction hypothesis on $Y(m)$ it is clear that

$$|D_\xi^\alpha Z| \leqslant K_\alpha(1+s) \quad \text{for } \alpha \neq 0, \tag{4.12}$$

and

$$|D_\xi^\alpha D_x^\beta Z| \leqslant K_{\alpha\beta} \quad \text{for } \beta \neq 0. \tag{4.13}$$

Also, using (4.8) and choosing t_{-1} still larger if necessary, we easily get

$$|Z| \geqslant K_0(1+|x|+s) \quad \text{with } K_0 > 0. \tag{4.14}$$

So using (4.14) and condition (vii) of Sec. 3

$$\begin{aligned} |Y(m+1, t_0, t, x, \xi)| &\leqslant \int_{t_0}^{t} ds |W(Z, \xi)| \\ &\leqslant K\int_{t_0}^{t} ds\,(1+|x|+s)^{-\delta} \\ &\leqslant K(1+t)^{1-\delta}. \end{aligned}$$

Thus we have proved (i) and (ii) for $m+1$ when $\alpha+\beta = 0$.

Now let us assume that $\alpha+\beta \geqslant 1$. Then one can very easily prove, by induction on $\alpha+\beta$, the following:

CLAIM: $D_\xi^\alpha D_x^\beta(W(Z,\xi))$ is a finite linear combination of terms of the form $[W_{i,j}(Z,\xi)]\cdot(\partial_\xi^{a_1}\partial_x^{b_1}Z)^{\lambda_1}\cdots(\partial_\xi^{a_p}\partial_x^{b_p}Z)^{\lambda_p}$, where

(a) $W_{i,j}(x,\xi) = (D_\xi^j D_x^i W)(x,\xi)$.

(b) $1 \leqslant i+j \leqslant \alpha+\beta$.

(c) the product $(\partial_\xi^{a_1}\partial_x^{b_1}Z)^{\lambda_1}\cdots(\partial_\xi^{a_p}\partial_x^{b_p}Z)^{\lambda_p}$ may or may not appear; if the product appears, then:

(d) $\lambda_1 + \cdots + \lambda_p \leqslant i$, $\quad \lambda_1, ..., \lambda_p \geqslant 1$;

(e) $1 \leqslant a_1 + b_1, ..., a_p + b_p \leqslant \alpha+\beta$;

(f) if $\beta \neq 0$, then $b_1 + b_2 + \cdots + b_p \geqslant 1$.

We sketch the proof of the CLAIM. For $\alpha+\beta = 1$ either $\alpha = 0, \beta = 1$ or $\alpha = 1, \beta = 0$ in which case the claim is easily verified. Assume the claim to be true for all α, β with $\alpha+\beta = \gamma$. Now $\gamma+1$ can be obtained as $(\alpha+1)+\beta$ or $\alpha+(\beta+1)$ where $\alpha+\beta = \gamma$. In either case, using induction hypothesis the CLAIM is verified.

Now let us consider a typical term as in the CLAIM. Using (4.12), (4.13), (4.14), and condition (vii) of Sec. 3, we have

$$\int_{t_0}^{t} ds |W_{i,j}(Z,\xi)(\partial_\xi^{a_1}\partial_x^{b_1}Z)^{\lambda_1}\cdots(\partial_\xi^{a_p}\partial_x^{b_p}Z)^{\lambda_p}| \leqslant K\int_{t_0}^{t} ds\,(1+|x|+s)^{-i-\delta}(1+s)^i \quad \text{if } \beta = 0$$

or

$$\leqslant K\int_{t_0}^{t} ds\,(1+|x|+s)^{-i-\delta}(1+s)^{i-1} \quad \text{if}\ \beta\neq 0$$

$$\leqslant K\int_{t_0}^{t} ds\,(1+s)^{-\delta} \text{ or } K\int_{t_0}^{t} ds(1+|x|+s)^{-1-\delta} \text{ as}$$

$\beta=0$ or $\beta\neq 0$, respectively,

$\leqslant K(1+t)^{1-\delta}$ or $K(1+|x|)^{-\delta}$ as $\beta=0$ or $\beta\neq 0$, respectively. (4.15)

Now the result follows from (4.11), (4.15), and the CLAIM.

(iii) Put $Z=x+(t-t_0)h_0'(\xi)+\partial_\xi Y(m_0,t_0,t,x,\xi)$. Then choose r_0 large so that, by using (4.8) and (i)

$$|Z|\geqslant K_0(1+|x|+t) \quad \text{for some } K_0>0. \tag{4.16}$$

Clearly,

$$|D_\xi^\alpha Z|\leqslant K_\alpha(1+t), \quad \alpha\neq 0 \ \text{by (i)}, \tag{4.17}$$

and

$$|D_\xi^\alpha D_x^\beta Z|\leqslant K_{\alpha\beta}, \quad \beta\neq 0, \ \text{by (ii)}. \tag{4.18}$$

Now the result follows as in (i) and (ii), by using CLAIM, (4.16), (4.17), and (4.18).

(iv) Put $Z=x+th_0'(\xi)+\partial_\xi Y(m_0,t_0,t,x,\xi)$,

$U(x,\xi)=-(\nabla_x W)(x,\xi)\cdot\nabla h_0(\xi)$,

so that, by condition (vii) of Sec. 3, for any compact set B of R^n

$$\sup_{\xi \text{ in } B}|D_\xi^\alpha D_x^\beta U(x,\xi)|\leqslant K(B,\alpha,\beta)(1+|x|)^{-|\beta|-1-\delta}. \tag{4.19}$$

Then,

$$D_\xi^\alpha D_x^\beta\{W(x+(t-t_0)h_0'(\xi)+\partial_\xi Y(m_0,t_0,t,x,\xi),\xi) - W(x+th_0'(\xi)+\partial_\xi Y(m_0,t_0,t,x,\xi),\xi)\}$$
$$=t_0\int_0^1 d\rho\, D_\xi^\alpha D_x^\beta U(Z-\rho t_0 h_0'(\xi),\xi). \tag{4.20}$$

By (4.8), (i), and (ii) we have

$$|Z-\rho t_0 h_0'(\xi)| \geqslant K_0(1+|x|+t-\rho t_0)-K(1+t)^{1-\delta}$$

for some $K_0>0$

$$\geqslant \tfrac{1}{2}K_0(1+|x|+t) \quad \text{if } r_0 \text{ is large}, \tag{4.21}$$

$$|D_\xi^\alpha\{Z-\rho t_0 h_0'(\xi)\}|\leqslant K_\alpha(1+t) \quad \text{for } \alpha\neq 0, \tag{4.22}$$

and

$$|D_\xi^\alpha D_x^\beta\{Z-\rho t_0 h_0'(\xi)\}|\leqslant K_{\alpha\beta} \quad \text{for } \beta\neq 0. \tag{4.23}$$

Now the result follows by (4.19), to (4.23) as in (i), (ii).

(v) We prove the result by induction on m. For $m=1$ the result follows from (i) and (ii). Now assume the result to be true for m with $1\leqslant m\leqslant m_0-1$.

$$Y(m,t_0,t,x,\xi)-Y(m-1,t_0,t,x,\xi) = \int_0^1 d\rho\int_{t_0}^{t} ds\,(\nabla_x W)(Z,\xi)\cdot U, \tag{4.24}$$

where

$$Z=Z(\rho,t_0,s,x,\xi) = x+sh_0'(\xi)+\rho\partial_\xi Y(m-1,t_0,s,x,\xi) + (1-\rho)\partial_\xi Y(m-2,t_0,s,x,\xi)$$

and

$$U=U(t_0,s,x,\xi) = \nabla_\xi Y(m-1,t_0,s,x,\xi)-\nabla_\xi Y(m-2,t_0,s,x,\xi).$$

Note that by induction hypothesis we have

$$|D_\xi^\alpha D_x^\beta U|\leqslant K_{\alpha\beta}(1+s)^{1-(m-1)\delta} \quad \text{for } s\geqslant t_0. \tag{4.25}$$

Also by (i) and (ii)

$$|D_\xi^\alpha D_x^\beta Z|\leqslant K_{\alpha\beta}(1+s) \quad \text{if } \beta=0,\ \alpha\neq 0, \tag{4.26}$$

$$\leqslant K_{\alpha\beta} \quad \text{if } \beta\neq 0. \tag{4.27}$$

Furthermore, by choosing t_{-1} large, if necessary, we have by (4.8)

$$|Z|\geqslant K_0(1+|x|+s) \quad \text{for } s\geqslant t_0, \text{ with } K_0>0. \tag{4.28}$$

Now the result follows from the techniques of the proof of (i) and (ii) by using (4.24)–(4.28) and condition (vii) of Sec. 3.

(vi) Follows from (v) by writing

$$W(x+th_0'(\xi)+\partial_\xi Y(m_0,t_0,t,x,\xi),\xi) - W(x+th_0'(\xi)+\partial_\xi Y(m_0-1,t_0,t,x,\xi),\xi)$$
$$=\int_0^1 d\rho\,\frac{d}{d\rho}W(x+th_0'(\xi)+\rho\partial_\xi Y(m_0,t_0,t,x,\xi) + (1-\rho)\partial_\xi Y(m_0-1,t_0,t,x,\xi),\xi)$$

and using the techniques of the proof of (v).

(vii) Follows from (vi) by taking $\alpha+\beta=0$, $x=0$.

(viii) $|\partial\{Y(m_0,t_0,t,x,\xi)-Y(m_0,t_0,t,0,\xi)\}/\partial t|$

$$=\Big|\int_0^1 d\rho(\nabla_x W)(\rho x+t\nabla h_0(\xi) + \nabla_\xi Y(m_0-1,t_0,t,\rho x,\xi),\xi) \cdot\{x+(\nabla_x\nabla_\xi Y)(m_0-1,t_0,t,\rho x,\xi)\}\Big|,$$

$$\leqslant\int_0^1 d\rho(1+|\rho x+t\nabla h_0(\xi) + \nabla_\xi Y(m_0-1,t_0,t,\rho x,\xi)|)^{-1-\delta}(|x|+K)$$

by (ii) and condition (vii) of Sec. 3,

$$\leqslant(1+|t|)^{-1-\delta}(|x|+K) \quad \text{by (i) and (4.8)}. \tag{4.29}$$

The result follows from (4.29) by noting $\delta>0$.

(ix) Clearly for τ very large we get by (i) and (4.6)

$|\tau h_0'(\xi)+\partial_\xi Y(m_0-1,t_0,\tau,0,\xi)|\geqslant 2a\tau$ with $a>0$ so that

$$|W(\tau h_0'(\xi)+\partial_\xi Y(m_0-1,t_0,\tau,0,\xi),\xi)| \leqslant K(1+\tau)^{-\delta}. \tag{4.30}$$

By (4.30) we get

$$|Y(m_0,t_0,t+s,0,\xi)-Y(m_0,t_0,t,0,\xi)| \leqslant K\int_t^{t+s} d\tau\,(1+\tau)^{-\delta}.$$

The result follows by noting $\lim_{t\to\infty}\int_t^{t+s} d\tau\,(1+\tau)^{-\delta}=0$. Q.E.D.

We define the modification for the free evolution to be $Y(m_0, t, \xi)$, where

$$Y(m, t, \xi) = Y(m, 0, t, 0, \xi) \quad \text{for } t \geqslant 0. \tag{4.31}$$

Then we have

Lemma 4.2: (i) For C and t_0 as in Lemma 4.1 and $0 \leqslant m \leqslant m_0$, we get

(a) $\sup\{|D_\xi^\alpha\{Y(m, t, \xi) - Y(m, t_0, t, 0, \xi)\}|: t \geqslant t_0, \xi \in C_{2c}\} \leqslant K_\alpha$,

(b) $\lim_{t\to\infty} Y(m, t, \xi) - Y(m, t_0, t, 0, \xi)$ exists for ξ in C_{2c}.

(ii) $\lim_{t\to\infty} Y(m_0, t, \xi) - Y(m_0, t_0, t, 0, \xi)$ exists for ξ in C_{2c}.

(iii) $\lim_{t\to\infty} Y(m_0, t+s, \xi) - Y(m_0, t, \xi) = 0$ for all s and all ξ in G given by (4.1).

Proof: (i) We prove by induction on m. For $m = 0$ it is easy to verify. Assume the result to be true for m with $0 \leqslant m \leqslant m_0$. For $t \geqslant t_0$ define $\psi(m, t, \xi)$ by

$$\psi(m, t, \xi) = Y(m, t, \xi) - Y(m, t_0, t, 0, \xi). \tag{4.32}$$

Then by induction hypothesis

$$\sup\{|D_\xi^\alpha \psi(m, t, \xi)|: t \geqslant t_0, \xi \in C_{2c}\} \leqslant K_\alpha < \infty, \tag{4.33}$$

and

$$\lim_{t\to\infty} \psi(m, t, \xi) \text{ exists for } \xi \text{ in } C_{2c}. \tag{4.34}$$

Then by (4.32) we have for $t \geqslant t_0$

$$\begin{aligned}\psi(m+1, t, \xi) &= \int_0^{t_0} ds\, W(sh_0'(\xi) + \partial_\xi Y(m, s, \xi), \xi) \\ &\quad + \int_{t_0}^t ds \int_0^1 d\rho\, (\nabla_x W)(sh_0'(\xi) + \rho\partial_\xi \psi(m, s, \xi) \\ &\quad + \partial_\xi Y(m, t_0, s, 0, \xi), \xi)\nabla_\xi \psi(m, s, \xi).\end{aligned}$$

Now the result follows by (4.32), (4.33), (4.34) and using the technique of the proof of Lemma 4.1(i).

(ii) Follows from (i)(b) by taking $m = m_0$.

(iii) Let C be as in Lemma 4.1. Then by Lemma 4.1(ix) and Lemma 4.2(ii), we get for ξ in C

$$\lim_{t\to\infty} \{Y(m_0, t+s, \xi) - Y(m_0, t, \xi)\} = 0.$$

The result follows by noting G is a union of such sets C. Q.E.D.

Remark 4.3: We can define for $t \leqslant t_0 \leqslant 0$ the function Y by (4.9), (4.10) and prove results similar to (a) Lemma 4.1 when (4.4) is replaced by $x\cdot h_0'(k) \leqslant 0$ and (b) Lemma 4.2.

5. EXISTENCE OF THE WAVE OPERATOR

This section closely follows Ref. 16. Let $Y(m_0, t_0, t, 0, \xi)$ and $Y(m_0, t, \xi)$ be as in (4.10) and (4.31), respectively. Define $X(t, \xi)$, $X(t_0, t, \xi)$ by

$$X(t, \xi) = th_0(\xi) + Y(m_0, t, \xi) \quad \forall \xi \in G \text{ of (4.1)}, \tag{5.1}$$

$$X(t_0, t, \xi) = th_0(\xi) + Y(m_0, t_0, t, 0, \xi) \quad \forall \xi \in G,\ t \geqslant t_0. \tag{5.2}$$

Let $H_0 = h_0(P)$, H be the free and total Hamiltonians given by conditions (i)–(iv), statement of Lemma 3.1, (vii), (ix), and (x) of Sec. 3. Define the free, modified free, total unitary evolutions U_t, Z_t, V_t, and for $t \geqslant t_0$ a unitary evolution $Z(t_0, t)$ by

$$\begin{aligned} U_t &= \exp(-itH_0), \quad Z_t = \exp[-iX(t, P)], \\ V_t &= \exp(-itH), \quad Z(t_0, t) = \exp[-iX(t_0, t, P)]. \end{aligned} \tag{5.3}$$

The aim of this section is to show

(i) $\Omega_+ = \text{s-}\lim_{t\to\infty} V_t^* Z_t$ exist,

(ii) $V_t\Omega_+ = \Omega_+ U_t\ \forall t$,

(iii) Range $\Omega_+ \subset \mathscr{H}_{ac}(H)$, the absolutely continuous space for H.

Lemma 5.1: Let the compact set C and $c > 0$ be as in (4.3) and further assume that $\xi \to \nabla h_0(\xi)$ is a diffeomorphism on a neighborhood of C_{3c}. Let $\phi \in C_0^\infty(G)$ be such that

$$\operatorname{supp} \phi \subset \{\xi \in R^n: \operatorname{dist}(\xi, C) < c\}. \tag{5.4}$$

Let $f \in \mathscr{S}(R^n)$. Then

(i) $\int_{t_0}^\infty dt\, \|(1 + |Q|)^{-1-\epsilon_0} Z(t_0, t)\phi(P)f\| < \infty$.

(ii) $\int_{t_0}^\infty dt\, \|W_S(Q, P)Z(t_0, t)\phi(P)f\| < \infty$.

(iii) $\int_{t_0}^\infty dt\, \|\{W(th_0'(P) + \partial Y(m_0 - 1, t_0, t, 0, P)/\partial P, P) - W(\partial X(t_0, t, P)/\partial P, P)\}\phi(P)\| < \infty$.

(iv) $\int_{t_0}^\infty dt\, \|\{W(Q, P) - W(\partial X(t_0, t, P)/\partial P, P)\} \times Z(t_0, t)\phi(P)f\| < \infty$.

(v) $\Omega_+^\phi = \text{s-}\lim_{t\to\infty} V_t^* Z(t_0, t)\phi(P)$ exists.

Proof: (*i*) Let $a > 0$ be as in (4.6). Then (i) follows if we can show

$$\int_{t_0}^\infty dt\, \|F(|Q| \leqslant at)Z(t_0, t)\phi(P)f\| < \infty.$$

Clearly,

$$(Z(t_0, t)\phi(P)f)(q) = \int d\xi\, \phi(\xi)\hat{f}(\xi)\exp(i[q\cdot\xi - X(t_0, t, \xi)]). \tag{5.5}$$

For ξ in C_{2c} and $|q| \leqslant at$, we have

$$\begin{aligned} |\nabla_\xi\{q\cdot\xi - X(t_0, t, \xi)\}| &\geqslant t|\nabla h_0(\xi)| - |q| - |\nabla_\xi Y(m_0, t_0, t, 0, \xi)| \\ &\geqslant 2at - K(1 + t)^{1-\delta} \quad \text{by Lemma 4.1(ii) and (4.6)} \\ &\geqslant \tfrac{3}{2}at \quad \text{for } t \geqslant t_0 \text{ with } t_0 \text{ large;} \end{aligned} \tag{5.6}$$

also by Lemma 4.1(i)

$$|D_\xi^\alpha\{q\cdot\xi - X(t_0, t, \xi)\}| \leqslant Kt \quad \text{for } |\alpha| \geqslant 1. \tag{5.7}$$

Now apply the stationary phase Lemma A.I of Ref. 16 or Lemma I of the Appendix of Ref. 12 to (5.5) using (5.6) and (5.7) to get

$$|F(|q|\leqslant at)\{Z(t_0,t)\phi(P)f\}(q)|$$
$$\leqslant K_M(1+t)^{-M}F(|q|\leqslant at)\sum_{|\alpha|\leqslant M}\|D_\xi^\alpha\{\phi(\xi)\hat f(\xi)\}\|_\infty$$
$$\leqslant K_M(f)(1+t)^{-M}F(|q|\leqslant at). \tag{5.8}$$

Choosing $M=[\frac{1}{2}n]+3$ in (5.8), we get

$$\int_{t_0}^\infty dt\,\|F(|Q|\leqslant at)Z(t_0,t)\phi(P)f\|<\infty$$

and so the result follows.

(ii) We deduce from (i) and Lemma 3.1. Put $\psi(\xi)=\phi(\xi)(1+\xi^2)^{N/2}$ with N as in Lemma 3.1. Then note that $\psi\in C_0^\infty(G)$ and supp $\phi=$ supp ψ so that as in (i) we have

$$\int_{t_0}^\infty dt\,\|(1+|Q|)^{-1-\epsilon_0}\psi(P)Z(t_0,t)f\|<\infty. \tag{5.9}$$

By Lemma 3.1 we have

$$\|W_S(Q,P)\phi(P)Z(t_0,t)f\|$$
$$=\|W_S(Q,P)(1+P^2)^{-N/2}(1+|Q|)^{1+\epsilon_0}$$
$$\times(1+|Q|)^{-1-\epsilon_0}\psi(P)Z(t_0,t)f\|$$
$$\leqslant K\|(1+|Q|)^{-1-\epsilon_0}\psi(P)Z(t_0,t)f\|\quad\text{for } K<\infty. \tag{5.10}$$

The result follows from (5.9) and (5.10).

(iii) By Lemma 4.1(vii) we get

$$\|\{W(t\nabla h_0(P)+\partial Y(m_0-1,t_0,t,0,P)/\partial P,P)$$
$$-W(\partial X(t_0,t,P)/\partial P,P)\}\phi(P)\|$$
$$\leqslant K(1+t)^{-(m_0+1)\delta}.$$

Now the result follows since $(m_0+1)\delta>1$ by (4.2).

(iv) *Step 1*: (integration near t_0). By Lemma 3.1(ii) and condition (vii) of Sec. 3 we have

$$\sup\{\|W(Q,P)\phi(P)\|+\|W(\partial X(t_0,t,P)/\partial P,P)$$
$$\phi(P)\|:t\geqslant t_0\}<\infty.$$

So we very easily get for every finite $t_1\geqslant t_0$

$$\int_{t_0}^{t_1}dt\,\|\{W(Q,P)-W(\partial X(t_0,t,P)/\partial P,P)\}$$
$$\times Z(t_0,t)\phi(P)f\|<\infty.$$

So from now on we assume that $t\geqslant t_1\geqslant t_0$, where t_1 is large.

Step 2: (qt^{-1} outside an annulus): Define $g(t,q)$ by

$$g(t,q)=[\{W(Q,P)-W(\partial X(t_0,t,P)/\partial P,P)\}$$
$$\times\phi(P)Z(t_0,t)f](q) \tag{5.11}$$

so that by (3.4)

$$g(t,q)=\int d\xi\,\{W(q,\xi)-W(\partial_\xi X(t_0,t,\xi),\xi)\}$$
$$\times\phi(\xi)\hat f(\xi)\exp[i\{q\cdot\xi-X(t_0,t,\xi)\}].$$

Let

$$a=\inf\{t^{-1}|X'_\xi(t_0,t,\xi)|:\xi\in C_c,\ t\geqslant t_1\},$$
$$b=\sup\{t^{-1}|X'_\xi(t_0,t,\xi)|:\xi\in C_c,\ t\geqslant t_1\}$$

so that $0<a<b<\infty$.

By Lemma 4.1(i) it is easily seen that, for $|q|\notin[\frac{1}{2}at,2bt]$ and ξ in C_c,

$$|\nabla_\xi\{q\cdot\xi-X(t_0,t,\xi)\}|\geqslant K^*(1+|q|+t)\quad\text{with } K^*>0, \tag{5.12}$$
$$|D_\xi^\alpha\{q\cdot\xi-X(t_0,t,\xi)\}|\leqslant K_\alpha(1+|q|+t)\quad\text{for } |\alpha|\geqslant 1. \tag{5.13}$$

Now apply the stationary phase Lemma A.1 of Ref. 16 or Lemma I of the Appendix of Ref. 12 to the integral expression for $g(t,q)$, using (5.12) and (5.13), to get

$$|g(t,q)F(|q|\notin[\tfrac{1}{2}at,2bt])|$$
$$\leqslant K_MF(|q|\notin[\tfrac{1}{2}at,2bt])(1+|q|+t)^{-M}$$
$$\times\Big\{\sum_{|\alpha|\leqslant M}\|D_\xi^\alpha\{W(q,\xi)\phi(\xi)\hat f(\xi)\}\|_\infty$$
$$+\sum_{|\alpha|\leqslant M}\|D_\xi^\alpha\{W(X'_\xi(t_0,t,\xi),\xi)\phi(\xi)\hat f(\xi)\}\|_\infty\Big\}$$
$$\leqslant K_M(f)(1+|q|+t)^{-M}.$$

Choosing $M=[\frac{1}{2}n]+3$, we see that

$$\int_{t_1}^\infty dt\,\|g(t,\cdot)F(|\cdot|\notin[\tfrac{1}{2}at,2bt])\|<\infty. \tag{5.14}$$

Step 3: (qt^{-1} is inside the annulus but away from critical points): Let $\chi_0\in C_0^\infty(R^n)$ be such that $0\leqslant\chi_0\leqslant 1$, $\chi_0=1$ on $|x|\leqslant 1$, 0 on $|x|\geqslant 2$. Then, for any $L>0$, σ in $(0,\frac{1}{2})$, we get

$$1=(1-\chi_0)(t^{\sigma-1}L^{-1}[q-X'_\xi(t_0,t,\xi)])$$
$$+\chi_0(t^{\sigma-1}L^{-1}[q-X'_\xi(t_0,t,\xi)]). \tag{5.15}$$

Then

$$g(t,q)=g_1(t,q)+g_2(t,q), \tag{5.16}$$
$$g_1(t,q)=\int d\xi\,[W(q,\xi)-W(X'_\xi(t_0,t,\xi),\xi)]\phi(\xi)\hat f(\xi)$$
$$\times(1-\chi_0)(t^{\sigma-1}L^{-1}[q-X'_\xi(t_0,t,\xi)])$$
$$\times\exp(i[q\cdot\xi-X(t_0,t,\xi)]), \tag{5.17}$$
$$g_2(t,q)=\int_0^1 d\rho\int d\xi\,(\nabla_x W)(\rho q+(1-\rho)X'_\xi(t_0,t,\xi),\xi)$$
$$\cdot[q-\nabla_\xi X(t_0,t,\xi)]\phi(\xi)\hat f(\xi)$$
$$\times\chi_0(t^{\sigma-1}L^{-1}[q-X'_\xi(t_0,t,\xi)])$$
$$\times\exp(i[q\cdot\xi-X(t_0,t,\xi)])$$
$$=\int_0^1 d\rho\,g_2(\rho,t,q). \tag{5.18}$$

As in Step 2, we have

$$|g_1(t,q)F(|q|\in[\tfrac{1}{2}at,2bt])|$$
$$\leqslant K_M(f)F(|q|\in[\tfrac{1}{2}at,2bt])(1+Lt^{1-\sigma})^{-M}(1+t^\sigma)^M$$
$$\leqslant K_M(f)t^{-M(1-2\sigma)}F(|q|\in[\tfrac{1}{2}at,2bt]).$$

Now choosing M large so that $-M(1-2\sigma)+\frac{1}{2}n+1<0$, we easily see that

$$\int_{t_1}^\infty dt\,\|g_1(t,\cdot)F(|\cdot|\in[\tfrac{1}{2}at,2bt])\|<\infty. \tag{5.19}$$

Step 4 (q is nearer the critical points): As in Ref. 16, choose ξ_0 in C_c and a diffeomorphism ψ defined in a neighborhood of 0 such that for $|q|$ in $[\frac{1}{2}at,2bt]$

$$q=X'_\xi(t_0,t,\xi_0),$$
$$\psi(0)=\xi_0,$$

$$q\cdot\psi(y) - X(t_0, t, \psi(y)) = q\xi_0 - X(t_0, t, \xi_0) + t\langle Ay, y\rangle,$$
$$A = X''_\xi(t_0, t, \xi_0)/2t. \tag{5.20}$$

Note that ξ_0 depends on t, q; A depends on t, q but under our assumptions it varies in a compact subset of $\mathrm{GL}(n,R)$; ψ also depends on t, q. Further by Lemma A.6 of Ref. 16, we have

$$|(D_y^\alpha\psi)(0)| \leqslant K_\alpha \quad \text{for all } \alpha. \tag{5.21}$$

Now, applying Lemmas A.4 and A.2 of Ref. 16 to (5.18), we get, for every $M \geqslant 2$,

$$\begin{aligned}
&|g_2(\rho, t, q)F(|q|\in[\tfrac{1}{2}at, 2bt])|\\
&\leqslant \sum_{j=1}^{M-1} K_j t^{-j-n/2} |\langle A^{-1}D, D\rangle^j\\
&\quad\times\{(\nabla_x W)(\rho q + (1-\rho)X'_\xi(t_0, t, \psi(y)), \psi(y))\\
&\quad\cdot[q - X'_\xi(t_0, t, \psi(y))]\phi(\psi(y))\hat{f}(\psi(y))|\det\psi'(y)|\\
&\quad\times\chi_0(t^{\sigma-1}L^{-1}[q - X'_\xi(t_0, t, \psi(y))])\}|\\
&\quad + K_M t^{-M-n/2}\sum_{|\alpha|\leqslant s}\|D_\xi^\alpha\\
&\quad\{(\nabla_x W)(\rho q + (1-\rho)X'_\xi(t_0, t, \xi), \xi)\\
&\quad\times[q - X'_\xi(t_0, t, \xi)]\\
&\quad\times\phi(\xi)\hat{f}(\xi)\chi_0(t^{\sigma-1}L^{-1}[q - X'_\xi(t_0, t, \xi)])\}\|_\infty,
\end{aligned}$$

where $s > 2M + \frac{1}{2}n$ and D is the differential operator w.r.t. y at $y = 0$.

By Lemma 4.1 (i), condition (vii) of Sec. 3, and (5.21), we see that

$$\begin{aligned}
&|g_2(\rho, t, q)F(|q|\in[\tfrac{1}{2}at, 2bt])|\\
&\leqslant\Big\{K_j\sum_{j=1}^{M-1} t^{-j-n/2-1-\delta+1-\sigma+2j\sigma}\\
&\quad + K_M\sum_{|\alpha|\leqslant s} t^{-M-n/2-1-\delta+(1-\sigma)+s\sigma}\Big\}\\
&\quad\times F(|q|\in[\tfrac{1}{2}at, 2bt])\\
&\leqslant\{K_M^* t^{-1-n/2-\delta+\sigma} + K_M t^{-M-n/2-\delta+(s-1)\sigma}\}\\
&\quad\times F(|q|\in[\tfrac{1}{2}at, 2bt]) \quad \text{since } \sigma < \tfrac{1}{2}.
\end{aligned}$$

Now it is easily seen that if σ is in $(0, \min\{\delta, \frac{1}{2}\})$ and $-M - \delta + (2M + n - 1)\sigma < -1$, then

$$\int_{t_1}^\infty dt\,\|g_2(t, \cdot)F(|\cdot|\in[\tfrac{1}{2}at, 2bt])\| < \infty. \tag{5.22}$$

By step 1, (5.11), (5.14), (5.16), (5.19), and (5.22) the result follows.

(v) By using (ii)–(iv) we see that for f in $\mathscr{S}(R^n)$

$$\int_{t_0}^\infty dt\left|\left|\frac{d}{dt}\{V_t^* Z(t_0, t)\phi(P)f\right|\right| < \infty$$

so that $\text{s-lim}_{t\to\infty} V_t^* Z(t_0, t)\phi(P)f$ exists. The result follows since $\mathscr{S}(R^n)$ is dense in $L^2(R^n)$.

Theorem 5.2: (i) $\Omega_+ = \text{s-lim}_{t\to\infty} V_t^* Z_t$ exists,
(ii) Ω_+ is an isometry and $V_t\Omega_+ = \Omega_+ U_t$ for all t,
(iii) Range $\Omega_+ \subset \mathscr{H}_{\text{ac}}(H)$.

Proof: (i) For ϕ as in Lemma 5.1, using Lemma 4.2 (ii) we get the existence of $\text{s-lim}_{t\to\infty} V_t^* Z_t \phi(P)$. The result is clear since $\cup\{\text{Range }\phi(P): \phi, C \text{ as in Lemma 5.1}\}$ is dense in $L^2(R^n)$ by condition (iv) of Sec. 3.

(ii) It is clear that Ω_+ is an isometry. As in Ref. 16, using Lemma 4.2 (iii), we get $V_t\Omega_+ = \Omega_+ U_t$.

(iii) Follows from (ii) and the proof of Proposition 1, XI.3 of Ref. 4. Q.E.D.

Remark 5.3: By the remark 4.3 we can define Z_t for $t \leqslant 0$. Then we have Ω_- satisfying Theorem 5.2 for Ω_-. Note that we have proved Theorem 1(b).

6. APPROXIMATING THE TOTAL EVOLUTION BY AN AUXILIARY POSITION-MOMENTUM-DEPENDENT EVOLUTION ON A HALF-SPACE

Let the compact set C of Lemma 5.1 be a closed sphere, i.e., $\{x\in R^n: |x - x_0| \leqslant r\}$ for some $x_0 \in R^n$ and $r > 0$, let $Y(m_0, t_0, t, x, \xi)$ be as in (4.10), and let (4.3) to (4.7) be satisfied. The operators H and H_0 are as in Sec. 5. Define for $t \geqslant t_0$

$$X(t_0, t, x, \xi) = x\cdot\xi + (t - t_0)h_0(\xi) + Y(m_0, t_0, t, x, \xi). \tag{6.1}$$

Let T be as in Sec. 2 with the constant c^* of (2.1) less than or equal to the constant c of (4.3). Define the sets $E(C, \pm, r)$ by

$$\begin{aligned}
E(C, \pm, r) = \{&(x, k)\in R^n\times R^n: k\in C,\\
&x\cdot h'_0(k) \lesseqgtr 0,\ |x| \geqslant r\}.
\end{aligned}\tag{6.2}$$

For ϕ in $C_0^\infty(G)$ [with G as in (4.1)], define the integral kernels $I_1,\ldots,I_5$ of (q, x) depending on t_0, t, ϕ, k by

$$\begin{aligned}
&I_1(t_0, t, \phi, k, q, x)\\
&\quad = \int d\xi\,\phi(\xi)\hat{\eta}(\xi - k)\exp(i[q\cdot\xi - X(t_0, t, x, \xi)]),\\
&I_2(t_0, t, \phi, k, q, x)\\
&\quad = \int d\xi\,\phi(\xi)\hat{\eta}(\xi - k)\{\exp(i[q\cdot\xi - X(t_0, t, x, \xi)])\}.\\
&\qquad\times\{W(X'_\xi(t_0, t, x, \xi), \xi) - W(x + th'_0(\xi)\\
&\qquad + \partial_\xi Y(m_0 - 1, t_0, t, x, \xi), \xi)\},\\
&I_3(t_0, t, \phi, k, q, x)\\
&\quad = \int d\xi\,\phi(\xi)\hat{\eta}(\xi - k)\{W(q, \xi)\\
&\qquad - W(X'_\xi(t_0, t, x, \xi), \xi)\}\\
&\qquad\times\exp(i[q\cdot\xi - X(t_0, t, x, \xi)]),\\
&I_4(t_0, t, \phi, k, q, x)\\
&\quad = \int d\xi\,\phi(\xi)\hat{\eta}(\xi - k)W(q, \xi)\\
&\qquad\times\exp(i[q\cdot\xi - X(t_0, t, x, \xi)]),\\
&I_5(t_0, t, \phi, k, q, x)\\
&\quad = \int d\xi\,\phi(\xi)\hat{\eta}(\xi - k)W(X'_\xi(t_0, t, x, \xi), \xi)\\
&\qquad\times\exp(i[q\cdot\xi - X(t_0, t, x, \xi)])
\end{aligned}\tag{6.3}$$

so that $I_3 = I_4 - I_5$.

Define the operator $E_1(t_0, t, C, \phi, +, r)$ by

$\{E_1(t_0, t, C, \phi, +, r)f\}(q)$

$$= \int_{E(C,+,r)} dx\, dk\, \langle \eta_{xk} | f \rangle I_1(t_0, t, \phi, k, q, x). \tag{6.4}$$

Note that

$$\{\phi(P) T(E(C, +, r)) f\}(q)$$
$$= \int_{E(C,+,r)} dx\, dk\, \langle \eta_{xk} | f \rangle \int d\xi\, \phi(\xi) \hat{\eta}(\xi - k) \times \exp(i[q\cdot\xi - x\xi])$$

so that

$$E_1(t_0, t_0, C, \phi, +, r) = \phi(P) T(E(C, +, r)). \tag{6.5}$$

The operators $E_1(t_0, t, C, \phi, +, r)$ and $T(E(C, +, r))$ can be compared with, respectively, $E_{+,r}(t, s)$, $P_{+,r}$ of Ref. 30. We wish to prove statements similar to Theorem 4.4 and Propositions 4.5 and 4.6 of Ref. 30. To this end and let us note that, by (4.10),

$$\left[-iV_{t-t_0} \frac{d}{dt} \{ V^*_{t-t_0} E_1(t_0, t, C, \phi, +, r) f \} \right](q)$$
$$= \{ W_S(Q, P) E_1(t_0, t, C, \phi, +, r) f \}(q) + \int_{E(C,+,r)} dx\, dk\, \langle \eta_{xk} | f \rangle (I_2 + I_3)(t_0, t, \phi, k, q, x), \tag{6.6}$$
$$= \{ [W_S(Q, P) + W(Q, P)] E_1(t_0, t, C, \phi, +, r) f \}(q) + \int_{E(C,+,r)} dx\, dk\, \langle \eta_{xk} | f \rangle (I_2 - I_5)(t_0, t, \phi, k, q, x). \tag{6.7}$$

Define integral kernels $J_1, ..., J_5$ by

$$J_j(t_0, t, \phi, r, k, q, x) = F(x\cdot h_0'(k) \geqslant 0, |x| \geqslant r) I_j(t_0, t, \phi, k, q, x). \tag{6.8}$$

For any kernel $J = J(q, x)$ let $\|J(.,.)\|$ denote operator norm of the operator on $L^2(R^n)$ induced by the kernel J.

The motivation for the following Lemma 6.1 whose proof has a striking similarity to the proof of Lemma 5.1 is clear from (6.6)–(6.8).

Lemma 6.1: Let T, ϕ, C be as above. Then:

(i) $\displaystyle \lim_{r\to\infty} \int_{t_0}^{\infty} dt\, \|(1 + |Q|)^{-1-\epsilon_0} E_1(t_0, t, C, \phi, +, r)\| = 0;$

(ii) $\displaystyle \lim_{r\to\infty} \int_{t_0}^{\infty} dt\, \| W_S(Q, P) E_1(t_0, t, C, \phi, +, r)\| = 0;$

(iii) $\displaystyle \lim_{r\to\infty} \int_{t_0}^{\infty} dt \sup_{k\in C} \|J_2(t_0, t, \phi, r, k, \cdot,)\| = 0.$

For every finite $t_1 \geqslant t_0$:

(iv) $\displaystyle \lim_{r\to\infty} \int_{t_0}^{t_1} dt\, \| W(Q, P) E_1(t_0, t, C, \phi, +, r)\| = 0;$

(v) $\displaystyle \lim_{r\to\infty} \int_{t_0}^{t_1} dt \sup_{k\in C} \|J_5(t_0, t, \phi, r, k, \cdot, \cdot)\| = 0;$

(vi) $\displaystyle \lim_{r\to\infty} \int_{t_0}^{t_1} dt \left|\left| \frac{d}{dt} \{ V^*_{t-t_0} E_1(t_0, t, C, \phi, +, r) \right|\right| = 0.$

If further diameter of C_{3c} is small, then:

(vii) $\displaystyle \lim_{r\to\infty} \int_{t_1}^{\infty} dt \sup_{k\in C} \|J_3(t_0, t, \phi, r, k, \cdot, \cdot)\| = 0$ for some $t_1 \geqslant t_0$;

(viii) $\displaystyle \lim_{r\to\infty} \int_{t_0}^{\infty} dt \left|\left| \frac{d}{dt} \{ V^*_{t-t_0} E_1(t_0, t, C, \phi, +, r)\} \right|\right| = 0;$

(ix) $\displaystyle \lim_{r\to\infty} \sup_{t>t_0} \| V_{t-t_0} \phi(P) T(E(C, +, r)) - E_1(t_0, t, C, \phi, +, r)\| = 0;$

(x) $\displaystyle \lim_{S\to\infty} S^{-1} \int_0^S dt\, \| T(E(C, +, 0)) \phi(P) V_t^* f\| = 0$ for each f in $\mathscr{H}_c(H)$.

Proof (i) Define $J_1^0(t_0, t, \phi, r, k, q, x) = (1 + |q|)^{-1-\epsilon_0} J_1(t_0, t, \phi, r, k, q, x)$. By (6.4) and Lemma 2.4 we need only to show that

$$\lim_{r\to\infty} \int_{t_0}^{\infty} dt \sup_{k\in C} \|J_1^0(t_0, t, \phi, r, k, \cdot, \cdot)\| = 0,$$

which we do in three steps.

Step 1: We show that $\sup_{t>t_0} \|E_1(t_0, t, C, \phi, +, r)\| < \infty$ for large r.

By Lemma 4.1 (ii) we get for $x\cdot h_0'(k) \geqslant 0$, k in C, $|\xi - k| \leqslant \frac{1}{8} c^*$, and $|x| \geqslant r$, r large

$$|\nabla_\xi \nabla_x X(t_0, t, x, \xi) - I| \leqslant \tfrac{1}{2}$$

and

$$|D_\xi^\alpha \nabla_x X(t_0, t, x, \xi)| \leqslant K_\alpha \quad \text{for } |\alpha| \geqslant 1.$$

By Theorem 2.5,

$$\sup_{t>t_0} \sup_{k\in C} \|J_1(t_0, t, \phi, r, k, \cdot, \cdot)\| < \infty. \tag{6.9}$$

Now Step 1 follows by Lemma 2.4.

Step 2 (q is near the origin): When $x\cdot h_0'(k) \geqslant 0$, $|\xi - k| \leqslant \frac{1}{8} c^*$, and $|x| \geqslant r$, r large, we get by Lemma 4.1(i) and (4.8)

$$|\nabla_\xi X(t_0, t, x, \xi)| \geqslant 2K^*(|x| + t - t_0) \quad \text{for some } K^* > 0.$$

If, further, $|q| \leqslant K^*(|x| + t - t_0)$, then

$$|\nabla_\xi \{q\cdot\xi - X(t_0, t, x, \xi)\}| \geqslant K^*(|x| + t - t_0),$$
$$|D_\xi^\alpha \{q\cdot\xi - X(t_0, t, x, \xi)\}| \leqslant K_\alpha(|x| + t - t_0) \quad \text{for } |\alpha| \geqslant 1.$$

Now an application of the stationary phase Lemma A.1 of Ref. 16 or Lemma I of the Appendix of Ref. 12 gives, for each $M \geqslant 1$, there exists K_M such that

$$|F(|q| \leqslant K^*(|x| + t - t_0)) J_1^0(t_0, t, \phi, r, k, q, x)| \leqslant K_M(|x| + t - t_0)^{-M} F(|q| \leqslant K^*(|x| + t - t_0)).$$

By taking the Hilbert–Schmidt norm we get, for $M \geqslant n + 2$

$$\left[\int\!\!\int dq\, dx\, \{ F(|q| \leqslant K^*(|x| + t - t_0)) \times |J_1^0(t_0, t, \phi, r, k, q, x)| \}^2 |^{1/2} \right] \leqslant K(r + t - t_0)^{-2} \quad \text{for all } k \text{ in } C. \tag{6.10}$$

Step 3 (q is away from the origin): Clearly for $|x| \geqslant r$

$$F(|q| \geqslant K^*(|x| + t - t_0)) \leqslant F(|q| \geqslant K^*(r + t - t_0))$$

so that, using (6.9), we have

operator norm of the kernel

$$(1+|q|)^{-\theta}F(|q|)\geqslant K^*(|x|+t-t_0)J_1(\cdots)$$
$$\leqslant K(r+t-t_0)^{-\theta} \quad \text{for all } k \text{ in } C,\ \theta\geqslant 0. \tag{6.11}$$

By (6.10) and (6.11) we get the operator norm of the kernel $(1+|q|)^{-\theta}J_1(t_0,t,\phi,r,k,q,x)\leqslant K(r+t-t_0)^{-\theta}$ for θ in [0,2], k in C. Now by Lemma 2.4 we get

$$\|(1+|Q|)^{-\theta}E_1(t_0,t,C,\phi,+,r)\|$$
$$\leqslant K(r+t-t_0)^{-\theta} \quad \text{for } 0\leqslant\theta\leqslant 2. \tag{6.12}$$

The result follows by taking $\theta=1+\epsilon_0$.

(ii) We deduce from (i) and Lemma 3.1(i). Choose ϕ_1 in $C_0^\infty(G)$ such that $\phi_1\phi=\phi$. Then by (6.4) and (6.3) we have

$$\phi_1(P)E_1(t_0,t,C,\phi,+,r)=E_1(t_0,t,C,\varphi,\phi,+,r)$$
$$=E_1(t_0,t,C,\phi,+,r). \tag{6.13}$$

It is easy to see by commutation rules that for any ψ in $\mathscr{S}(R^n)$ the operator $(1+|Q|)^{-2}\psi(P)(1+|Q|)^2$ is bounded and so, by interpolation, $(1+|Q|)^{-\theta}\psi(P)(1+|Q|)^\theta$ is bounded for θ in [0,2]. Now from Lemma 3.1(i) we see that

$$\|W_S(Q,P)\phi_1(P)(1+|Q|)^{1+\epsilon_0}\|<\infty. \tag{6.14}$$

By (6.13) and (6.14)

$$\|W_S(Q,P)E_1(t_0,t,C,\phi,+,r)\|$$
$$\leqslant K\|(1+|Q|)^{-1-\epsilon_0}E_1(t_0,t,C,\phi,+,r)\|$$

and now the result follows from (i).

(iii) In Step 1 of the proof of (i) we have for $x\cdot h_0'(k)\geqslant 0$, k in C, $|\xi-k|\leqslant\frac{1}{8}c^*$, $|x|\geqslant r$, r large

$$|\nabla_\xi\nabla_x X(t_0,t,x,\xi)-I|\leqslant\tfrac{1}{2}$$

and

$$|D_\xi^\alpha\nabla_x X(t_0,t,x,\xi)|\leqslant K_\alpha \quad \text{for } |\alpha|\geqslant 1.$$

Now by Theorem 2.5, Lemma 4.1(iv), (vi)

$$\sup_{k\in C}\|J_2(t_0,t,\phi,r,k,\cdot,\cdot)\|$$
$$\leqslant K\,[t_0(1+r+t)^{-1-\delta}$$
$$+(1+t)^{1-m_0\delta}(1+t+r)^{-1-\delta}].$$

The result follows since $(m_0+1)\delta>1$ by (4.2).

(iv) Put $\theta=\delta$ in (6.12) to get

$$\|(1+|Q|)^{-\delta}E_1(t_0,t,C,\phi,+,r)\|\leqslant K(r+t-t_0)^{-\delta}.$$

Now the proof is similar to the proof of (ii) by using Lemma 3.1(ii).

(v) As in (iii) using Theorem 2.5, Lemma 4.1(iii)

$$\sup_{k\in C}\|J_5(t_0,t,\phi,r,k,\cdot)\|\leqslant K(r+t-t_0)^{-\delta}.$$

Now the result is immediate.

(vi) Follows from (6.7), (ii), (iii), (iv), (v), and Lemma 2.4.

(vii) This is the central part of this publication. The proof is lengthy and so we split it into seven steps. We always take t_1 large such that t_1-t_0 is large.

Step 1 (q near the origin): Using the condition (vii) of Sec. 3 and Lemma 4.1(iii), we see as in Step 2 of proof of (i)

$\sup_{k\in C}$ operator norm of

$$F(|q|\leqslant K^*(|x|+t-t_0))J_3(t_0,t,\phi,r,k,q,x)$$
$$\leqslant K(r+t-t_0)^{-2}, \quad \text{where } K^* \text{ is as in (6.10).} \tag{6.15}$$

Step 2 (q away from the origin and away from the critical points): We split I_3 into two parts. Choose σ in $(0,\frac{1}{2})$, χ_0 in $C_0^\infty(R^n)$ such that $0\leqslant\chi_0\leqslant 1$, 1 on $|x|\leqslant 1$, 0 on $|x|\geqslant 2$. Define $J_6(\sigma,t_0,t,\cdots)$ by

$$\begin{aligned}
&J_6(\sigma,t_0,t,\phi,r,k,q,x)\\
&\quad=F(|q|\geqslant K^*[|x|+t-t_0])F(x\cdot h_0'(k)\geqslant 0,\ |x|\geqslant r)\\
&\qquad\times\int d\xi\,\phi(\xi)\hat{\eta}(\xi-k)\{W(q,\xi)\\
&\qquad-W(X_\xi'(t_0,t,x,\xi),\xi)\}\\
&\qquad\times\{1-\chi_0((t-t_0)^{\sigma-1}L^{-1}[q-X_\xi'(t_0,t,x,\xi)])\}\\
&\qquad\times\exp(i[q\cdot\xi-X(t_0,t,x,\xi)]),
\end{aligned} \tag{6.16}$$

and, for $0\leqslant\rho\leqslant 1$, define $J_7(\rho,\sigma,\cdots)$ by

$$\begin{aligned}
&J_7(\rho,\sigma,t_0,t,\phi,r,k,q,x)\\
&\quad=F(|q|\geqslant K^*[|x|+t-t_0])F(x\cdot h_0'(k)\geqslant 0,|x|\geqslant r)\\
&\qquad\times\int d\xi\,\phi(\xi)\hat{\eta}(\xi-k)\\
&\qquad\times(\nabla_x W)(\rho q+(1-\rho)X_\xi'(t_0,t,x,\xi),\xi)\\
&\qquad\cdot[q-X_\xi'(t_0,t,x,\xi)]\\
&\qquad\times\chi_0((t-t_0)^{\sigma-1}L^{-1}[q-X_\xi'(t_0,t,x,\xi)])\\
&\qquad\times\exp(i[q\cdot\xi-X(t_0,t,x,\xi)]),
\end{aligned} \tag{6.17}$$

so that

$$\begin{aligned}
&F(|q|\geqslant K^*[|x|+t-t_0])J_3(t_0,t,\phi,r,k,q,x)\\
&\quad=J_6(\sigma,t_0,t,\phi,r,k,q,x)\\
&\qquad+\int_0^1 d\rho\,J_7(\rho,\sigma,t_0,t,\phi,r,k,q,x).
\end{aligned} \tag{6.18}$$

$J_7(\rho,\sigma,\cdots)$ will be estimated from the next step onwards and $J_6(\sigma,\ldots)$ will be estimated now.

Make a change of variable $\xi\to\xi(t-t_0)^{-\sigma}$ in (6.16) and apply the stationary phase Lemma A.1 of Ref. 16 or Lemma I of the Appendix of Ref. 12 to get, using the condition (vii) of Sec. 3 and Lemma 4.1(iii),

$$\begin{aligned}
&|J_6(\sigma,t_0,\ldots)|\\
&\quad\leqslant K_M(t-t_0)^{-n\sigma}(t-t_0+r)^{-\delta}\\
&\qquad\times\inf_{\xi\in C_c}|q-X_\xi'(t_0,t,x,\xi)|^{-M}\\
&\qquad\times F\Big(\inf_{\xi\in C_c}|q-X_\xi'(t_0,t,x,\xi)|\geqslant L(t-t_0)^{1-\sigma}\Big).
\end{aligned} \tag{6.19}$$

Let $C_c(t-t_0)=\{\lambda\in R^n$: all the coordinates of λ are integers and $\lambda(t-t_0)^{-\sigma}$ is in $C_c\}$. Note that the number of elements in $C_c(t-t_0)$ is $K(t-t_0)^{n\sigma}$, where K is independent of t,t_0,σ. Now

$$F(\xi \text{ in } C_c)\leqslant\sum_{\lambda\in C_c(t-t_0)}F(\xi:|\xi-\lambda(t-t_0)^{-\sigma}|_1\leqslant(t-t_0)^{-\sigma}), \tag{6.20}$$

$$|q - X'_\xi(t_0, t, x, \xi)| F(\xi: |\xi - \lambda (t-t_0)^{-\sigma}|_1 \leqslant (t-t_0)^{-\sigma})$$
$$\geqslant \{ |q - X'_\xi(t_0, t, x, \lambda (t-t_0)^{-\sigma})|$$
$$- |X'_\xi t_0, t, x, \lambda (t-t_0)^{-\sigma})$$
$$- X'_\xi(t_0, t, x, \xi)| \} F(\xi: \cdots)$$
$$\geqslant \{ |q - X'_\xi(t_0, t, x, \lambda (t-t_0)^{-\sigma})|$$
$$- [K_1(t-t_0) + K_2(1+t)^{1-\delta}](t-t_0)^{-\sigma}\} F(\xi: \cdots), \tag{6.21}$$

where we have used boundedness of h''_0 in C_{3c}, Lemma 4.1(i), and the mean value theorem in the last step. Now choose L such that

$$\tfrac{1}{2}L \leqslant \sup_{t > t_1} (t-t_0)^{-1}[K_1(t-t_0) + K_2(1+t)^{1-\delta}].$$

Since, in (6.19), $|q - X'_\xi(t_0, t, x, \xi)| \geqslant L(t-t_0)^{1-\sigma}$, we get from (6.21), when λ is in $C_c(t-t_0)$,

$$|q - X'_\xi(t_0, t, x, \xi)| F(\xi: \cdots)$$
$$\geqslant \{ |q - X'_\xi(t_0, t, x, \lambda (t-t_0)^{-\sigma})|$$
$$- \tfrac{1}{2} L (t-t_0)^{1-\sigma} \} F(\xi: \cdots)$$
$$\geqslant \tfrac{1}{2} |q - X'_\xi(t_0, t, x, \lambda (t-t_0)^{-\sigma})|$$
$$\times F(\xi: |\xi - \lambda (t-t_0)^{-\sigma}|_1 \leqslant t - t_0)^{-\sigma}). \tag{6.22}$$

From (6.19), (6.20), and (6.22),

$$|J_6(\sigma, t_0, t, \phi, r, k, q, x)|$$
$$\leqslant K_M (t-t_0)^{-n\sigma}(t-t_0+r)^{-\delta} \sum_{\lambda \in C_c(t-t_0)}$$
$$F(|q - X'_\xi(t_0, t, x, \lambda (t-t_0)^{-\sigma})| \geqslant L(t-t_0)^{1-\sigma})$$
$$\times |q - X'_\xi(t_0, t, x, \lambda (t-t_0)^{-\sigma})|^{-M}. \tag{6.23}$$

Clearly for λ in $C_c(t-t_0)$

$$\int dq\, F(|q - X'_\xi(t_0, t, x, \lambda (t-t_0)^{-\sigma})| \geqslant L(t-t_0)^{1-\sigma})$$
$$\times |q - X'_\xi(t_0, t, x, \lambda (t-t_0)^{-\sigma})|^{-M}$$
$$\leqslant \int dq\, F(|q| \geqslant L(t-t_0)^{1-\sigma})|q|^{-M}$$
$$\leqslant [L(t-t_0)^{1-\sigma}]|^{-M+n} \quad \text{a.e.x.} \tag{6.24}$$

Also

$$\int dx\, F(|q - X'_\xi(t_0, t, x, \lambda (t-t_0)^{-\sigma})| \geqslant L(t-t_0)^{1-\sigma})$$
$$\times |q - x'_\xi(t_0, t, x, \lambda (t-t_0)^{-\sigma})|^{-M}$$
$$= \int dy\, F(|q-y| \geqslant L(t-t_0)^{1-\sigma})|q-y|^{-M}$$
$$\times |\det[(\nabla_x \nabla_\xi X)(t_0, t, x, \lambda (t-t_0)^{-\sigma})]|^{-1}$$
$$\leqslant K \int dy\, F(|q-y| \geqslant L(t-t_0)^{1-\sigma})|q-y|^{-M},$$

where K depends only on C_c due to $|\nabla_x \nabla_\xi X(t_0, t, x, \xi) - I| \leqslant \tfrac{1}{2}$,

$$\leqslant K[(t-t_0)^{1-\sigma}]^{-M+n}. \tag{6.25}$$

By (6.23)–(6.25) and Lemma 2.3, we have

$$\sup_{k \in C} \|J_6(\sigma, t_0, t, \phi, r, k, \cdot, \cdot)\|$$
$$\leqslant K(L)(t-t_0)^{(-M+n)(1-\sigma)}(t-t_0+r)^{-\delta}. \tag{6.26}$$

Step 3 (q nearer the critical points: An expansion using Morse's lemma): Put

$$W_j(x, \xi) = \partial W(x, \xi)/\partial x_j,$$
$$\chi_j(x) = x_j \chi_0(x) \tag{6.27}$$

so that by (6.17)

$$(t-t_0)^{\sigma-1} L^{-1} J_7(\rho, \sigma, t_0, t, \phi, r, k, q, x)$$
$$= \Big\{ \sum_{j=1}^n \int d\xi \phi(\xi) \hat{\eta}(\xi - k)$$
$$\times W_j(\rho q + (1-\rho) X'_\xi(t_0, t, x, \xi), \xi)$$
$$\times \chi_j((t-t_0)^{\sigma-1} L^{-1}[q - X'_\xi(t_0, t, x, \xi)])$$
$$\times \exp(i[q \cdot \xi - X(t_0, t, x, \xi)]) \Big\}$$
$$\times F(x \cdot h'_0(k) \geqslant 0,\ |x| \geqslant r) F(|q| \geqslant K^*[|x| + t - t_0]). \tag{6.28}$$

As in Step 4 of the proof of Lemma 5.1(iv), let $\xi_0 = \xi_0(t_0, t, q, x)$ in C_c be the unique critical point, i.e.,

$$q = X'_\xi(t_0, t, x, \xi_0), \quad \xi_0 \in C_c. \tag{6.29}$$

Apply Lemma A.6 of Ref. 16 to the function $\xi \to (t-t_0)^{-1}\{q \cdot \xi - X(t_0, t, x, \xi)\}$ to get a function ψ defined in a neighborhood of 0 such that

$$\psi(0) = \xi_0, \tag{6.30}$$

$$q \cdot \psi(y) - X(t_0, t, x, \psi(y))$$
$$= q \cdot \xi_0 - X'_\xi(t_0, t, x, \xi_0) + (t-t_0)\langle Ay, y \rangle, \tag{6.31}$$

$$A = X''_\xi(t_0, t, x, \xi_0)/(2(t-t_0)). \tag{6.32}$$

Note that A, ψ depend on ξ_0 and therefore on t_0, t, q, x. Since ξ_0 is always in C_c, the real symmetric matrix A varies in a compact subset of $GL(n, R)$. The signature of a matrix is a continuous function of the matrix. So, by shrinking C_c if necessary, we can assume that the signature of A is constant for all t, t_0, q, x.

Applying Lemma A.2 and A.4 of Ref. 16, we have, for $M \geqslant 2$,

rhs (6.28)

$$= F(x \cdot h'_0(k) \geqslant 0,\ |q| \geqslant K^*[|x| + t - t_0], |x| \geqslant r)$$
$$\times \exp(i[q \cdot \xi_0 - X(t_0, t, x, \xi_0)])$$
$$\times \sum_{m=1}^n \sum_{j=0}^{M-1} K_j (t-t_0)^{-n/2-j}$$
$$\times (\det A)^{-1/2}[\langle A^{-1}D, D \rangle^j$$
$$\times \{\phi(\psi(y)) \hat{\eta}(\psi(y) - k) W_m(\rho q + (1-\rho)$$
$$\times X'_\xi(t_0, t, x, \psi(y)), \psi(y)) |\det \psi'(y)|$$
$$\times \chi_m((t-t_0)^{\sigma-1} L^{-1}[q - X'_\xi(t_0, t, x, \psi(y))])\}(0)]$$
$$+ M\text{th term}, \tag{6.33}$$

where (i) the K_j are constants independent of t_0, t, q, x, (ii) D is the differential operator $(\partial/\partial y_1, \ldots, \partial/\partial y_n)$, and

$|M\text{th term}|$

$$\leqslant K_M(t-t_0)^{-n/2-M}$$
$$\times F(x\cdot h_0'(k)\geqslant 0,\ |q|\geqslant K^*[|x|+t-t_0],\ |x|\geqslant r)$$
$$\times \sum_{m=1}^{n}\sum_{|\alpha|\leqslant 2M+n} \|D_\xi^\alpha\{\phi(\xi)\hat{\eta}(\xi-k)$$
$$\times W_m(\rho q+(1-\rho)X_\xi'(t_0,t,x,\xi),\xi)$$
$$\times \chi_m((t-t_0)^{\sigma-1}L^{-1}[q-X_\xi'(t_0,t,x,\xi)])\}\|_\infty . \quad (6.34)$$

Now we estimate the norm of the M th term. The norm of the other terms of the rhs of (6.33) will be estimated from next step onwards. For (6.34) we have

$$|\rho q+(1-\rho)X_\xi'(t_0,t,x,\xi)|$$
$$\geqslant |X_\xi'(t_0,t,x,\xi)|-\rho|q-X_\xi'(t_0,t,x,\xi)|$$
$$\geqslant 2^{-1/2}[|x|+2a(t-t_0)]-K_1(1+t)^{1-\delta}$$
$$-2L(t-t_0)^{1-\sigma}$$

by (4.8), Lemma 4.1(i), and the presence of χ_m in (6.34)

$\geqslant a(|x|+t-t_0)$ with some $a>0$

if $t\geqslant t_1$ with t_1-t_0 large. (6.35)

By (6.35) and condition (vii) of Sec. 3.

$|M$ th term of the rhs of (6.33)$|$

$$\leqslant K_M(t-t_0)^{-n/2-M+(2M+n)\sigma}(t-t_0+r)^{-1-\delta}$$
$$\times F(x\cdot h_0'(k)\geqslant 0,\ |x|\geqslant r)$$
$$\times F\Big(\sup_{\xi\in C_c}|q-X_\xi'(t_0,t,x,\xi)|\leqslant 2L(t-t_0)^{1-\sigma}\Big). \quad (6.36)$$

It is easily seen, by using (4.7), that

$$F\Big(\sup_{\xi\in C_c}|q-X_\xi'(t_0,t,x,\xi)|\leqslant 2L(t-t_0)^{1-\sigma}\Big)$$
$$\leqslant \sum_{\lambda\in C_c(t-t_0)} F(|q-X_\xi'(t_0,t,x,\lambda(t-t_0)^{-\sigma})|$$
$$\leqslant (2L+2+a)(t-t_0)^{1-\sigma}). \quad (6.37)$$

Again using the techniques in (6.24) and (6.25) we see that for each λ in $C_c(t-t_0)$

operator norm of

$$F(|q-X_\xi'(t_0,t,x,\lambda(t-t_0)^{-\sigma})|\leqslant (2L+2+a)(t-t_0)^{1-\sigma})$$
$$\leqslant K(L)(t-t_0)^{n(1-\sigma)}. \quad (6.38)$$

By (6.36)–(6.38)

operator norm of M th term of rhs of (6.33) for all k in C, $1\geqslant\rho\geqslant 0$,

$$\leqslant K_M(L)(r+t-t_0)^{-1-\delta}$$
$$\times(t-t_0)^{-n/2-M+(2M+n)\sigma+n(1-\sigma)+n\sigma}$$
$$\leqslant K_M(L)(r+t-t_0)^{-1-\delta}(t-t_0)^{-M(1-2\sigma)+n(\sigma+1/2)}. \quad (6.39)$$

Step 4 (bounds on the derivatives of ξ_0 and ψ): Note that since $q=X_\xi'(t_0,t,x,\psi(0))$ and $\chi_m(0)=0$ for each m

jth term of the rhs of (6.33) for $j=0$ vanishes

for each $m=1,2,\dots,n$. (6.40)

By Sec. 8 we get if C_c is sufficiently small, then (i) for each fixed x the map $\psi=\psi(t_0,t,q,x;y)$ is a C^∞ function of q and y, (ii) $\xi_0=\xi_0(t_0,t,q,x)$ is a C^∞ function of q and x for fixed t, t_0, (iii) $|D_q^\alpha\xi_0|\leqslant K_\alpha$ for all t_0,t,q,x under consideration, (iv) $|D_q^\alpha A(t_0,t,q,x)|+|D_q^\alpha A^{-1}(t_0,t,q,x)|\leqslant K_\alpha$ for all t_0,t,q,x under consideration, (v) $|D_q^\alpha D_y^\beta\psi|$ at $y=0\leqslant K_{\alpha\beta}$ for all t_0,t,q,x under consideration, (vi) $|D_q^\alpha\{\xi_0(q,x_1)-\xi_0(q,x_2)\}|\leqslant K_\alpha|x_1-x_2|(t-t_0)^{-1}$ for all α,q,x_1,x_2,t,t_0, and (vii) $|\xi_0(q,x_1)-\xi_0(q,x_2)|\geqslant K_0^*|x_1-x_2|(t-t_0)^{-1}$ for some $K_0^*>0$ for all t_0,t,q,x_1,x_2.

The proof of (i)–(vii) is not short. Since we do not want to break the continuity of the proof of the present lemma, we prove these results in Sec. 8.

Step 5 [simplifying the terms on the rhs of (6.33)]: Define for $j=1,2,\dots,M-1$, $m=1,2,\dots,n$,

$$f_{m,j}(q,x)=f_{m,j}(\rho,\sigma,t_0,t,k,q,x)$$
$$=(\det A)^{-1/2}[\langle A^{-1}D,D\rangle^j$$
$$\times\{\phi(\psi(y))\hat{\eta}(\psi(y)-k)|\det\psi'(y)|$$
$$\times W_m(\rho q+(1-\rho)X_\xi'(t_0,t,x,\psi(y)),\psi(y))$$
$$\times\chi_m((t-t_0)^{\sigma-1}L^{-1}[q-X_\xi'(t_0,t,x,\psi(y))]]\}(0) \quad (6.41)$$

Then by the assertions (i)–(v) of Step 4, (6.35),

$$\sum_{m=1}^{n} f_{m,j}(q,x)=(t-t_0+r)^{-1-\delta}(t-t_0)^{2j\sigma}g_j(q,x), \quad (6.42)$$

where $g_j=g_j(\rho,\sigma,t_0,t,k,q,x)$ satisfies

$|D_q^\alpha g_j(q,x)|\leqslant K_\alpha$ for all $\alpha,\rho,\sigma,t_0,t,k,q,x$;

further, for every fixed t_0,t,k,ρ,x the function g_j has compact support in q. (6.43)

Put

$$h_j(q,x)=\exp(i[q\cdot\xi_0-X(t_0,t,x,\xi_0)])g_j(q,x),$$
$$=h_j(\rho,\sigma,t_0,t,k,q,x). \quad (6.44)$$

Step 6 (calculating the operator norm of h_j): Let $H_j=H_j(\rho,\sigma,t_0,t,k)$ be the operator induced by the kernel h_j. Then

(the kernel of $H_j^*H_j$)(q,x)

$$=\int dy\,\bar{h}_j(y,q)h_j(y,x)$$
$$=\int dy\,\bar{g}_j(y,q)g_j(y,x)\exp(i[y\cdot\xi_0(y,x)$$
$$-X(t_0,t,x,\xi_0(y,x))-y\cdot\xi_0(y,q)$$
$$+X(t_0,t,q,\xi_0(y,q))]), \quad (6.45)$$

∇_y (the phase in the rhs of (6.45))

$$=\nabla_y\{y\cdot\xi_0(y,x)-X(t_0,t,x,\xi_0(y,x))$$
$$-y\cdot\xi_0(y,q)+X(t_0,t,q,\xi_0(y,q))\}$$
$$=\xi_0(y,x)-\xi_0(y,q)\quad\text{by (6.29)}. \quad (6.46)$$

By (6.46) and assertions (vi) and (vii) of Step 4, we have

$|D_y^\alpha$ (the phase on rhs of (6.45))$|$

$\leqslant K_\alpha|q-x|(t-t_0)^{-1}$ for $|\alpha|\geqslant 1$,

$|\nabla_y$ (the phase on rhs of (6.45))$|$

$\geqslant K^*|q-x|(t-t_0)^{-1}$ for some $K^*>0$.

Now one can apply the stationary phases Lemma A.1 of Ref. 16 or Lemma I of the Appendix of Ref. 12 to the rhs of (6.45) using (6.43) to get

$$|\text{rhs of (6.45)}| \leqslant K_N(1+|q-x|(t-t_0)^{-1})^{-N}. \qquad (6.47)$$

Take $N=n, n+1$ in (6.47) and interpolate to get 1 the integral kernel of $(H_j^* H_j)$ at $(q, x)| \leqslant K(1+|q-x|(t-t_0)^{-1})^{-n-\delta/4}$ so that

$$\|h_j(\cdot,\cdot)\| \leqslant K(t-t_0)^{n/2+\delta/8} \quad \text{for all } k, \rho, \sigma, t_0, r. \qquad (6.48)$$

Now from (6.33), (6.39), (6.40), (6.41), (6.42), (6.44), and (6.48) we get

Operator norm of rhs of (6.28)

$$\begin{aligned}
&\leqslant K_M(t-t_0+r)^{-1-\delta}(t-t_0)^{-M(1-2\sigma)+n(\sigma+1/2)} \\
&\quad + \sum_{j=1}^{M-1} K_j(t-t_0)^{-n/2-j}(t-t_0+r)^{-1-\delta} \\
&\quad \times (t-t_0)^{2j\sigma}(t-t_0)^{n/2+\delta/8} \\
&\leqslant K_M(t-t_0+r)^{-1-\delta}[(t-t_0)^{-M(1-2\sigma)+n(\sigma+1/2)} \\
&\quad + (t-t_0)^{-1+2\sigma+\delta/8}] \\
&\qquad \text{since } \sigma \text{ is in } (0, \tfrac{1}{2}).
\end{aligned} \qquad (6.49)$$

Step 7 (the concluding step): By (6.15), (6.18), (6.26), (6.28), and (6.49) we get

$$\begin{aligned}
&\sup_{k\in C} \|J_3(t_0, t, \phi, r, k, \cdot, \cdot)\| \\
&\leqslant K\{(r+t-t_0)^{-2} + (t-t_0+r)^{-\delta}(t-t_0)^{(-M+n)(1-\sigma)} \\
&\quad + (t-t_0+r)^{-1-\delta}(t-t_0)^{-M(1-2\sigma)+n(\sigma+1/2)+1-\sigma} \\
&\quad + (t-t_0+r)^{-1-\delta}(t-t_0)^{-1+2\sigma+\delta/8+1-\sigma}.
\end{aligned} \qquad (6.50)$$

Now choose $\sigma = \frac{1}{8}\delta$, M large, such that

$$(-M+n+1)(1-\tfrac{1}{8}\delta) \leqslant -1,$$

and $-M(1-\frac{1}{4}\delta)+n(\frac{1}{8}\delta+\frac{1}{2})+1-\frac{1}{8}\delta \leqslant 0$ so that from (6.50)

$$\begin{aligned}
&\sup_{k\in C} \|J_3(t_0, t, \phi, r, k, \cdot, \cdot)\| \\
&\leqslant K\{(t-t_0+r)^{-2} + (t-t_0+r)^{-\delta}(t-t_0)^{-1} \\
&\quad + (t-t_0+r)^{-1-\delta} + (t-t_0+r)^{-1-\delta}(t-t_0)^{\delta/4}\} \\
&\qquad \text{for all } t \geqslant t_1 \geqslant t_0 \text{ with } t_1 - t_0 \text{ large.}
\end{aligned}$$

Now the result follows by noting $\delta > 0$.

(viii) By (6.6), (6.8), (ii), (iii), (vii), and Lemma 2.4, we have

$$\lim_{r\to\infty} \int_{t_1}^{\infty} dt \left\| \frac{d}{dt} \{V_{t-t_0}^* E_1(t_0, t, C, \phi, +, r)\right\| = 0.$$

Now the result follows from (vi).

(ix) Follows from (iii) and (6.5).

(x) By (6.12) and (ix) we get

$$\lim_{r\to\infty} \sup_{t>t_0} \|(1+|Q|)^{-2} V_{t-t_0}\phi(P) T(E(C,+,r))\| = 0.$$

This leads to, since Range$(1+|Q|)^{-2}$ is dense in $L^2(R^n)$,

$$\lim_{r\to\infty} \sup_{t\geqslant 0} \|T(E(C,+,r))\phi(P) V_t^* f\| = 0$$
$$\text{for each } f \text{ in } L^2(R^n). \qquad (6.51)$$

By Lemma 2.1(iii) and the RAGE theorem,[4] we have for f in $\mathscr{H}_c(H)$

$$\begin{aligned}
&\lim_{S\to\infty} S^{-1} \int_0^S dt \, \|T\{|x| \leqslant r, k\in C, x\cdot h_0'(k) \geqslant 0\} \\
&\quad \times \phi(P) V_t^* f\| = 0 \\
&\qquad \text{for each fixed } r.
\end{aligned} \qquad (6.52)$$

The result follows from (6.51) and (6.52). Q.E.D.

Lemma 6.2: Let ϕ be as in (5.4), C, c, T be as in Lemma 6.1(viii) and $Z(t)$, $Z(t_0, t)$ as in (5.3). Then:

(i) $\phi(P) E_1(t_0, t, C, \phi, +, r) = E_1(t_0, t, C, \phi^2, +, r)$;

(ii) $\text{s-lim}_{t\to\infty} Z(t_0, t)^* E_1(t_0, t, C, \phi, +, r)$ exists;

(iii) $w_1(t_0, \phi, C, +, r)$
$= \text{s-lim}_{t\to\infty} Z(t-t_0)^* E_1(t_0, t, C, \phi^2, +, r)$ exists;

(iv) $\Omega_1(t_0, \phi, C, +, r)$
$= \text{s-lim}_{t\to\infty} V_{t-t_0}^* E_1(t_0, t, C, \phi^2, +, r)$ exists;

(v) $\Omega_1(t_0, \phi, C, +, r) = \Omega_+ w_1(t_0, \phi, C, +, r)$;

(vi) $\lim_{r\to\infty} \|\phi^2(P) T(E(C, +, r)) - \Omega_1(t_0, \phi, C, +, r)\| = 0$;

(vii) $\lim_{r\to\infty} \|(1-\Omega_+\Omega_+^*)\phi^2(P) T(E(C, +, r))\| = 0.$

Proof: (i) Follows from (6.3) and (6.4).

(ii) We deduce this result by verifying the conditions of Lemma 2.7 by using Theorem 2.5 and results of Sec. 4. A simple calculation shows that for f in $L^2(R^n)$

$$\begin{aligned}
&Z(t_0, t)^* E_1(t_0, t, C, \phi, +, r) \\
&= \int dx \, dk \, \langle \eta_{xk} | f \rangle I(t_0, t, C, \phi, +, r, k, q, x),
\end{aligned} \qquad (6.53)$$

$$\begin{aligned}
&I(t_0, t, C, \phi, +, r, k, q, x) \\
&= F(E(C, +, r)) \int d\xi \, \phi(\xi) \hat{\eta}(\xi - k) \\
&\quad \times \exp(i[q\cdot\xi - A(t_0, t, x, \xi)]), \\
&A(t_0, t, x, \xi) = x\cdot\xi - t_0 h_0(\xi) + Y(m_0, t_0, t, x, \xi) \\
&\quad - Y(m_0, t_0, t, 0, \xi),
\end{aligned} \qquad (6.54)$$

where Y is as in (4.10).

The integral operator induced by the kernel $I(t_0, t, C, \phi, +, r, k, q, x)$ will be denoted by $I(t_0, t, C, \phi, +, r, k)$. The presence of ϕ and $\hat{\eta}$ forces that for some bounded subset B of R^n:

$$I(t_0, t, C, \phi, +, r, k) = 0 \quad \text{for } k \text{ not in } B. \qquad (6.55)$$

Further, by Theorem 2.5 and Lemma 4.1(ii), we see that there is some $r_0 > 0$ so that

$$\sup\{\|I(t_0, t, C, \phi, +, r, k)\| : r \geqslant r_0, t \geqslant t_0, k \text{ in } R^n\} < \infty. \qquad (6.56)$$

When x and ξ are as in (4.4) and (4.5), it is easy to verify as in (4.29) that for each α

$$\sup_{x,\xi}\left|D_\xi^\alpha \frac{\partial}{\partial t}\{A(t_0,t,x,\xi)\}\right|$$
$$\leqslant K_\alpha(1+|t|)^{-1-\delta}(1+|x|).$$

By the last inequality, Lemma 4.1(ii), and Theorem 2.5, we arrive at, for each $s>0$,

$$\int_{t_0}^\infty dt \text{ operator norm of the kernel}$$
$$\frac{d}{dt}I(t_0,t,C,\phi,+,r,k,q,x)F(|x|\leqslant s)<\infty. \quad (6.57)$$

By (6.56) and (6.57) we have

$$\text{s-}\lim_{t\to\infty} I(t_0,t,C,\phi,+,r,k) \text{ exists for each } k. \quad (6.58)$$

Now the result follows from (6.53), (6.55), (6.56), (6.58), and Lemma 2.7.

(iii) Clearly s-$\lim_{t\to\infty} Z(t-t_0)^*Z(t_0,t)\phi(P)$ exists by Lemma 4.2(ii), (iii).

The result follows from (i), (ii).

(iv) Follows from Lemma 6.1 (viii).

(v) Follows from (iii) and (iv).

(vi) Follows from (iv) and Lemma 6.1(ix).

(vii) Since $1-\Omega_+\Omega_+^*$ is bounded, we get from (vi) that

$$\lim_{r\to\infty}\|(1-\Omega_+\Omega_+^*)\{\phi^2(P)T(E(C,+,r))$$
$$-\Omega_1(t_0,C,\phi,+,r)\}\|=0.$$

Since $(1-\Omega_+\Omega_+^*)\Omega_+=0$, we get by (v) that $(1-\Omega_+\Omega_+^*)\Omega_1(t_0,C,\phi,+,r)=0$ for each r and now the result is clear. Q.E.D.

Remark 6,3: By Remark 5.3 we have the operator Ω_-. Using the techniques of the proof of Lemma 6.1 and 6.2, we can prove Lemma 6.1 (x) with V_t^* replaced by V_t and $+$ replaced by $-$; Lemma 6.2(vii) with $+$ replaced by $-$.

More precisely we have proved the following

Lemma 6.4: Let G be given by (4.1) and q be in G. Then there exists a compact sphere (of nonzero radius) A_q with center q and a constant $a_q>0$ such that $\{p: \text{dist}(p,A_q)\leqslant a_q\}\subset G$. Let T be *any* operator of Sec. 2 with [refer to (2.1)] supp $\hat\eta\subset\{k:|k|\leqslant\frac18 b\}$ such that $0<b\leqslant a_q$. Let ϕ be in $C_0^\infty(G)$. Then:

(i) $\lim_{r\to\infty}\|(1-\Omega_\pm\Omega_\pm^*)\phi^2(P)T(E(A_q,\pm,r))\|=0$

when supp $\phi\subset\{p: \text{dist}(p,A_q)<a_q\}$;

(ii) $\lim_{S\to\infty} S^{-1}\int_0^S dt\,\|T(E(A_q,\pm,0))\phi^2(P)V_{\mp t}f\|=0$

for f in $\mathscr{H}_c(H)$.

7. PROOF OF THEOREM 1(c), (d), (e), (f)

Let the Hamiltonians H and H_0 be as in Sec. 6 and further satisfying condition (xi) of Sec. 3 and let G be as in (4.1).

Theorem 7.1: Let A be any compact subset of G. Then there exists an operator T of Sec. 2 depending only on A such that

(i) $(1-\Omega_\pm\Omega_\pm^*)\phi^2(P)T\{x\cdot h_0'(k)\geqslant 0\}$ is compact

for each ϕ in $C_0^\infty(G)$ with supp $\phi\subset A$,

(ii) $\lim_{S\to\infty} S^{-1}\int_0^S dt\,\|T\{x\cdot h_0'(k)\gtrless 0\}\phi^2(P)V_{\pm t}^*f\|=0$

for f in $\mathscr{H}_c(H)$, ϕ as in (i).

Proof: (i) *Step 1*: Let q be in G, A_q, a_q, T as in Lemma 6.4 and ψ in $C_0^\infty(G)$. Then we prove that

$$\lim_{r\to\infty}\|(1-\Omega_\pm\Omega_\pm^*)\psi^2(P)T(E(A_q,\pm,r)\|=0. \quad (7.1)$$

Choose a smooth ϕ such that $\phi=1$ on $\{p: \text{dist}(p,A_q)\leqslant\frac14 b\}$ and supp $\phi\subset\{p: \text{dist}(p,A_q)<a_q\}$. Then by Lemma 6.4

$$\lim_{r\to\infty}\|(1-\Omega_\pm\Omega_\pm^*)\psi^2(P)\phi^2(P)T(E(A_q,\pm,r))\|=0. \quad (7.2)$$

By Lemma 2.2(v), $[1-\phi^2(P)]T(E(A_q,\pm,r))=0$ for each r. Now (7.1) follows from (7.2).

Step 2: Let A be compact in G. Then there is an operator T of Sec. 2 depending only on A such that for all ψ in $C_0^\infty(G)$

$$(1-\Omega_\pm\Omega_\pm^*)\psi^2(P)T(E(A,\pm,0)) \text{ is compact.} \quad (7.3)$$

Let for q in G A_q, a_q be as in Lemma 6.4. A is compact. So A can be covered by finitely many A_q's; call them $A_1,\dots,A_m$. Let $0<b\leqslant\frac12\{a_1,\dots,a_m\}$. Choose T of Sec. 2 with supp $\hat\eta\subset\{k: |k|\leqslant\frac18 b\}$. Then for this T by Step 1 we have for $j=1,\dots,m$

$$\lim_{r\to\infty}\|(1-\Omega_\pm\Omega_\pm^*)\psi^2(P)T(E(A_j,\pm,r))\|=0. \quad (7.4)$$

By Lemma 2.2(ii), (iv) we get

$$\lim_{r\to\infty}\|(1-\Omega_\pm\Omega_\pm^*)\psi^2(P)T(E(A,\pm,r))\|=0. \quad (7.5)$$

By (7.5) and Lemma 2.1(iii) we get (7.3).

Step 3: Let A be compact in G. Choose $\epsilon>0$ such that $A_\epsilon=\{p: \text{dist}(p,A)\leqslant\epsilon\}\subset G$. By the proof of Step 2, there is some number $a>0$ such that for all b with $0<b\leqslant a$ the operator $(1-\Omega_\pm\Omega_\pm^*)\psi^2(P)T(E(A_\epsilon,\pm,0))$ is compact for all T with supp $\hat\eta\subset\{k: |k|\leqslant\frac18 b\}$. Now choose $0<b\leqslant\min\{\epsilon,a\}$.

Let $\phi\in C_0^\infty(G)$ be with supp $\phi\subset A$. Then

$$\begin{aligned}\phi^2(P)T\{x\cdot h_0'(k)\lesseqgtr 0\} &= \phi^2(P)T(E(A_\epsilon,\pm,0))\\ &\quad+\phi^2(P)T(E(R^n\backslash A_\epsilon,\pm,0))\\ &=\phi^2(P)T(E(A_\epsilon,\pm,0)) \quad \text{by Lemma 2.2(v).} \end{aligned}\quad (7.6)$$

Now the result follows from (7.6) and the compactness of $(1-\Omega_\pm\Omega_\pm^*)\phi^2(P)T(E(A_\epsilon,\pm,0))$. This proves (i).

(ii) The proof uses techniques similar to the proof of (i). As in Step 2 of (i), using Lemma 2.2(i), (iii), we get for A compact in G, ϕ in $C_0^\infty(G)$, f in $\mathscr{H}_c(H)$,

$$\lim_{S\to\infty} S^{-1}\int_0^S dt\,\|T(A,\mp,0)\phi^2(P)V_{\pm t}f\|=0.$$

Now the argument is similar to Step 3 of proof of (i). Q.E.D.

Theorem 7.2:

$\mathscr{H}_c(H)$

$$= \left\{ f \in \mathscr{H}_c(H) : \lim_{S\to\infty} S^{-1} \int_0^S dt \, \|\phi(P) V_{\pm,t} f\| = 0 \text{ for each } \phi \text{ in } C_0^\infty(G) \right\} \oplus \text{Range } \Omega_\pm .$$

Proof: We prove only for the $+$ sign; for the $-$ sign the proof is similar. Let us remark that $\oplus$ stands for sum of two closed orthogonal subspaces. Let

$$M_+ = \left\{ f \in \mathscr{H}_c(H) : \lim_{S\to\infty} S^{-1} \int_0^S dt \, \|\phi(P) V_t f\| = 0 \text{ for each } \phi \text{ in } C_0^\infty(G) \right\}.$$

It is clear that M_+ is a closed linear subspace of $\mathscr{H}_c(H)$. Since Ω_+ is an isometry, Range Ω_+ is closed.

Step 1: to show the orthogonality of M_+ and Range Ω_+: Let $f \in M_+$, $g \in$ Range Ω_+. Choose h in $L^2(R^n)$ so that $g = \Omega_+ h$, i.e., $\|V_t g - Z_t h\| \to 0$ as $t \to \infty$ with Z_t as in Theorem 5.2. From the identity

$$\langle f, g \rangle = \langle V_t f, V_t g - Z_t h \rangle + \langle \bar\phi(P) V_t f, Z_t h \rangle + \langle V_t f, Z_t [1 - \phi(P)] h \rangle$$

we easily get

$$|\langle f, g \rangle| \leqslant \|[1 - \phi(P)]h\| \quad \text{for each } \phi \text{ in } C_0^\infty(G).$$

Now we get $\langle f, g \rangle = 0$ by condition (iv) of Sec. 3.

It is clear from Theorem 5.2(iii) that $M_+ \oplus \text{Range } \Omega_+ \subset \mathscr{H}_c(H)$. Thus to prove the theorem it is enough to show:

Step 2: $\mathscr{H}_c(H) \ominus \text{Range } \Omega_+ \subset M_+$: Let $f \in \mathscr{H}_c(H) \ominus \text{Range } \Omega_+$ and $\phi \in C_0^\infty(G)$ be real valued. Take $A = \text{supp } \phi$ in Theorem 7.1 to get an operator T such that $(1 - \Omega_+ \Omega_+^*) \phi^2(P) T\{x \cdot h_0'(k) \geqslant 0\}$ is compact. Now, applying RAGE theorem,[4] we get

$$\lim_{S\to\infty} S^{-1} \int_0^S dt \, \|(1 - \Omega_+ \Omega_+^*) \phi^2(P) T\{x \cdot h_0'(k) \geqslant 0\} \times \phi^2(P) V_t f\| = 0. \tag{7.7}$$

Also

$$\langle \Omega_+ \Omega_+^* \phi^2(P) T\{x \cdot h_0'(k) \geqslant 0\} \phi^2(P) V_t f, V_t f \rangle = 0 \quad \text{since } f \perp \text{Range } \Omega_+ \tag{7.8}$$

Clearly,

$$\begin{aligned} \|\phi^2(P) V_t f\|^2 &= \langle (1 - \Omega_+ \Omega_+^*) \phi^2(P) T\{x \cdot h_0'(k) \geqslant 0\} \times \phi^2(P) V_t f, V_t f \rangle \\ &\quad + \langle \Omega_+ \Omega_+^* \phi^2(P) T\{x \cdot h_0'(k) \geqslant 0\} \phi^2(P) V_t f, V_t f \rangle \\ &\quad + \langle T\{x \cdot h_0'(k) \leqslant 0\} \phi^2(P) V_t f, \phi^2(P) V_t f \rangle. \end{aligned} \tag{7.9}$$

By (7.7), (7.8), Theorem 7.1(ii), and (7.9), we have

$$\lim_{S\to\infty} S^{-1} \int_0^S dt \, \|\phi^2(P) V_t f\|^2 = 0$$

for each real-valued ϕ in $C_0^\infty(G)$. The proof of Step 2 is now obvious. Q.E.D.

Theorem 7.3: Let H and H_0 be as in Sec. 6 and satisfying condition (xi) of Sec. 3. Then Theorem 1(c), (d), (e), and (f) are true.

Proof of Theorem 1(c): We prove for $+$ sign; for $-$ *ve* sign the proof is similar. We deduce the proof from Theorem 7.2.

Let $C_v = \text{closure } \{h_0(\xi) : h_0'(\xi) = 0 \text{ or } \det h_0''(\xi) = 0\}$. Then C_v is a countable closed subset of R. Let $\psi \in C_0^\infty(R \setminus C_v)$ and define $\phi(P) = \psi(h_0(P)) = \psi(H_0)$. Then $\phi \in C_0^\infty(G)$ so that for f in M_+ of Theorem 7.2 (refer the proof) we get $\lim_{S\to\infty} S^{-1} \int_0^S dt \, \|\psi(H_0) V_t f\| = 0$. But by RAGE theorem[4] and Lemma 3.4(vii), for f in M_+, $\lim_{S\to\infty} S^{-1} \int_0^S dt \, \|[\psi(H) - \psi(H_0)] V_t f\| = 0$. So we have $\psi(H) f = 0$ for each f in M_+ and for each ψ in $C_0^\infty(R \setminus C_v)$ forcing $M_+ = \{0\}$. Theorem 1(c) now follows from Theorem 7.2.

Proof of Theorem 1(d) and (e): Theorem 1(d) and (e) follow from (c), Theorem 5.2, and Remark 5.3.

Proof of Theorem 1(f): Let E be the (orthogonal) projection onto the point spectral subspace for H. The result will easily follow if $\psi^2(H) E$ is compact for each ψ in $C_0^\infty(R \setminus C_v)$.

Let $\psi \in C_0^\infty(R \setminus C_v)$. By (c) $E = 1 - \Omega_\pm \Omega_\pm^*$ so that by taking $\phi(P) = \psi(H_0) = \psi(h_0(P))$ and applying Theorem 7.1(i), we have that $E\psi^2(H_0)$ is compact. Now, using Lemma 3.4(vii), we get $E\psi^2(H)$ is compact. Q.E.D.

Remark 7.4: One can try to use the techniques developed in this publication to prove similar results when (i) the potentials are time-dependent potentials as in Ref. 30 or (ii) h_0 of Sec. 3 is replaced by a vaguely elliptic function of Ref. 11 and W_S is replaced by a regular perturbation of Ref. 11.

8. BOUNDS ON THE DERIVATIVES OF ξ_0, ψ OF SEC. 6.

Lemma 8.1: Let $M = M(n, R)$, $M_s = M_s(n, R)$, $GL(n, R)$ be the set of all $n \times n$, $n \times n$ symmetric, $n \times n$ invertible matrices over the reals. If $A_0 \in M_s \cap GL(n, R)$, then there exists a C^∞ map g defined in a neighborhood of (A_0, A_0) in $M_s \times M_s \to GL(n, R)$ such that $g(A_0, A_0) = 1$ and $g(A, B)^* A g(A, B) = B$ for all (A, B) in the domain of g.

Proof: We follow Lemma A.3 of Ref. 16. Define F: $M \times M_s \times M_s \to M_s$ by $F(R, A, B) = R^* A R - B$. Then F is C^∞, $F(1, A_0, A_0) = 0$, and $F'(1, A_0, A_0)(R, 0, 0) = R^* A_0 + A_0 R$. For any matrix E in M_s, if we take $R = \frac{1}{2} A_0^{-1} E$, then $F'(1, A_0, A_0)(R, 0, 0) = E$. Thus $F'(1, A_0, A_0) : M \times 0 \times 0 \to M_s$ is onto. By the implicit function theorem the result follows. Q.E.D.

Let $q, x, \xi_0 = \xi_0(t_0, t, q, x)$ and C, c be as in Step 4 of the proof of Lemma 6.1(vii). In what follows we shall keep on shrinking C; since we do the shrinking only *finitely* many times, we shall be left with a positive radius for C.

Lemma 8.2: If C_c is sufficiently small, there exists a function $Z = Z(t_0, t, q, x, \xi)$ such that, for $A = [2(t - t_0)]^{-1} X_\xi''(t_0, t, q, \xi_0)$, we get $q \cdot \xi - X(t_0, t, x, \xi) = q\xi_0 - X(t_0, t, x, \xi_0) + (t - t_0) \langle AZ, Z \rangle$ for all ξ in C_c. Further $Z(\xi_0(q, x)) = 0$ and for every fixed q, x the map Z is a C^∞ diffeomorphism in ξ.

Proof: Let ξ^* be the center of C and $A_0 = [2(t - t_0)]^{-1} X_\xi''(t_0, t, x, \xi^*)$. Let g be as in Lemma 8.1. Choose C_c small so that $[2(t - t_0)]^{-1} X_\xi''(t_0, t, x, \xi_0(q, x))$ is in the domain of g which is possible since $(t - t_0)^{-1} \nabla_\xi X_\xi''(t_0, t, x, \xi)$ is uniformly bounded in t_0, t, x, ξ. Define Z by

$$Z(t_0, t, q, x, \xi) = g(A, B(\xi))(\xi - \xi_0(q, x)), \tag{8.1}$$

$$B = B(\xi) = B(t_0, t, q, x, \xi)$$
$$= (t - t_0)^{-1}\int_0^1 d\theta\,(1-\theta)X''_\xi(t_0, t, x, \xi_0(q, x) + \theta(\xi - \xi_0(q, x))). \tag{8.2}$$

Then from (8.1), (8.2), and Lemma 8.1 the result follows, as in Lemma A.3 of Ref. 16. Q.E.D.

Lemma 8.3: Let Ω be an open subset of R^m and U: $\Omega \to R^m$ any C^∞, one–one map. If $U'(x)$ is nonsingular for every x, then $U(\Omega)$ is open and the inverse map V: $U(\Omega) \to \Omega$ given $U(V(y)) = y$, $V(U(x)) = x$ for y in $U(\Omega)$, and x in Ω is a C^∞ map.

Suppose further $\sup\{|D^\alpha U(x)|: x \in \Omega\} \leqslant K_\alpha < \infty$ for $|\alpha| \geqslant 1$ and $\sup\{|[U'(x)]^{-1}|: x \in \Omega\} \leqslant K_0^* < \infty$. Then $\sup\{|(D^\alpha V)(y)|: y \in U(\Omega)\} \leqslant L_\alpha < \infty$ for each $|\alpha| \geqslant 1$. Further L_α depends only on K_0^* and K_β for $|\beta| \leqslant \alpha$.

Proof: Obvious. Q.E.D.

Lemma 8.4: The map $\xi_0 = \xi_0(q, x)$, where $q = X'_\xi(t_0, t, x, \xi_0(q, x))$ is C^∞ in q, x.

Proof: Define U, V by $U(q, x, \xi) = (X'_\xi(t_0, t, x, \xi), x, q)$ and $V(q, x, y) = (y, x, \xi_0(q, x))$. Then it is clear that $U(V(q, x, y)) = (q, x, y)$. U is clearly C^∞ in q, x, y. Since X''_ξ is nonsingular, we get U' is nonsingular at every point. By Lemma 8.3 the map V is C^∞ proving the result. Q.E.D.

Lemma 8.5: (i) $\sup_{q,x} |D_q^\alpha(\xi_0(q, x))| \leqslant K_\alpha < \infty$, for each α, t_0, t.

(ii) $\sup_{q,x} |D_q^\alpha A(q, x)| + |D_q^\alpha A^{-1}(q, x)| \leqslant K_\alpha < \infty$ for each α, t_0, t, where A is as in Lemma 8.2.

(iii) $\sup_{q,x} |D_q^\alpha \nabla_x \xi_0(q, x)| \leqslant K_\alpha (t - t_0)^{-1}$ for each α, t_0, t.

By shrinking C_c, if necessary:

(iv) We can find a closed sphere $S \subset \mathrm{GL}(n, R)$ such that $(t - t_0)\nabla_x \xi_0(q, x)$ is in S for all q, x, t_0, t.

(v) $\sup_q |D_q^\alpha\{\xi_0(q, x_1) - \xi_0(q, x_2)\}| \leqslant K_\alpha |x_1 - x_2| (t - t_0)^{-1}$ for each α.

(vi) $\inf_q |\xi_0(q, x_1) - \xi_0(q, x_2)| \geqslant K_0^* |x_1 - x_2| (t - t_0)^{-1}$ for some $K_0^* > 0$.

Proof: (i) For each fixed x the map $\xi_0(\cdot, x)$ is the inverse of the map $X'_\xi(t_0, t, x, \cdot)$. Clearly, $|D_\xi^\alpha(X'_\xi(t_0, t, x, \xi))| \leqslant K_\alpha (t - t_0)$ for each $|\alpha| \geqslant 1$, q, ξ, t, t_0 and $\|[X''_\xi(t_0, t, x, \xi)]^{-1}\| \leqslant K^*(t - t_0)^{-1}$ for all x, ξ. The result now follows from the proof of second part Lemma 8.3 for $|\alpha| \geqslant 1$. Since ξ_0 is in C_c the result is trivially true for $\alpha = 0$.

(ii) Follows from (i), $|D_\xi^\alpha X(t_0, t, x, \xi)| \leqslant K_\alpha (t - t_0)$ for $|\alpha| \geqslant 2$ and the smoothness of the map $E \to E^{-1}$: $\mathrm{GL}(n, R) \to \mathrm{GL}(n, R)$.

(iii) Differentiating the equation $q = X'_\xi(t_0, t, x, \xi_0(q, x))$ w.r.t. x and simplifying, we get

$$2(t - t_0)\nabla_x \xi_0(q, x) = -A^{-1}(t_0, t, q, x)\nabla_x \nabla_\xi X(t_0, t, x, \xi_0(q, x)). \tag{8.3}$$

The result now follows from (i), (ii), $|D_\xi^\alpha \nabla_x \nabla_\xi X| \leqslant K_\alpha$ for all α, x, ξ, and induction on $|\alpha|$.

(iv) Follows from 8.3, $\|\nabla_x \nabla_\xi X(t_0, t, x, \xi) - I\| \leqslant \frac{1}{2}$ for all x, ξ, t_0, t and that A varies over a compact subset of $\mathrm{GL}(n, R)$.

(v) Follows from (iii) and the mean value theorem.

(vi) By (iv) and the fundamental theorem of calculus, we get

$$(t - t_0)[\xi_0(q, x_1) - \xi_0(q, x_2)] = f(q, x_1, x_2)(x_1 - x_2),$$

where f is in S so that $\|f^{-1}\| \leqslant K \leqslant \infty$ for all q, x_1, x_2. The result now easily follows. Q.E.D.

By Lemma 8.2 let $\psi = \psi(y) = \psi(t_0, t, q, x; y)$ be the inverse of $Z = Z(\xi) = Z(t_0, t, q, x, \xi)$ so that

$$\psi(Z(\xi)) = \xi \quad \text{for } \xi \text{ in } C_c$$

and

$$Z(\psi(y)) = y \quad \text{for } y \text{ near } 0.$$

Lemma 8.6: For each fixed x the function ψ is C^∞ in q and y.

Proof: For each fixed x define U, V by the rule $U(q, \xi) = (q, Z(q, x, \xi))$, $V(q, y) = (q, \psi(q, x, y))$. Then it is clear that U, V are inverses of each other. U is clearly C^∞ in q, ξ by (8.1), (8.2), and Lemmas 8.1 and 8.5. Further by (8.1) one gets

$$-I + \frac{\partial}{\partial\xi} Z(q, x, \xi) = \frac{\partial}{\partial B} g(A(q, x), B(q, x, \xi)) \frac{\partial}{\partial \xi} B(q, x, \xi) \times (\xi - \xi_0(q, x)) + g(A(q, x), B(q, x, \xi)) - I, \tag{8.4}$$

where I is the identity matrix. Now $g(A_0, A_0) = I$. So we can assume, by shrinking domain of g, if necessary (this will shrink C_c if done), that

$$\|g(\widetilde{A}, \widetilde{B}) - I\| \leqslant \tfrac{1}{4} \quad \text{for all } (\widetilde{A}, \widetilde{B}) \text{ in domain of } g. \tag{8.5}$$

By (8.2) we have $\|\partial B(q, x, \xi)/\partial\xi\| \leqslant K$ for all q, x, ξ under consideration. Since g is C^∞, we can have, by shrinking C_c if necessary.

$$\|\text{first term of rhs of (8.4)}\| \leqslant \tfrac{1}{4}. \tag{8.6}$$

From (8.4)–(8.6) we have

$$\left\| -I + \frac{\partial}{\partial\xi} Z(q, x, \xi) \right\| \leqslant \tfrac{1}{2}, \tag{8.7}$$

so that $\partial Z(q, x, \xi)/\partial\xi$ is nonsingular which in turn implies $U'(q, \xi)$ is nonsingular.

By Lemma 8.3 the map V is C^∞ and so the function ψ is C^∞ in q, y. Q.E.D.

Lemma 8.7: $|D_q^\alpha D_y^\beta \psi(t_0, t, q, x; y)|$ at $y = 0 \leqslant K_{\alpha\beta}$ for all $\alpha, \beta, t_0, t, q, x$.

Proof: Since $\psi(t_0, t, q, x; 0) = \xi_0$, the result is obvious when $\alpha + \beta = 0$. Let U, V be as in the proof of Lemma 8.6. It is easily seen that

$$[U'(q, \xi)]^{-1} = \begin{bmatrix} 1 & 0 \\ -\nabla_q Z(q, x, \xi)\{\nabla_\xi Z(q, x, \xi)\}^{-1} & \{\nabla_\xi Z(q, x, \xi)\}^{-1} \end{bmatrix} \tag{8.8}$$

Note that, by (8.7), $\nabla_\xi Z(q, x, \xi)$ is invertible and that the inverse is uniformly bounded in t_0, t, q, x, ξ. By (8.1), (8.2), and Lemma 8.5 it is verified that $\nabla_q Z(q, x, \xi)$ is uniformly bounded in q, x, ξ, t_0, t. Then $[U'(q, \xi)]^{-1}$ is uniformly bounded. Again by (8.1), (8.2), and Lemma 8.5 it is clear that $|D^{\alpha}_{q,\xi} U(q,\xi)| \leqslant K_\alpha$ for $|\alpha| \geqslant 1$ for all q, x, ξ, t_0, t. By applying Lemma 8.3 we get the result. Q.E.D.

ACKNOWLEDGMENT

I wish to thank the referee for his kind comments.

[1] W. O. Amrein, J. M. Jauch, and K. B. Sinha, *Scattering Theory in Quantum Mechanics* (Benjamin, Reading, MA, 1977).
[2] T. Kato, *Perturbation Theory for Linear Operators* (Springer-Verlag, Berlin, 1966).
[3] J. Weidmann, *Linear Operators in Hilbert Spaces* (Springer-Verlag, New York, Heidelberg, Berlin, 1980).
[4] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, III. Scattering Theory* (Academic, New York, 1979).
[5] Ref. 4, *Vol. IV. Analysis of Operators* (Academic, New York, 1978).
[6] V. Enss, "Asymptotic completeness for quantum Mechanical potential scattering. I. Short range potentials," Comm. Math. Phys. **61**, 285–291 (1978).
[7] E. B. Davies, "On Enss' approach to scattering theory," Duke Math. J. **47**, 171–185 (1980).
[8] E. Mourre, "Link between the geometrical and the spectral transformation approaches in scattering theory," Comm. Math. Phys. **68**, 91–94 (1979).
[9] P. A. Perry, "Mellin transforms and scattering theory, I. Short range potentials," Duke Math. J. **47**, 187–193 (1980).
[10] K. B. Sinha, private communications.
[11] B. Simon, "Phase space analysis of simple scattering systems: extensions of some work of Enss," Duke Math. J. **46**, 119–168 (1979).
[12] T. Umeda, "Scattering theory for Pseudo differential operators. I. The existence of wave operators," Osaka J. Math **18**, 361–392 (1981).
[13] T. Umeda, "Scattering theory for Pseudo-differential Operators. II. The completeness of wave operators," Osaka J. Math **19**, 511–526 (1982).
[14] P. Alsholm, "Wave operators for long range scattering," J. Math. Anal. Appl. **59**, 550 (1977).
[15] V. S. Buslaev and V. B. Matveev, "Wave operators for the Schrodinger equation with a slowly decreasing potential," Theoret. Math. Phys. **2**, 266–274 (1970).
[16] L. Hormander, "The existence of wave operators in scattering theory," Math. Z. **146**, 69–91 (1976).
[17] A. M. Berthier and P. Collet, "Wave operators for momentum dependent long range potential," Ann. Inst. H. Poicaré, Sec. A, **27**, 279–293 (1977).
[18] R. Lavine, "Absolute continuity of positive spectrum for Schrödinger Operators with long range potentials," J. Func. Anal. **12**, 30–54 (1973).
[19] (a) H. Kitada, "Scattering theory for Schrödinger operators with long range potentials II," J. Math. Soc. Jpn. **30**, 603–632 (1978); (b) Ikebe-Isazaki, "Completeness of modified wave operators for long range potentials," Publ. R.I.M.S., Kyoto Univ. **15**, 679–718 (1979).
[20] S. Agmon, "Some new results in spectral and scattering theory of differential operators in $L^2(R^n)$," lecture notes, Seminaire Goulaic-Schwartz 1978–79, Centre de Mathematiques—Paliseau.
[21] W. O. Amrein, Ph. A. Martin, and B. Misra, "On the asymptotic condition of scattering theory," Helv. Phys. Acta **43**, 313–344 (1970).
[22] L. E. Thomas, "On the algebraic theory of scattering," J. Func. Anal. **15**, 364–377 (1974).
[23] V. Enss, "Asymptotic completeness for quantum mechanical potential scattering—II. Singular and long range potentials," Ann. Phys. (N.Y.) **119**, 117–132 (1979).
[24] V. Enss, "Geometric methods in spectral and scattering theory of Schrödinger operators," in *Rigorous Atomic and Molecular Physics*," edited by G. Velo and A. Wightman (Plenum, New York, 1981), Sec. 7.
[25] P. A. Perry, "Propagation of states in dilation analytic potentials and asymptotic completeness," Comm. Math. Phys. **81**, 243–259 (1981).
[26] Pl. Muthuramalingam and K. B. Sinha, "Asymptotic evolution of certain observables and completeness in Coulomb scattering" J. Funct. Anal. (to be published).
[27] Pl. Muthuramalingam and K. B. Sinha, "Asymptotic completeness in long range scattering—II," Ann. Sci. Ecol. Norm. Sup. (submitted).
[28] Pl. Muthuramalingam, "Spectral and scattering theory for Schrödinger operator with a class of momentum dependent long range potentials," thesis, Indian Statistical Institute, New Delhi, 1981.
[29] Pl. Muthuramalingam, "Asymptotic completeness for the Coulomp potential," Technical Report, Indian Statistical Institute, New Delhi, July 1982.
[30] H. Kitada and K. Yajima, "A scattering theory for time dependent long range potentials," Duke Math. J. **49**, 341–376 (1982).
[31] J. Ginibre, "La Methode 'dependent du temps' dans le probleme de la completude asymptotique," Preprint LPTHE 80/10, Université Paris-Sud, 1980.
[32] E. B. Davies, *Quantum Theory of Open Systems* (Academic, New York, 1976).
[33] A. P. Calderon and R. Vaillan Court, "A class of bounded pseudo differential operators," Proc. Natl. Acad. Sci. USA **69**, 1185–1187 (1972).
[34] Pl. Muthuramalingam, Lectures on spectral properties of the (two-body) Schrödinger operator $-\frac{1}{2}\Delta + W(Q)$ on $L^2(R^n)$ using time dependent scattering theory in quantum mechanics, Indian Statistical Institute, New Delhi, June 1983.

Inverse scattering for optical couplers. Exact solution of Marchenko equations

G. P. Bava and G. Ghione
Dipartimento di Elettronica, Politecnico di Torino, Torino, Italy

(Received 30 November 1983; accepted for publication 13 January 1984)

An exact solution of Marchenko equations for rational scattering data of arbitrary order is developed, with reference to the case of two coupled waves propagating in a lossless (optical) waveguide. Some numerical examples are presented for the synthesis of coupling structures having a prescribed frequency response.

PACS numbers: 03.80. + r, 02.30.Rz, 42.80.Lt, 42.80.Sa

I. INTRODUCTION

The evolution of two propagating modes, coupled through some kind of nonuniform interaction, is very often described in terms of the Zakharov–Shabat (ZS) scattering problem, in the generalized form examined by Ablowitz *et al.*[1,2] In these cases, the existence of a suitable inversion technique allows us to synthesize the interaction parameters, so that some desired response is achieved. In the present paper such a synthesis procedure will be examined, with reference to the design of optical couplers.

The coupling between two electromagnetic modes propagating in an optical waveguide will be assumed to admit the following description.[3,4] Let a_i ($i = 1,2$) be the complex amplitudes of the modes having propagation constants β_i and let x be the axis of the guiding structure; we can write

$$\frac{da_1}{dx} + i\beta_1 a_1 = ih_{11}a_1 + h_{12}a_2, \tag{1a}$$

$$\frac{da_2}{dx} + i\beta_2 a_2 = ih_{22}a_2 + h_{21}a_1, \tag{1b}$$

where h_{ii} are the self-coupling coefficients and h_{ij} $(i\neq j)$ are the coupling coefficients. The coupling phenomenon is caused by some kind of waveguide perturbation, such as deformation of the boundaries, electro-optic effects, and so forth. Without loss of generality, it is convenient to assume the length x, the propagation constants, and the coupling coefficients to be dimensionless quantities; the corresponding actual parameters are xL, β_i/L, h_{ij}/L, respectively, where L is some typical length related to the problem (e.g., the center-band wavelength).

As it is well known from the coupled mode theory, Eqs. (1) give a satisfactory picture of the interaction if the two modes considered are near the phase-matching condition in the wavelength range of interest. Such a synchronous condition can be achieved, even when $\beta_1 \neq \beta_2$, by means of a periodic modulation of the coupling, which can be explicitly shown by means of the substitution:

$$h_{ij} = k_{ij}(x)e^{-i\beta_0 x} \quad (i\neq j). \tag{2}$$

The x-dependence of k_{ij} in (2) is included to account for slow phase or amplitude variations of the coupling coefficient. Thus, with $2k = \beta_1 - \beta_2 - \beta_0$ (synchronism parameter) and

$$c_j = a_j \exp\left(i\left\{(-1)^j kx + \beta_j x - \int_0^x h_{jj}\, dx'\right\}\right), \tag{3}$$

Eqs. (1) can be rewritten as

$$i\begin{pmatrix} \dfrac{d}{dx} & -Q \\ R & -\dfrac{d}{dx} \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = k\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}, \tag{4}$$

that is, in the ZS form, with

$$Q = k_{12}\exp\left\{i\int_0^x (h_{22} - h_{11})dx'\right\}, \tag{5a}$$

$$R = k_{21}\exp\left\{i\int_0^x (h_{11} - h_{22})dx'\right\}. \tag{5b}$$

It can be shown that for lossless, passive structures the relation $R = \mp Q^*$ holds, where * means complex conjugate and the upper (lower) sign applies when modes 1 and 2 carry power in the same direction (in opposite directions). From now on, the analysis will be confined to this case.

The inversion procedure for the ZS problem, via the solution of Marchenko integral equations, allows us to establish the characteristics of the coupling region (Q and R as functions of x) for a given set of scattering data (scattering parameters, to be discussed later on), thereby synthesizing a nonuniform coupler between two propagating modes. The design procedure based upon the inverse scattering technique is exact in principle, whereas the synthesis methods presented so far in the literature are either approximate or heavily relying on data obtained through analysis.[5–7]

In this paper, we discuss in detail the inversion problem; in particular, an exact procedure for the solution of Marchenko equations will be presented. This method is a generalization of the technique developed in Ref. 8, which applies to the simpler problem of a nonuniform lossless transmission line. We will not report details of the general ZS problem, since they can be found in the literature (see for instance Ref. 9).

II. INVERSE SCATTERING FOR OPTICAL COUPLERS

A. Interpretation of scattering data

The definition of the scattering data[9] and their connection with the physical problem under examination are more simply obtained by introducing the Jost solutions ϕ, $\bar{\phi}$, ψ, $\bar{\psi}$. As it is well known, these are solutions of (4) having the asymptotic behaviors

0022-2488/84/061900-05$02.50

$$x \to -\infty \begin{cases} \phi \to \begin{pmatrix}1\\0\end{pmatrix} e^{-ikx} \\ , \\ \bar{\phi} \to \begin{pmatrix}0\\1\end{pmatrix} e^{+ikx} \end{cases} \tag{6a}$$

$$x \to +\infty \begin{cases} \psi \to \begin{pmatrix}0\\1\end{pmatrix} e^{+ikx} \\ . \\ \bar{\psi} \to \begin{pmatrix}1\\0\end{pmatrix} e^{-ikx} \end{cases} \tag{6b}$$

The Jost solutions are linearly related[9] as

$$\bar{\psi} = \bar{a}\phi - b\bar{\phi}, \tag{7a}$$

$$\psi = a\bar{\phi} - \bar{b}\phi, \tag{7b}$$

with

$$a\bar{a} - b\bar{b} = 1 \tag{7c}$$

in a proper range of definition. If the structure is lossless, it can be shown that

$$\bar{a}(k) = a^*(k^*), \quad \bar{b}(k) = \mp b^*(k^*). \tag{8}$$

Now, two different cases can be considered according to whether the two interacting modes flow in opposite directions (A) or in the same direction (B). It is worth remarking that the physical direction of the power flow depends on the group velocity of the mode and has therefore no relation with the (apparent) propagation direction as it would be suggested by the asymptotic behavior (6) (cf. the definition of k).

Case A. In this case, we assume that the two modes travel as it is shown in Fig. 1(a). Equation (7a) suggests the scattering process depicted in Fig. 1(b): mode 1, traveling from the left, interacts with the coupling region, giving rise to a reflected wave (mode 2) with reflection coefficient $r = -b/\bar{a}$, and is transmitted to the right with transmission coefficient $t = 1/\bar{a}$.

Case B. In this case, we assume that the two modes travel in the same direction [Fig. 2(a)]. Again Eq. (7a) suggests the scattering process which is depicted in Fig. 2(b): mode 1 impinges from the right on the interaction region and is transmitted partly as mode 1, with transmission coefficient $t' = \bar{a}$, and partly as mode 2, with transmission coefficient $t'' = -b$; in this case $r = -b/\bar{a}$ is the partition coefficient.

For real k, (7c) and (8) yield, in case A

$$|r|^2 + |t|^2 = 1, \tag{9}$$

while, in case B

$$|t'|^2 + |t''|^2 = 1. \tag{10}$$

Equations (9) and (10) are statements of energy conservation.

The inverse scattering procedure for the ZS problem, given $r = -b/\bar{a}$ as a function of k (that is, the reflection coefficient for case A and the partition coefficient for case B), and assuming that the structure does not admit any bound states, leads[9] to the construction of the coupling coefficient $k_{12}(x)$ (through R). We must assume that, in the frequency range of interest, k_{12} does not depend on wavelength; this assumption represents a satisfactory approximation whenever narrow-band couplers are considered.

B. Inversion procedure: Marchenko equations

The inversion procedure is carried out according to the following steps[9]:

1. Given $\Gamma(k) = -\bar{b}(k)/a(k) = \pm b^*(k^*)/a^*(k^*) = \mp r^*(k^*)$, compute

$$F(x) = -\left(\frac{1}{2\pi}\right)\int_\gamma \frac{\bar{b}(k)}{a(k)} e^{-ikx}\, dk, \tag{11}$$

where the contour γ follows the real k axis apart from infinitesimal deformations encircling clockwise the real zeros of a (if any) in the upper k half-plane. $\Gamma(k)$ must be defined in a proper neighborhood of the reak k axis.

2. Solve the coupled Marchenko integral equations in the simplified form which holds when $R = \mp Q^*$:

$$\left.\begin{aligned} &E(x,y) \mp F^*(x+y) \\ &\qquad \mp \int_{-\infty}^{x} G(x,s)\, F^*(s+y)\, ds = 0 \qquad (12a) \\ &G(x,s) + \int_{-\infty}^{x} E(x,u)\, F(u+s)\, du = 0 \qquad (12b) \end{aligned}\right\} y,s \leqslant x$$

in the unknowns $E(x,y)$ and $G(x,s)$.

3. The coupling coefficient in the interaction region can now be obtained as

$$R(x) = 2E(x,x). \tag{13}$$

In Ref. 9 the conditions to be satisfied in order to achieve a solution to the inverse problem are thoroughly discussed. Here we mention the condition, that $r(k) = O[1/|k|]$ at least, $|k| \to \infty$. Moreover, in case A (discussed in the last section), we must have $|r(k)| \leqslant 1$ for real k. In case B, $r(k)$ is permitted to have real poles; in fact, according to (10), $r(k) = 0$ implies $t'' = 0$ (complete direct transmission), while at a pole of $r(k)$, we have $t' = 0$ (complete cross transmission). Finally, in case A, $r(k)$ is not allowed to have poles in the (open) lower k half-

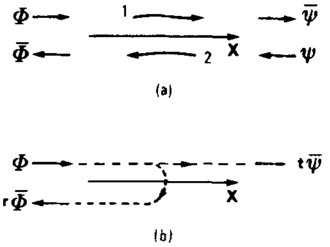


FIG. 1. Wave propagation for case A (a); scattering scheme and definition of r and t (b).

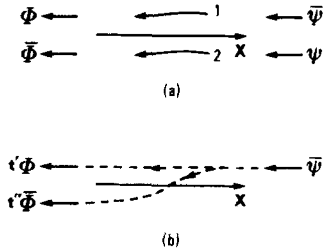


FIG. 2. Wave propagation for case B (a); scattering scheme and definition of t' and t'' (b).

plane [and, conversely, $\Gamma(k)$ must not have poles in the (open) upper k half-plane].

In the next section, a procedure is developed for solving Eqs. (12) exactly for the case, wherein the scattering data are represented by (or approximated with) a rational function.

III. SOLUTION OF THE PROBLEM

A. Exact solution of Marchenko equations

Let us approximate the desired $\Gamma(k)$ by [or let us assume that $\Gamma(k)$ is] a rational function of k,[8]

$$\Gamma(k) = N(k)/D(k), \tag{14}$$

where

$$D(k) = \prod_{m=1}^{n}(k - k_m), \quad \mathrm{Im}(k_m) \leqslant 0 \tag{15a}$$

and

$$N(k) = C\prod_{p=1}^{M}(k - K_p), \quad M < n. \tag{15b}$$

We assume that $\Gamma(k)$ has simple poles k_m. Then, if $u(x)$ is the step function ($u = 0$ for $x < 0$, $u = 1$ otherwise) we have from (11)

$$F(x) = F'(x)u(x), \tag{16a}$$

$$F'(x) = -i\sum_p C_p \exp(-ik_p x), \tag{16b}$$

$$C_p = N(k_p)\prod_{m\neq p}\frac{1}{k_p - k_m}. \tag{16c}$$

By using this expression for $F(x)$ in (12), it turns out that E and G must take the form

$$E(x,y) = E'(x,y)u(x+y), \tag{17a}$$

$$G(x,s) = G'(x,s)u(x+s). \tag{17b}$$

Thus, system (12) can be rewritten as follows:

$$E'(x,y) \mp F'^*(x+y) \mp \int_{-y}^{x} G'(x,s)\,F'^*(s+y)ds = 0, \tag{18a}$$

$$G'(x,s) + \int_{-s}^{x} E'(x,u)\,F'(u+s)du = 0. \tag{18b}$$

By a generalization of the items of Ref. 8, we shall assume that the solutions to (18) can be expanded in the form

$$E'(x,y) = \sum_\alpha f_\alpha(x)\exp(a_\alpha y), \tag{19a}$$

$$G'(x,s) = \sum_\alpha g_\alpha(x)\exp(b_\alpha s), \tag{19b}$$

where $f_\alpha, g_\alpha, a_\alpha, b_\alpha$, and the range of α are unknowns to be determined.

Substituting expressions (19) and (16) into (18) and carrying out the integrations we find that, if for every a_α a corresponding $b_\alpha = -a_\alpha$ is defined, the system can be decomposed into two subsystems:

$$\left.\begin{aligned} &f_\alpha(x) \pm i\sum_p \frac{C_p^*}{-a_\alpha + ik_p^*}g_\alpha(x) = 0 \quad &(20a)\\ &i\sum_p \frac{C_p}{a_\alpha + ik_p}f_\alpha(x) + g_\alpha(x) = 0 \quad &(20b)\end{aligned}\right\}\forall\alpha$$

$$\left.\begin{aligned} &\sum_\alpha \frac{\exp(-a_\alpha x)}{-a_\alpha + ik_p^*}g_\alpha(x) = -1 \quad &(21a)\\ &\sum_\alpha \frac{\exp(a_\alpha x)}{a_\alpha - ik_p}f_\alpha(x) = 0 \quad &(21b)\end{aligned}\right\}\forall p.$$

The homogeneous system (20) admits a nontrivial solution only if the corresponding determinant vanishes, i.e.,

$$\sum_{p,q}\frac{C_p^* C_q}{(-a_\alpha + ik_p^*)(a_\alpha - ik_q)} = \pm 1. \tag{22a}$$

This equation can be rewritten in a more compact form via the theorem of residues; one finds

$$r(-ia_\alpha)r^*(ia_\alpha^*) = \mp 1, \tag{22b}$$

which is an algebraic equation of order $2n$ in a_α giving $2n$ complex solutions for $\alpha = 1,2,\ldots,2n$. By taking the complex conjugate of (22b) it can be shown that, if a_α is a solution, also $-a_\alpha^*$ satisfies (22a). Then, the functions $f_\alpha(x)$ and $g_\alpha(x)$ are obtained by solving the linear system (21). Elimination of $g_\alpha(x)$ leads to the form

$$\left.\begin{aligned} &\sum_\alpha \frac{r^*(ia_\alpha^*)}{-a_\alpha + ik_p^*}\exp(-a_\alpha x)f_\alpha(x) = \mp 1 \quad &(23a)\\ &\sum_\alpha \frac{1}{a_\alpha - ik_p}\exp(+a_\alpha x)f_\alpha(x) = 0 \quad &(23b)\end{aligned}\right\}\forall p.$$

Equations (22) and (23) give a complete solution to our problem; the coupling coefficient can now be written as

$$R(x) = 2E(x,x) = 2u(x)\sum_\alpha f_\alpha(x)\exp(a_\alpha x). \tag{24}$$

Although this expression for $R(x)$ has been obtained under seemingly restrictive assumptions upon the structure of $E(x,y)$ and $G(x,y)$, a theorem of existence and uniqueness for the solution of Marchenko equations,[9] holding for scattering data of the form assumed here, ensures that (24) is the unique solution to the inverse scattering problem under examination.

B. Some properties

We summarize here some characteristics of system (4) and properties of the solutions of (22) and (23), which are relevant to the physical interpretation of the problem at hand.

1. If the condition $r(-k^*) = r^*(k)$ holds (that is, if r is real on the imaginary k axis) Eq. (22b) takes the form

$$r(-ia_\alpha)r(ia_\alpha) = \mp 1. \tag{25}$$

Therefore, if a_α is a solution, $-a_\alpha$ is also a solution. But it has already been shown that, if a_β is a solution, also $-a_\beta^*$ is a root of (22); it follows that, under the aforementioned condition, (25) admits solutions which are imaginary ($\pm i|a_\alpha|$) and/or quadrantally symmetrical ($\pm a_\alpha, \pm a_\alpha^*$). Moreover, if the reality condition $r(-k^*) = r^*(k)$ is satisfied, inspection of system (23) reveals that f_α corresponding to pairs of complex conjugate roots are complex conjugate; therefore, it follows from (24) that the coupling coefficient $R(x)$ is real. It is worth noticing that the converse of this property is also true, as it can be easily deduced from some general properties of Jost solutions (cf., e.g., Ref. 9, Sec. 6.1).

2. If the degree of $D(k)$ exceeds the degree of $N(k)$ by

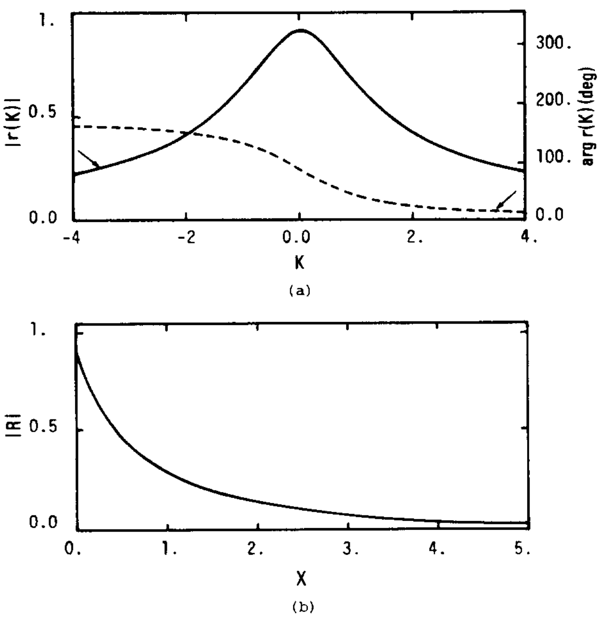


FIG. 3. Single-pole response (a) and corresponding $|R|$ (b) for case A: $C = 0.95$; $\eta = 1$.

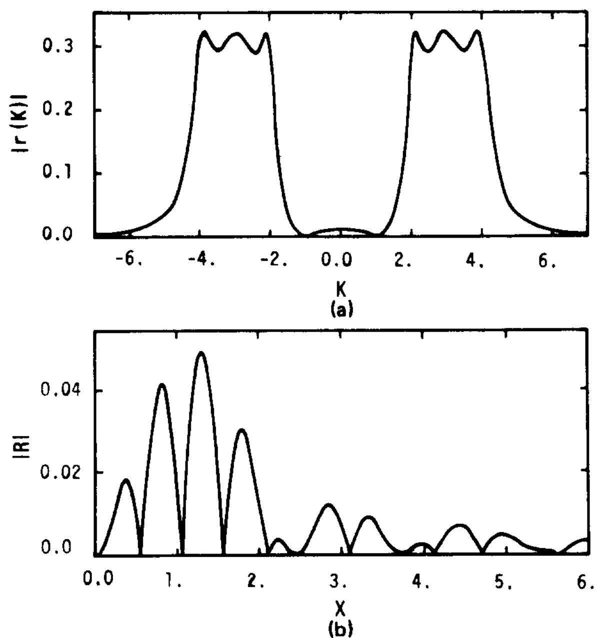


FIG. 5. Reflection coefficient with symmetrical passband and equiripple lateral stopbands (a) and corresponding $|R|$ (b). Poles of $r(k)$: $k_1 = 2. + i0.2$, $k_2 = 2.9 + i0.55$, $k_3 = 4. + i0.3$, $k_4 = -k_1^*$, $k_5 = -k_2^*$, $k_6 = -k_3^*$; zeros of $r(k)$: $K_1 = 1.$, $K_2 = -1.$; $C = 1$.

more than one, $R(x)$ [given by (24)] is continuous in $x = 0$. In fact, $R(x)$ is continuous in $x = 0$ only if $E'(0,0) = 0$, which implies in turn [from (18)] that $F'(0) = 0$. From (16), this can be shown to require that the sum of the residues of all the poles of $r(k)$ must vanish. This amounts to the property stated above.

3. If the assigned scattering data $r(k)$ are related to the coupling coefficient $R(x)$, then $r(k - k_0)$ leads to $R(x)\exp(ik_0x)$. Hence, the scattering data for an optical coupler can always be assumed to be assigned on a wavelength band centered around the origin $k = 0$. A shift of the central wavelength can be achieved with a different choice of the waveguiding structure (that is, with a different $\beta_1 - \beta_2$) or by introducing a proper periodicity in the coupling mechanism (that is, by changing β_0).

IV. EXAMPLES

A. Single-pole scattering data

Owing to the impossibility of solving analytically algebraic equations of order higher than four, a fully analytical treatment of the inversion technique previously developed is fraught with difficulties, unless the degree of the rational function representing the scattering data is one or two. In particular, the case wherein $r(k)$ has a single pole can be easi-

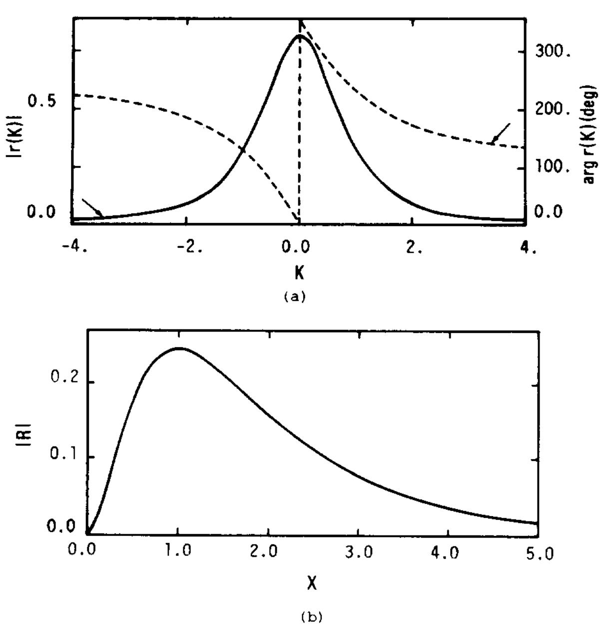


FIG. 4. Three-pole response (a) and corresponding $|R|$ (b) for case A. Poles of $r(k)$: $k_1 = i$, $k_2 = 1.1i$, $k_3 = 1.11i$; $C = i$.

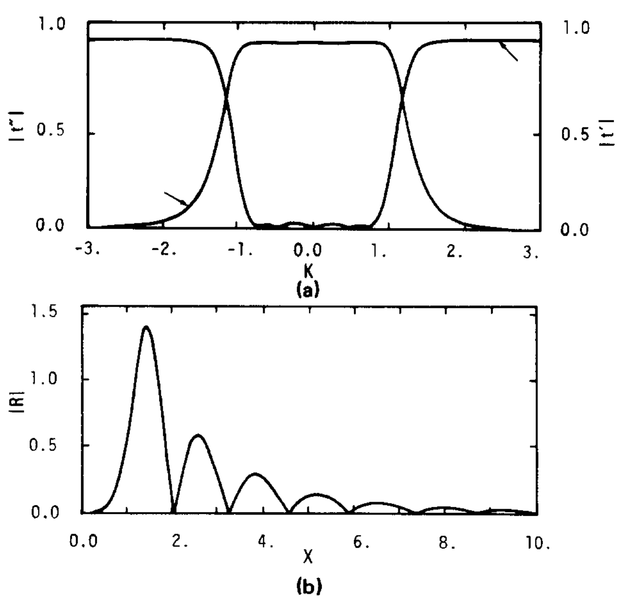


FIG. 6. Direct ($|t'|$) and cross ($|t''|$) transmission coefficients (a) and corresponding $|R|$ (b) for case B. Poles of $r(k)$: $k_1 = 0$, $k_2 = 0.5$, $k_3 = 0.75$, $k_4 = -0.5$, $k_5 = -0.75$; $C = i$.

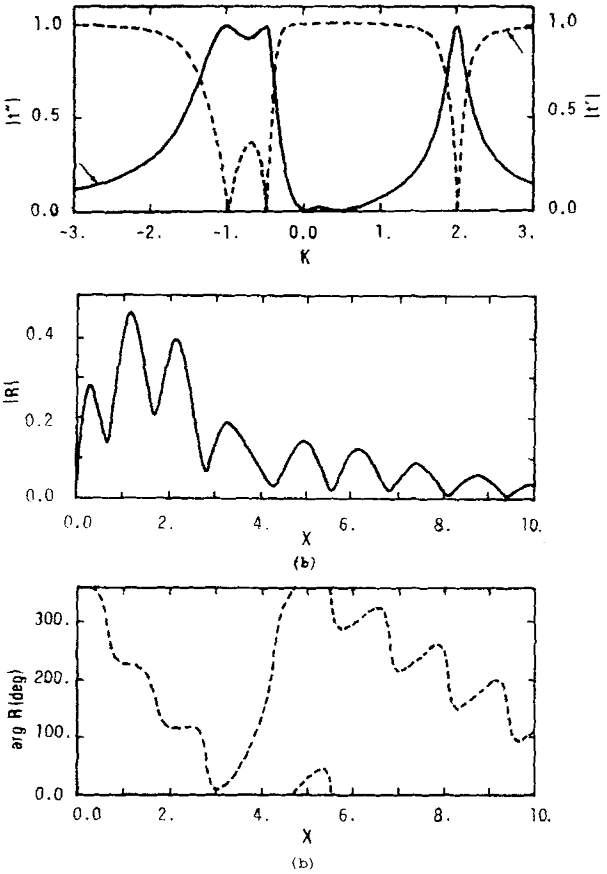


FIG. 7. Direct ($|t'|$) and cross ($|t''|$) transmission coefficients (a) and corresponding R (b) for case B. Poles of $r(k)$: $k_1 = -0.5$, $k_2 = -1.$, $k_3 = 2.$, $k_4 = 2.i$; zeroes of $r(k)$: $K_1 = 0.$, $K_2 = 0.5$; $C = 1$.

ly handled without having to resort to numerical treatment of Eqs. (22) and (23). We assume

$$r(k) = C/(k - k_1), \quad \mathrm{Im}(k_1) \geqslant 0, \tag{26}$$

where we can always assume Re $k_1 = 0$ (cf. Sec. IIIB, property 3). Therefore

$$|r|^2 = \frac{|C|^2}{k^2 + \eta^2}, \quad k \text{ real}, \; k_1 = i\eta \tag{27}$$

has a Lorentzian shape centered around the origin. In case A, we must have $|r| \leqslant 1$, which leads to the condition $|C/\eta| \leqslant 1$. The roots of Eq. (22) are

$$a_1 = (\pm |C|^2 + \eta^2)^{1/2}, \quad a_2 = -a_1, \tag{28}$$

where the upper and the lower sign refer to the cases B and A, respectively; both roots are always real. From (23) and (24) we obtain

$$R(x) = \pm 2iC^* u(x) \frac{a_1 \exp(2a_1 x)}{a_1 \cosh 2a_1 x - \eta \sinh 2a_1 x}. \tag{29}$$

A typical behavior of $r(k)$ and the corresponding $R(x)$ are shown in Fig. 3 for case A.

B. Higher-order scattering data

When scattering data (14) are considered of order higher than two, a numerical treatment of (22) becomes indispensable, and it is also convenient to solve system (23) numerically.

In Fig. 4, an example concerning a reflection coefficient $r(k)$ (case A) having three poles on the imaginary k-axis is reported. As expected, the related coupling coefficient is continuous at the origin. A more complex example of scattering data (reflection coefficient) and of the corresponding $R(x)$ is given in Fig. 5. Owing to the symmetry of $r(k)$ the resulting $R(x)$ is real (positive or negative) in both cases.

Figure 6 shows the amplitudes of the direct and cross transmission coefficients t' and t'' (case B) and the related coupling coefficient, for a symmetrical equiripple stopband–passband coupler. Again, the symmetry of the characteristics leads to real $R(x)$. Conversely, in Fig. 7 an example is reported, wherein asymmetrical assigned t' and t'' lead to a complex coupling coefficient, whose amplitude and phase are shown in Fig. 7b.

V. CONCLUSIONS

In the present paper an exact technique for the solution of Marchenko equations has been presented, with the aim of obtaining a suitable procedure for the design of narrow-band optical couplers. Some preliminary examples of application for the inversion technique are presented, also starting from scattering data of realistic shape and practical interest. However, further considerations are needed in order to analyze the feasibility of the implementation of the coupling coefficients thereby obtained in real structures.

[1]V. E. Zakharov and A. B. Shabat, Zh. Eksp. Teor. Fiz. **61**, 118 (1971) [Sov. Phys. JETP **34**, 62 (1972)].
[2]M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Stud. Appl. Math. **53**, 294 (1974).
[3]D. Marcuse, *Theory of Dielectric Optical Waveguides* (Academic, New York, 1974).
[4]D. Marcuse, Bell Syst. Tech. J. **54**, 985 (1975).
[5]H. Kogelnik, Bell Syst. Tech. J. **55**, 109 (1976).
[6]M. G. F. Wilson and G. A. Teh, IEEE Trans. Microwave Theory Tech. **MTT-23**, 85 (1975).
[7]M. Matsuhara, K. O. Hill, and A. Watanabe, J. Opt. Soc. Am. **65**, 804 (1975).
[8]K. R. Pechenick and J. M. Cohen, J. Math. Phys. **22**, 1513 (1981).
[9]R. K. Dodd, J. K. Eilbeck, J. D. Gibbon and H. C. Morris, *Solitons and Nonlinear Wave Equations* (Academic, New York, 1982).

The time-dependent inverse source problem for the acoustic and electromagnetic equations in the one- and three-dimensional cases[a)]

H. E. Moses
University of Lowell, Center for Atmospheric Research, Lowell, Massachusetts 01854

(Received 23 February 1983; accepted for publication 2 September 1983)

The object of the time-dependent inverse source problem of electromagnetic theory and acoustics is to find time-dependent sources and currents, which are turned on at a given time and then off to give rise to prescribed radiation fields. In an early paper for the three-dimensional electromagnetic case, the present writer showed that the sources and currents are not unique and gave conditions which make them so. The ideas of that paper are reformulated for the three-dimensional electromagnetic case and extended to the acoustical three-dimensional case and the one-dimensional electromagnetic and acoustic cases. The one-dimensional cases show very explicitly the nature of the ambiguity of the choice of sources and currents. This ambiguity is closely related to one which occurs in inverse scattering theory. The ambiguity in inverse scattering theory arises when one wishes to obtain the off-shell elements of the T matrix from some of the on-shell elements (i.e., from the corresponding elements of the scattering operator). In inverse scattering theory prescribing of the representation in which the potential is to be diagonal removes the ambiguity. For the inverse source problem a partial prescription of the time dependence of the sources and currents removes the ambiguity. The inverse source problem is then solved explicitly for this prescribed time dependence. The direct source problems for the one- and three-dimensional acoustic and electromagnetic cases are also given to provide a contrast with the inverse source problem and for use in later papers. Moreover, the present author's earlier work on the eigenfunctions of the curl operator is reviewed and used to simplify drastically the three-dimensional direct and inverse source problems for electromagnetic theory by splitting off the radiation field and its currents from the longitudinal field and its sources and currents. Finally, for a prescribed time dependence, the inverse source problem is solved explicitly in closed form. Using methods developed for the inverse source problem for three-dimensional electromagnetic theory, we solve the following important direct problem. Consider a sphere of finite radius and uniform charge density rotating about a fixed axis. Assume that the angular velocity is zero before a certain time, varies in an arbitrary fashion, and then becomes zero again. What is the final electromagnetic field?

PACS numbers: 03.80. + r

1. INTRODUCTION

We are concerned with the time-dependent acoustic and electromagnetic equations in one and three dimensions. In one dimension the equation is (taking the velocity $c = 1$ for simplicity)

$$\left[-\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial t^2} \right] f(x;t) = \rho(x;t), \tag{1.1}$$

where $f(x;t)$ is the overpressure or electric field in the acoustic and electromagnetic cases, respectively, and $\rho(x;t)$ is the corresponding source. The three-dimensional acoustic and electromagnetic equations are

$$\left[-\nabla^2 + \frac{\partial^2}{\partial t^2} \right] f(\mathbf{x};t) = 4\pi\rho(\mathbf{x};t) \tag{1.2}$$

and

$$\begin{aligned} &\nabla\times\mathbf{H}(\mathbf{x};t) = 4\pi\,\mathbf{j}(\mathbf{x};t) + \frac{\partial}{\partial t}\mathbf{E}(\mathbf{x};t), \\ &\nabla\times\mathbf{E}(\mathbf{x};t) = -\frac{\partial}{\partial t}\mathbf{H}(\mathbf{x};t), \\ &\nabla\cdot\mathbf{E}(\mathbf{x};t) = 4\pi\rho(\mathbf{x};t), \\ &\nabla\cdot\mathbf{H}(\mathbf{x};t) = 0, \end{aligned} \tag{1.3}$$

respectively, using obvious notation for the various physical quantities. The direct and inverse source problems can be treated in two distinct but obviously related approaches. They are the time-dependent and the time-independent approaches. We shall discuss the time-independent approach first.

A. Time-independent approach

In the time-independent approach, in which the inverse source problem has been treated most extensively, a Fourier transform with respect to time is taken for all physical quantities and the resultant time-independent equations are studied for a fixed frequency. For example, let us consider the acoustic equation (1.2) and write

$$\begin{aligned} f(\mathbf{x};t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(\mathbf{x},k)e^{ikt}\,dk, \\ \rho(\mathbf{x};t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} S(\mathbf{x},k)e^{ikt}\,dk, \end{aligned} \tag{1.4}$$

and thereby obtain the time-independent version of Eq. (1.2) as

$$[\nabla^2 + k^2]F(\mathbf{x};k) = 4\pi S(\mathbf{x},k). \tag{1.5}$$

[a)] The work in this paper was done under the National Science Foundation Grant MCS-8024640.

The direct problem is concerned with finding the solution of Eq. (1.5) under certain boundary conditions, often taken to be

$$\lim_{r\to\infty} F(\mathbf{x},k) = \frac{e^{ikr}}{r} A(\theta,\phi), \tag{1.5'}$$

where r,θ,ϕ are the spherical polar coordinates of $\mathbf{x}$, if the space in which Eq. (1.5) holds extends to infinity. If one has the solution, then, in particular, the amplitude of the outgoing spherical wave $A(\theta,\phi)$ would be known. Thus in the time-independent direct problem, one gives the sources $S(\mathbf{x},k)$ and seeks the radiation pattern given by $A(\theta,\phi)$.

In the inverse problem the radiation pattern $A(\theta,\phi)$ is prescribed and one wishes to find the source $S(\mathbf{x},k)$. In Ref. 1, for example, the inverse problem is treated in some detail. The solution is shown to be nonunique, but additional conditions can be given to make them so and a procedure for constructing the source $S(\mathbf{x},t)$ in terms of spherical harmonics is given. In Ref. 2 the same author discusses the electromagnetic analog. Reference 3 summarizes the work of others for the three-dimensional acoustic and electromagnetic theories, mostly in terms of the time-independent approach. The nonuniqueness of the inverse problem is stressed, and some ways of adding conditions which lead to unique solutions are discussed.

An important element of the time-independent approach is that one can prescribe the condition, one which is useful for practical realizations, that the source function $S(\mathbf{x},k)$ vanish outside a sphere $r=a$.

We shall now consider the time-dependent approach to the direct and inverse problems with sources.

B. Time-dependent approach

The time-dependent approach to the direct and inverse source problems is complementary to that for the time-independent approach. As we have discussed above, the time-independent approach relates the radiation pattern to the sources for a fixed frequency. The time-dependent approach, as we formulate it for this paper, is concerned with pulses. We assume that the sources are nonzero only for a finite interval of time $t_0 < t < t_1$. In the direct problem we are concerned with the initial value problem in which the field is prescribed as a solution of the homogeneous equation *before* the sources are turned on. One then solves the equation for times after the sources are turned on, using appropriate Riemann functions. After the sources are turned off, one again has a solution of the homogeneous equation which is determined by the initial solution for $t<t_0$ and by the known time-dependent sources.

The inverse problem, as we formulate it, is the following: Let us prescribe t_0 and t_1 and the solutions of the homogeneous equations for $t<t_0$ and $t>t_1$. We want to find the sources as functions of time and space which will lead us to the prescribed solution for $t>t_1$ from the initial solution for $t<t_0$. It is by no means clear *a priori* that *any* sources can be found for arbitrary initial and final solutions of the homogeneous equation. However, we shall show below that there are actually an infinite number of sources which satisfy our conditions. Moreover, we shall give conditions on the time dependence which will lead to unique sources as far as the space dependence is concerned. Using a particular form for the time dependence will lead to closed-form expressions for the source and currents.

The problem of finding the sources from initial and final fields and a description of the time dependence of the sources is closely analogous to the problem of finding the potential from appropriate elements of the scattering operator in inverse scattering theory in one and three dimensions (Ref. 4). The portion of the scattering operator used in inverse scattering gives some of the on-shell elements of the T matrix. One is required to find all of the off-shell elements, since the potential can then be found from them. The process of finding the off-shell elements is not unique. To make them unique, we must impose conditions on the potential which give the representation in which it is diagonal. In practice we require the potential to be diagonal in the x representation. In the case of the inverse source problem, that part of the source or current which is given by the initial and final solutions of the homogeneous equation is analogous to the on-shell components of the T matrix which one prescribes. Giving a partial description of the time dependence of the source is analogous to prescribing the representation in which the potential is to be diagonal.

As far as he can determine, the author of the present paper is the first and apparently still the only one who has worked in the time-dependent context described above. In Ref. 5, he solved the inverse source problem for Maxwell's equations (1.3), which are the most complicated equations of Eqs. (1.1)–(1.3). The methods found to work for that case can be adapted to Eqs. (1.1) and (1.2). In Ref. 5, when we discovered that the sources for the inverse problem were not unique, it occurred to us to assume that the sources were separable functions of space and time, i.e.,

$$\mathbf{j}(\mathbf{x};t) = \mathbf{j}(\mathbf{x})h(t). \tag{1.6}$$

This way of specifying the time dependence turned out to be not suitable if one imposed the condition that the sources be real. (This condition is analogous to requiring the potential to be real in the corresponding inverse scattering problem, as discussed in Ref. 4.) Thus one cannot impose the time dependence arbitrarily. A time-dependence condition which *was* satisfactory was to require

$$\mathbf{j}(\mathbf{x};t) = \mathbf{j}_e(\mathbf{x})h_e(t) + \mathbf{j}_o(\mathbf{x})h_o(t), \tag{1.7}$$

where $h_e(t)$ was a real, essentially arbitrary, function of t which was symmetric about the midpoint of the interval $[t_0,t_1]$ and $h_o(t)$ was similarly a real odd function about the midpoint. (The subscripts e and o stand for "even" and "odd," respectively.) We then showed that the inverse source problem gave $\mathbf{j}_e(\mathbf{x})$ and $\mathbf{j}_o(\mathbf{x})$ uniquely.

It should be mentioned that Bleistein and Cohen (Ref. 6) recently independently discovered the nonuniqueness of the solution by combining time-dependent and time-independent approaches and proposed the time-dependence equation (1.6), which will generally, however, not yield real currents and sources, as we have noted above.

C. Further discussion and synopsis of the paper

We shall now outline the rest of the paper. We shall treat Eqs. (1.1)–(1.3) in order of difficulty, which is the same as the order of the equation numbers, though the most difficult—Eq. (1.3), the electromagnetic case—was the first to be discussed by the author (Ref. 5).

For each of the equations we shall treat the direct problem first in order to clarify the notions for the inverse source problem. The importance of causality is evident from the direct problems. Because of causality it is possible to formulate the inverse source problem even for finite (rather than infinite) domains for which Eqs. (1.1)–(1.3) are valid. If there are reflectors or changes of indices of refraction, it is only after the radiation fields reach them that they become effective. For example, for some purposes one might want to shape an electromagnetic pulse within a wave guide *before* the pulse strikes the walls. This will be possible, provided the sources are also within the walls.

The discussions of the direct source problem for Eqs. (1.1) and (1.2), the one-dimensional equation and the three-dimensional acoustic equation, will be recognized as being restatements of the classical Cauchy problem as discussed, for example, in Ref. 7. For Maxwell's equations (1.3), however, our treatment will differ markedly from the usual treatments and also from our earlier treatment of Ref. 5. For Eq. (1.3) we are concerned with the transverse fields only and only in those portions of the currents which give rise to them. In our original treatment of Ref. 5, we rewrote Maxwell's equations in the form of Dirac's equations and used the eigenfunctions of the Dirac-like Hamiltonian to split off the transverse parts of the field and the corresponding currents. By contrast, in the present paper we shall use the eigenfunctions of the curl operator for the separation of the fields and currents.

We introduced the eigenfunctions of the curl operator in Ref. 8 as a means of treating electromagnetic theory, fluid dynamics, and other phenomena obeying vector field equations in the infinite domain. In Ref. 8 we discussed the situation for Maxwell's equations somewhat tersely and showed that the eigenfunctions separate the radiation from the longitudinal fields. In the present paper we shall review the properties of the eigenfunctions of the curl operator in both Cartesian and spherical polar coordinates and apply them first to the direct source problem and then the inverse source problem. It will be seen that the direct and inverse source problems for Eq. (1.3) differ little in difficulty from those of the one-dimensional problem.

After showing the ambiguity in the inverse problem, we shall prove that the specification of the time dependence of the sources in the form analogous to Eq. (1.7) and the requirement that the sources be real lead to unique solutions of the inverse problem in the sense that $\mathbf{j}_o(\mathbf{x})$ and $\mathbf{j}_e(\mathbf{x})$ or their analogs are unique.

For all of the inverse source problems treated in the present paper we shall give concrete examples. The existence of such examples assures us that our procedure is usable and not empty formalism.

In the time-independent version of the inverse source problem the sources are required to be contained in a given finite volume, i.e., have compact support. Among our examples are some in which the sources are contained in a finite volume but also some in which the sources go to infinity with various rates of decay. From the examples the extent of localization of the sources will be seen to depend on the time functions $h_o(t)$ and $h_e(t)$. One of the objects of future research will be to see what assumptions about time dependence, other than that given by Eq. (1.7), will prove useful and also give a prescribed localization of sources.

D. Applications

Before we leave the Introduction, it may be useful to discuss possible applications. There are obviously many of them, and we shall mention a few. The principal use of transmission lines is to propagate pulses of a particular shape. The inverse source problem tells us how to turn the sources on and off to achieve the desired pulse shape. The lack of uniqueness is now a positive asset, since engineering constraints can now be taken into account. For purposes of sounding in three dimensions, as, for example, the use of sonar to locate underwater objects or ionosondes to probe the ionosphere, it is useful to shape the incident pulses in a particular way. The present paper suggests methods by which this pulse shaping can be accomplished.

The accurate control of pulse shape could also be used to construct pulse shapes for secure communications or to extend the amount of information contained in a pulse. More generally, a pulse can be modified or even deleted by appropriate choices of the sources.

A problem which can be treated by extensions of the present work is the *source detection problem*. Here one wishes to characterize the source from the pulses which it emits. For example, one might want to obtain an estimate of the physical size or shape of the emitter. The additional information which is needed, besides the pulse shape, for an essentially unique determination of the source could be the subject of further research. Undoubtedly causality plays an essential role. On the other hand, if any pulsed source emits a radiating wave, we shall show that from this radiating wave alone, one can always construct a pulsed source (indeed in many ways) which will reproduce this wave. Thus one could, for example, manufacture sounds, which would seem to come from a source, but actually come from a source much smaller or at a different position. In this manner one can make up decoys to simulate the presence of a supposed object.

2. THE ONE-DIMENSIONAL PROBLEM

A. The direct problem

Here we are concerned with Eq. (1.1). We assume that the source $\rho(x;t)\equiv 0$ for $t < t_0$ and $t > t_1$ and for simplicity without loss of generality let $t_0 = -T$ and $t_1 = T$. We shall now introduce notation, which for the present case may seem overly elaborate. However, since we wish to stress the similarity of the treatment of the one-dimensional problem with that of the acoustic and electromagnetic problem in three dimensions, we find this notation very useful. Accordingly, we define

$$\chi(x|p) = (1/\sqrt{2\pi})e^{ipx} . \tag{2.1}$$

We also define

$$\gamma(p;t) = \int_{-\infty}^{+\infty} \chi^*(x|p)\rho(x;t)\,dx , \tag{2.2}$$

$$\psi(p,\lambda;t) = e^{-i\lambda\nu(t+T)}\psi(p,\lambda;-T) + \frac{i\lambda}{2\nu} e^{-i\lambda\nu t}\int_{-T}^{t} \gamma(p;t')e^{i\lambda\nu t'}\,dt' , \tag{2.3}$$

where

$$\lambda = \pm 1 , \quad -\infty < p < \infty , \nu = |p| . \tag{2.3'}$$

The following is easily shown. Let $f(x;t)$ be given by

$$f(x;t) = \sum_\lambda \int_{-\infty}^{+\infty} \chi(x|p)\psi(p,\lambda;t)\,dp . \tag{2.4}$$

The apparent divergence for $p \to 0$ in Eq. (2.3) is eliminated in Eq. (2.4). Then $f(x;t)$ satisfies Eq. (1.1).

Moreover, on defining $f_\pm(x;t)$ by

$$f(x;t) = f_-(x;t) \quad \text{for } t < -T , \qquad f(x;t) = f_+(x;t) \quad \text{for } t > T , \tag{2.4a}$$

we see that $f_\pm(x;t)$ are solutions of Eq. (1.1) without sources. These solutions are given explicitly by

$$f_\pm(x;t) = \sum_\lambda \int_{-\infty}^{+\infty} \chi(x|p)e^{-i\lambda\nu(t\mp T)}\psi(p,\lambda;\pm T)\,dp . \tag{2.4b}$$

We are now in a position to discuss the direct and inverse source problem. In the direct problem, one prescribes $f_-(x;t)$ and the source $\rho(x;t)$. One wishes to find $f_+(x;t)$. It is seen that the problem reduces to a quadrature.

In the inverse source problem, by contrast, one prescribes the initial function $f_-(x;t)$ and the final function $f_+(x;t)$. One then looks for sources $\rho(x;t)$ which, when used in the direct problem, would give the prescribed $f_+(x;t)$ when the initial wave function $f_-(x;t)$ was used.

The functions $f_\pm(x;t)$ can themselves be obtained in terms of $f_\pm(x;\pm T)$ and $(\partial/\partial t)f_\pm(x;\pm T)$. This is, in fact, the usual statement of the initial and final value problems for the source-free equation. Then in the direct problem we prescribe $f_-(x;-T)$ and the source function $\rho(x;t)$. We then seek $f_+(x;T)$ and $(\partial/\partial t)f_+(x;T)$. In the inverse problem we prescribe $f_\pm(x;\pm T)$ and $(\partial/\partial t)f_\pm(x;\pm T)$ and want the source function $\rho(x;t)$.

We shall now show how $f_+(x;t)$ is obtained from $f_+(x;T)$ and $(\partial/\partial t)f_+(x;T)$ through the use of suitable Riemann functions. Of course, the function $f_-(x;t)$ can be treated analogously. From Eq. (2.4b)

$$f_+(x;T) = \sum_\lambda \int_{-\infty}^{+\infty} \chi(x|p)\psi(p,\lambda;T)\,dp , \tag{2.5}$$

$$\frac{\partial}{\partial t}f_+(x;T) = -i\sum_\lambda \lambda \int_{-\infty}^{+\infty} \chi(x|p)\nu\psi(p,\lambda;T)\,dp . \tag{2.5'}$$

Taking Fourier transforms leads to

$$\psi(p,+1;T) + \psi(p,-1;T) = \int_{-\infty}^{+\infty} \chi^*(x|p)f_+(x;T)\,dx ,$$
$$\psi(p,+1;T) - \psi(p,-1;T) = \frac{i}{\nu}\int_{-\infty}^{+\infty} \chi^*(x|p)\frac{\partial}{\partial t}f_+(x;T)\,dx . \tag{2.6}$$

From Eq. (2.6) we obtain $\psi(p,\lambda;T)$ in terms of $f_+(x;T)$ and $(\partial/\partial t)f_+(x;T)$:

$$\psi(p,\lambda;T) = \frac{1}{2}\left[\int_{-\infty}^{+\infty} \chi^*(x|p)f_+(x;T)\,dx + i\frac{\lambda}{\nu}\int_{-\infty}^{+\infty} \chi^*(x|p)\frac{\partial}{\partial t}f_+(x;T)\,dx\right] . \tag{2.7}$$

Thus from Eq. (2.4b)

$$f_+(x;t) = \int_{-\infty}^{+\infty} dx'\, G_1(x-x';t-T)\frac{\partial}{\partial t}f_+(x';T) + \int_{-\infty}^{+\infty} dx'\, G_2(x-x';t-T)f_+(x';T) , \tag{2.8}$$

where

$$G_1(x;t) = \frac{i}{4\pi}\sum_\lambda \int_{-\infty}^{+\infty} e^{ipx}e^{-i\lambda\nu t}\frac{\lambda}{\nu}\,dp = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{\sin pt}{p}e^{ipt}\,dp . \tag{2.9}$$

In Eq. (2.9) we have used the fact that $(\sin \nu t/\nu) = (\sin pt/p)$.

Likewise $G_2(x;t)$ is given by

$$G_2(x;t) = \frac{1}{4\pi}\sum_\lambda \int_{-\infty}^{+\infty} e^{ipx}e^{-i\lambda\nu t}\,dp = \frac{\partial}{\partial t}G_1(x;t) . \tag{2.10}$$

We shall now evaluate the expressions Eq. (2.9) in detail to show how the usual Riemann function makes its appearance. This derivation serves as a prototype for the analogous derivations of the three-dimensional problems, which we shall not give in detail.

The integrand of the second integral in Eq. (2.9) contains no singularities on the real axis. We deform the contour to consist of the entire real axis with a semicircular indentation into the negative imaginary half-plane about the point $p = 0$. Then

$$G_1(x;t) = \frac{1}{4\pi i}\left[\int_{-\infty}^{+\infty} \frac{e^{ip(x+t)}}{p}\,dp - \int_{-\infty}^{+\infty} \frac{e^{ip(x-t)}}{p}\,dp\right] . \tag{2.11}$$

The integrals are now readily evaluated by closing the contours at infinity. Thus

$$G_1(x;t) = \tfrac{1}{2}[\eta(x+t) - \eta(x-t)] , \tag{2.12}$$

where $\eta(x)$ is the Heaviside function defined by $\eta(x) = 1$ for $x > 0$ and $\eta(x) = 0$ for $x < 0$. We shall now simplify the expression Eq. (2.12). Let $t > 0$ and $x > 0$. Then since $x + t > 0$ and $1 - \eta(x) = \eta(-x)$, we have

$$G_1(x;t) = \tfrac{1}{2}\,\eta(t-x) = \tfrac{1}{2}\,\eta(|t| - |x|) . \tag{2.13}$$

Similarly for $t > 0$ and $x < 0$, since $x - t < 0$

$$G_1(x;t) = \tfrac{1}{2}\,\eta(x+t) = \tfrac{1}{2}\,\eta(|t| - |x|) , \tag{2.14}$$

as before. From similar considerations for $t < 0$

$$G_1(x;t) = -\tfrac{1}{2}\eta(|t| - |x|). \tag{2.15}$$

Then for all t and x

$$G_1(x;t) = \tfrac{1}{2}\,\mathrm{sgn}\,t\,\eta(|t| - |x|). \tag{2.16}$$

If a is any number $a > 0$, $\eta(ax) = \eta(x)$. Then since $|x| + |t| > 0$,

$$G_1(x;t) = \tfrac{1}{2}\,\mathrm{sgn}\,t\,\eta[(|t| + |x|)(|t| - |x|)]$$
$$= \tfrac{1}{2}\,\mathrm{sgn}\,t\,\eta(t^2 - x^2). \tag{2.17}$$

This is the usual invariant form.

From Eqs. (2.10) and (2.17)

$$G_2(x;t) = |t|\,\delta(t^2 - x^2). \tag{2.18}$$

The analog of Eq. (2.8) for $f_-(x;t)$ is

$$f_-(x;t) = \int_{-\infty}^{+\infty} dx'\, G_1(x - x';t + T)\frac{\partial}{\partial t}f_-(x'; -T) + \int_{-\infty}^{+\infty} dx'\, G_2(x - x';t + T)f_-(x'; -T). \tag{2.8a}$$

Similarly from Eq. (2.4) we obtain $f(x;t)$ in terms of the source $\rho(x;t)$ as follows:

$$f(x;t) = f_-(x;t) + \int_{-\infty}^{+\infty} dx' \int_{-T}^{t} dt'\, G_1(x - x';t - t')\rho(x';t'). \tag{2.19}$$

In particular, we have

$$f_+(x;t) = f_-(x;t) + \int_{-\infty}^{+\infty} dx' \int_{-T}^{+T} G_1(x - x';t - t')\rho(x';t'). \tag{2.20}$$

Equation (2.20) somewhat resembles the relationship in scattering theory in which the wave function for $t \to +\infty$ is obtained from the wave function for $t \to -\infty$ through the use of the scattering operator (see, e.g., Ref. 9). This analog becomes even closer if we write Eq. (2.20) in terms of $\psi(p,\lambda;T)$ and $\psi(p,\lambda,-T)$, for this relationship becomes [see Eq. (2.3)]

$$\psi(p,\lambda;T) = e^{-2i\lambda\nu T}\psi(p,\lambda;-T) + \frac{i\lambda}{2\nu}e^{-i\nu T}\int_{-T}^{+T}\gamma(p;t')e^{i\lambda\nu t'}\,dt'. \tag{2.21}$$

B. The inverse problem

We are now in a position to discuss the inverse source problem for the one-dimensional case. In the inverse problem we wish to specify $f_\pm(x;t)$ and T and obtain the source function $\rho(x,t)$. We note that if we specify T, we would get the same results for T_1, if $T_1 > T$, since the interval $-T < t < +T$ nests inside the interval $-T_1 < t < +T_1$ and, of course, the sources would vanish outside the larger time interval, also. Hence, specifying one time interval outside of which the source vanishes we specify all the larger intervals also.

Specification of $f_\pm(x;t)$ and T is equivalent from Eq. (2.4b) and Eq. (2.6) to specifying T and $\psi(p,\lambda;\pm T)$. We then wish to obtain $\gamma(p;t)$ which would enable us to find $\rho(x;t)$.

Let us define $F(p;k)$ by

$$F(p;k) = \frac{1}{\sqrt{2\pi}}\int_{-T}^{+T}\gamma(p;t')e^{ikt'}\,dt'. \tag{2.22}$$

If we knew $F(p;k)$, we would know $\rho(x;t)$ through

$$\rho(x;t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} dp \int_{-\infty}^{+\infty} dk\, \chi(x|p)e^{-ikt}F(p;k). \tag{2.23}$$

But from our assumptions we do not know $F(p,k)$ for all p and k. We only know $F(p;k)$ for $k = \lambda\nu$: from Eq. (2.21)

$$G(p,\lambda) \equiv F(p;\lambda\nu) = -\frac{2i\lambda\nu}{\sqrt{2\pi}}\,[e^{-i\lambda\nu T}\psi(p,\lambda;T) - e^{-i\lambda\nu T}\psi(p,\lambda;-T)]. \tag{2.24}$$

Knowing $F(p;k)$ only on the shell $k = \lambda\nu$ is analogous to knowing the T matrix of scattering theory only on the energy shell. In the present case, as in scattering theory, the on-shell information leads to nonunique sources $\rho(x;t)$, since *any* function $F(p;k)$ which has values on the shell given by Eq. (2.24) will yield suitable functions for the sources. In scattering theory the ambiguity is removed by prescribing the representation in which the potential is diagonal. We shall show that by giving the time dependence of $\rho(x;t)$ we obtain unique answers for the sources.

But first we must examine consequences of the conditions that $f(x;t)$ and $\rho(x;t)$ shall be real for all t and x. These conditions limit the kind of time dependence which one can have.

The reality condition for $f(x;t)$ leads to the requirement that

$$\psi^*(p,\lambda;t) = \psi(-p,-\lambda;t). \tag{2.25}$$

The reality of $\rho(x;t)$ is equivalent to the condition

$$F^*(p;k) = F(-p;-k). \tag{2.26}$$

Equations (2.25) and (2.26) are necessary and sufficient conditions. Both of these equations lead to the following condition on $G(p,\lambda)$:

$$G^*(p,\lambda) = G(-p,-\lambda). \tag{2.27}$$

To get unique results for $\rho(x;t)$, one is tempted to use separation of variables in the form

$$\rho(x;t) = \rho(x)h(t), \tag{2.28}$$

where $\rho(x)$ and $h(t)$ are real functions of their arguments. Indeed, we made an analogous assumption in our three-dimensional electromagnetic inverse source problem of Ref. 5. However, we soon saw that the reality conditions were violated. Our next assumption, one that worked, would in the present context be

$$\rho(x;t) = \rho_e(x)h_e(t) + \rho_o(x)h_o(t), \tag{2.29}$$

where $h_e(t)$ is a real even function of t and $h_o(t)$ is a real odd function of t. The functions $\rho_e(x)$ and $\rho_o(x)$ are real functions. We prescribe $h_e(t)$ and $h_o(t)$ and obtain $\rho_e(x)$ and $\rho_o(x)$ from our known function $G(p,\lambda)$, which is obtained from Eq. (2.24). We require, of course, that $h_{e,o}(t) \equiv 0$ when $t < -T$ or $t > +T$.

We shall see shortly that neither $h_e(t)$ nor $h_o(t)$ can be assumed to be identically zero. However, it can happen that either $\rho_o(x)$ or $\rho_e(x)$ can be identically zero. Another interesting and important aspect of the assumption of Eq. (2.29) is that if $h_e(t)$ and $h_o(t)$ are replaced, respectively, by $Ah_e(t)$ and $Bh_o(t)$, where A and B are arbitrary, nonzero constants, $\rho(x,t)$ will be independent of A and B, since $\rho_e(x)$ and $\rho_o(x)$ will be replaced by $\rho_e(x)/A$ and $\rho_o(x)/B$, respectively. Thus (roughly speaking) the *shapes*, not the *sizes* of the time-dependent functions are the important quantities.

Let us define

$$g_e(k) = \frac{1}{\sqrt{2\pi}} \int_{-T}^{+T} h_e(t) e^{ikt}\, dt ,$$

$$g_o(k) = \frac{1}{\sqrt{2\pi}} \int_{-T}^{+T} h_o(t) e^{ikt}\, dt . \tag{2.30}$$

The $g_e(k)$ is a *real, even* function of k and $g_o(k)$ is an *imaginary, odd* function of k. It follows that

$$g_e(\lambda\nu) = g_e(\nu), \quad g_o(\lambda\nu) = \lambda g_o(\nu) . \tag{2.31}$$

From Eq. (2.22)

$$G(p,\lambda) \equiv F(p;\lambda\nu) = F_e(p) g_e(\nu) + \lambda F_o(p) g_o(\nu), \tag{2.32}$$

where $F_e(p)$ and $F_o(p)$ are the Fourier transforms of $\rho_e(x)$ and $\rho_o(x)$, respectively, i.e.,

$$\rho_e(x) = \int_{-\infty}^{+\infty} \chi(x|\,p) F_e(p)\, dp ,$$

$$\rho_o(x) = \int_{-\infty}^{+\infty} \chi(x|\,p) F_o(p)\, dp . \tag{2.33}$$

For real $\rho_e(x)$ and $\rho_o(x)$ we shall need

$$F^*_{e,o}(p) = F_{e,o}(-p) . \tag{2.34}$$

We can now state our objective explicitly: from $\psi(p,\lambda, \pm T)$, $h_{e,o}(t)$, find $F_{e,o}(p)$ which satisfy Eq. (2.34) and thus real $\rho_{e,o}(x)$. From Eq. (2.32) we obtain unique solutions for $F_{e,o}(p)$.

$$F_e(p) = \frac{1}{2g_e(\nu)} \sum_\lambda G(p,\lambda), \quad F_o(p) = \frac{1}{2g_o(\nu)} \sum_\lambda \lambda G(p,\lambda) . \tag{2.35}$$

Our procedure for solving the inverse problem is thus to obtain $G(p,\lambda)$ in terms of $\psi(p,\lambda; \pm T)$ through the use of Eq. (2.24), then obtain $F_{e,o}(p)$ from Eq. (2.35) and finally $\rho_{e,o}(x)$ from Eq. (2.33). By having carefully taken care of the reality conditions, our results for $\rho_{e,o}(x)$ are real and unique.

Most of the present discussion including uniqueness is close to that for the three-dimensional acoustic and electromagnetic cases. When these cases are considered, some abridgement of the discussion for these latter cases will be made. We see, as mentioned previously, that we may not take either $h_e(t)$ or $h_o(t)$ to be zero, for then $g_e(k)$ or $g_o(k)$ would be zero. Nevertheless, F_e or F_o (and hence ρ_e or ρ_o) can be identically zero. We shall give examples of this situation shortly.

It is possible that $g_e(k)$ or $g_o(k)$ may have real zeros. This case can be treated as a generalization of the following example.

Assume $g_e(k)$ has a simple zero at $k = a > 0$. We multiply and divide the right-hand side of the first of Eq. (2.35) by $p^2 - a^2$. Then $\rho_e(x)$ will then be represented as a convolution of the Fourier transforms of $(p^2 - a^2)F_e(p)$ and $(p^2 - a^2)^{-1}$ defined as a principal value at the singularity. The Fourier transform of $(p^2 - a^2)^{-1}$ is a distribution with well-known properties. The matter of zeros in $g_{e,o}(k)$ as well as consequences on the analytic properties of these functions arising from the Wiener–Paley theorem as applied to functions $h_{e,o}(t)$ will be explored in later papers.

C. An example

For most of the examples of the present paper we shall use

$$h_e(t) = \delta(t) \quad \text{or} \quad h_e(t) = \delta''(t) \quad \text{and} \quad h_o(t) = \delta'(t) . \tag{2.36}$$

Thus from Eq. (2.30)

$$g_e(k) = 1/\sqrt{2\pi} \quad \text{or} \quad g_e(k) = -k^2/\sqrt{2\pi}$$
$$\text{and} \quad g_o(k) = ik/\sqrt{2\pi} . \tag{2.36'}$$

These functions have the appropriate dependence on k and also satisfy the requirements that g_e be real and g_o be imaginary. Moreover, though T can be any positive number, for simplicity in the expressions we shall let $T \to 0$.

We shall assume for the present purpose that $f_-(x,t) \equiv 0$. Hence

$$\psi(p,\lambda;0_-) \equiv 0 . \tag{2.37}$$

Thus we need only specify $f_+(x,t)$ or, equivalently, $\psi(p,\lambda;0_+)$. For our example, let us consider the function $F(x)$ given by

$$F(x) = \sin \kappa x \quad \text{for} \quad -a < x < a,$$
$$\equiv 0 \quad \text{elsewhere} \quad (\kappa a = n\pi, \ n \text{ any positive integer}) . \tag{2.38}$$

Then we choose

$$f_+(x;t) = F(x - t) . \tag{2.39}$$

Our example is thus a pulse of finite length, containing an integer number of wavelength $L = (2\pi/\kappa)$ moving in the positive x direction. [The restriction that n be a positive integer can be removed. One gets slightly more complicated expressions for the source $\rho(x;t)$.] This pulse shape is one of those most often used in transmission line theory. We can now give the pulsed source which can give rise to this pulsed wave.

We have

$$f_+(x;0_+) = \sin \kappa x \quad \text{for} \quad -a < x < a$$
$$\equiv 0 \quad \text{elsewhere},$$

$$\frac{\partial}{\partial t} f(x;0_+) = -\kappa \cos \kappa x \quad \text{for} \quad -a < x < a ,$$
$$\equiv 0 \quad \text{elsewhere} . \tag{2.40}$$

From Eq. (2.7)

$$\psi(p,\lambda;0_+) = -\frac{i}{2\nu} \frac{1}{\sqrt{2\pi}} \left[(\kappa\lambda + \nu) \frac{\sin(p-\kappa)a}{(p-\kappa)} + (\kappa\lambda - \nu) \frac{\sin(p+\kappa)a}{(p+\kappa)} \right] . \tag{2.41}$$

We see that $\psi(p,\lambda;0_+)$ satisfies condition Eq. (2.25), which is necessary and sufficient for $f_+(x;t)$ to be real for all time t.

On using $h_e(t)=\delta(t)$, $h_o(t)=\delta'(t)$, Eq. (2.24) for $G(p,\lambda)$, the appropriate expressions for $g_{e,o}(k)$ from Eq. (2.36'), we obtain finally from Eq. (2.35)

$$F_e(p) = -\frac{\kappa}{2}\frac{1}{\sqrt{2\pi}}\left[\frac{\sin(p-\kappa)a}{(p-\kappa)}+\frac{\sin(p+\kappa)a}{(p+\kappa)}\right],$$
$$F_o(p) = \frac{i}{2}\frac{1}{\sqrt{2\pi}}\left[\frac{\sin(p-\kappa)a}{(p-\kappa)}-\frac{\sin(p+\kappa)a}{(p+\kappa)}\right]. \tag{2.42}$$

Then $\rho(x;t)$ is given by Eqs. (2.29) and (2.33):

$$\rho(x;t) = -\tfrac{1}{2}[\kappa\delta(t)\cos\kappa x+\delta'(t)\sin\kappa x] \quad \text{for } -a<x<a\,,$$
$$\equiv 0 \quad \text{elsewhere}. \tag{2.43}$$

In this example, the source is contained in a finite region of x space—in fact the same region for which $f_+(x;0_+)$ is not identically zero. The finiteness of the support upon which the source function is not zero is important if one wishes to construct the source physically. Moreover, the fact that the size of the support is the same as that for $f_+(x;0_+)$ means that the dimensions of the source are reasonable. In higher dimensions it will be shown in a specific example that the choices of $h_e(t)$ and $h_o(t)$ are crucial in determining whether the support of the source is finite.

Equation (2.43) can be generalized considerably. Assume

$$f_+(x;0_+)=A(x),\quad \frac{\partial}{\partial t}f_+(x;0_+)=B(x). \tag{2.44}$$

Then, as can be shown using the previous analysis

$$\rho(x;t)=B(x)\delta(t)-A(x)\delta'(t). \tag{2.45}$$

The consequences for the compactness of the support of the density in space can now be read off in an obvious manner from the initial conditions (2.44).

3. THE THREE-DIMENSIONAL ACOUSTIC PROBLEM

The inverse source problem for the three-dimensional acoustic equation (1.2) is not essentially more difficult than the one-dimensional problem. However, the functions which satisfy the homogeneous equation, i.e., when the source is turned off, form a much richer class than those of the one-dimensional problem. Thus we can set up the inverse problem in two ways, corresponding to the use of a coordinate-independent formalism or of spherical polar coordinates. But first it behooves us to discuss the direct problem.

A. The direct problem

Let $\chi(\mathbf{x}|\mathbf{p})$ and $\psi(\mathbf{p},\lambda;t)$ be defined by

$$\chi(\mathbf{x}|\mathbf{p})=\frac{1}{(2\pi)^{3/2}}e^{i\mathbf{p}\cdot\mathbf{x}}, \tag{3.1}$$

$$\psi(\mathbf{p},\lambda;t)=e^{-i\lambda p(t+T)}\psi(\mathbf{p},\lambda;-T)+\frac{i\lambda}{2p}e^{-i\lambda pt}\int_{-T}^{t}\gamma(\mathbf{p};t')\times e^{-i\lambda pt'}\,dt', \tag{3.2}$$

where

$$p=|\mathbf{p}|,\quad \lambda=\pm 1 \tag{3.2'}$$

and

$$4\pi\rho(\mathbf{x};t)=\int\chi(\mathbf{x}|p)\gamma(\mathbf{p};t)\,d\mathbf{p},$$
$$\gamma(\mathbf{p};t)=4\pi\int\chi^*(\mathbf{x}|\mathbf{p})\rho(\mathbf{x};t)\,d\mathbf{x}. \tag{3.3}$$

Then on defining $f(\mathbf{x};t)$ by

$$f(\mathbf{x};t)=\sum_\lambda\int\chi(\mathbf{x}|\mathbf{p})\psi(\mathbf{p},\lambda;t)\,d\mathbf{p}, \tag{3.4}$$

it is seen that $f(x;t)$ satisfies Eq. (1.2) with

$$f(\mathbf{x};t)=f_-(\mathbf{x};t)\quad\text{for } t\leqslant -T,$$
$$f(\mathbf{x};t)=f_+(\mathbf{x};t)\quad\text{for } t>T, \tag{3.5}$$

where

$$f_\pm(\mathbf{x};t)=\sum_\lambda\int\chi(\mathbf{x}|\mathbf{p})e^{-i\lambda p(t\mp T)}\psi(\mathbf{p},\lambda;\pm T)\,d\mathbf{p} \tag{3.6}$$

and, of course, are solutions of the homogeneous equation corresponding to Eq. (1.2).

The quantities $\psi(\mathbf{p},\lambda;\pm T)$ can be obtained from $f_\pm(\mathbf{x};\pm T)\equiv f(\mathbf{x};\pm T)$ and $(\partial/\partial t)f_\pm(\mathbf{x};\pm T)$ from the following expressions [cf. Eq. (2.7) and its derivation]:

$$\psi(\mathbf{p},\lambda;\pm T)=\frac{1}{2}\left[\int\chi^*(\mathbf{x}|\mathbf{p})f_\pm(\mathbf{x};\pm T)\,d\mathbf{x}+i\frac{\lambda}{p}\int\chi^*(\mathbf{x}|\mathbf{p})\frac{\partial}{\partial t}f_\pm(\mathbf{x};\pm T)\,d\mathbf{x}\right]. \tag{3.7}$$

The functions $f_\pm(\mathbf{x};t)$ can be found from $f_\pm(\mathbf{x};\pm T)$ using influence functions as in Eq. (2.8). One has

$$f_\pm(\mathbf{x};t)=\int d\mathbf{x}'\,G_1(\mathbf{x}-\mathbf{x}';t\mp T)\frac{\partial}{\partial t}f_\pm(\mathbf{x}';\pm T)+\int d\mathbf{x}'\,G_2(\mathbf{x}-\mathbf{x}';t\mp T)f_\pm(\mathbf{x};\pm T), \tag{3.8}$$

where $G_1(\mathbf{x};t)$ and $G_2(\mathbf{x};t)$ are given by

$$G_1(\mathbf{x};t)=-\frac{1}{4\pi r}[\delta(t+r)-\delta(t-r)]$$
$$\equiv\frac{1}{2\pi}\operatorname{sgn} t\,\delta(t^2-r^2),$$
$$G_2(\mathbf{x};t)=\frac{\partial}{\partial t}G_1(\mathbf{x};t)=\frac{1}{\pi}|t|\delta'(t^2-r^2). \tag{3.9}$$

In Eq. (3.9), $r\equiv|\mathbf{x}|$.

In complete analogy to the one-dimensional equation (2.19), we can express $f(\mathbf{x};t)$ in terms of $f_-(\mathbf{x};t)$ and $\rho(\mathbf{x};t)$ and influence functions as

$$f(\mathbf{x};t)=f_-(\mathbf{x};t)+4\pi\int d\mathbf{x}'\int_{-T}^{t}dt'\,G_1(\mathbf{x}-\mathbf{x}';t-t')\rho(\mathbf{x}';t'). \tag{3.10}$$

Thus for $t>T$

$$f_+(\mathbf{x};t)=f_-(\mathbf{x};t)+4\pi\int d\mathbf{x}'\int_{-T}^{+T}dt'\,G_1(\mathbf{x}-\mathbf{x}';t-t')\rho(\mathbf{x}',t')\,. \tag{3.11}$$

As for the one-dimensional situation, the direct source problem is concerned with obtaining $f_+(\mathbf{x};t)$ from the initial conditions as given by $f_-(\mathbf{x};t)$, or equivalently by $f_-(\mathbf{x};-T)$ and $(\partial/\partial t)f_-(\mathbf{x};-T)$, and the source $\rho(\mathbf{x};t)$. In the inverse problem, one prescribes $f_\pm(\mathbf{x};t)$ and seeks to find $\rho(\mathbf{x};t)$.

B. The inverse problem. Coordinate-independent version

As in the direct problem for the three-dimensional case, the inverse problem uses its one-dimensional analog as its prototype.

Let us define $F(\mathbf{p};k)$ by

$$F(\mathbf{p};k)=\frac{1}{\sqrt{2\pi}}\int_{-T}^{+T}\gamma(\mathbf{p};t')e^{ikt'}\,dt'\,. \tag{3.12}$$

Of course, if we knew $F(\mathbf{p};k)$, we should know $\rho(\mathbf{x};t)$, since

$$4\pi\rho(\mathbf{x};t)=\frac{1}{\sqrt{2\pi}}\int d\mathbf{p}\int_{-\infty}^{+\infty}dk\,\chi(\mathbf{x}|\mathbf{p})e^{-ikt}F(\mathbf{p};k)\,. \tag{3.13}$$

From the information given in the inverse scattering problem we know $\psi(\mathbf{p},\lambda;\pm T)$ and hence $G(\mathbf{p},\lambda)$ defined by

$$G(\mathbf{p},\lambda)\equiv F(\mathbf{p};\lambda p)=-\frac{2i\lambda p}{\sqrt{2\pi}}\,[e^{i\lambda pT}\psi(\mathbf{p},\lambda;T)-e^{-i\lambda pT}\psi(\mathbf{p},\lambda;-T)]\,. \tag{3.14}$$

The requirement that the functions $f(\mathbf{x};t)$ and $\rho(\mathbf{x};t)$ be real lead to necessary and sufficient conditions on $\psi(\mathbf{p},\lambda;t)$ and $F(\mathbf{p};k)$, namely,

$$\psi^*(\mathbf{p},\lambda;t)=\psi(-\mathbf{p},-\lambda;t)\,,$$
$$F^*(\mathbf{p};k)=F(-\mathbf{p};-k)\,. \tag{3.15}$$

Both conditions lead to the requirement that

$$G^*(\mathbf{p},\lambda)=G(-\mathbf{p},-\lambda)\,. \tag{3.16}$$

As in the one-dimensional case we make the assumption

$$\rho(\mathbf{x};t)=\rho_e(\mathbf{x})h_e(t)+\rho_o(\mathbf{x})h_o(t)\,, \tag{3.17}$$

where the properties of the given functions $h_{e,o}(t)$ and their Fourier transforms $g_{e,o}(k)$ are the same as those for the one-dimensional problem.

If we now define $F_{e,o}(\mathbf{p})$ by

$$F_{e,o}(\mathbf{p})=\int\chi^*(\mathbf{x}|\mathbf{p})\rho_{e,o}(\mathbf{x})\,d\mathbf{x}$$

or (3.18)

$$\rho_{e,o}(\mathbf{x})=\int\chi(\mathbf{x}|\mathbf{p})F_{e,o}(\mathbf{p})\,d\mathbf{p}\,,$$

we have as the equation analogous to Eq. (2.35)

$$F_e(\mathbf{p})=\frac{1}{2g_e(p)}\sum_\lambda G(\mathbf{p},\lambda)\,,\quad F_o(\mathbf{p})=\frac{1}{2g_o(p)}\sum_\lambda\lambda G(\mathbf{p},\lambda)\,. \tag{3.19}$$

Thus to find $\rho_{e,o}(\mathbf{x})$, one specifies $\psi(\mathbf{p},\lambda;\pm T)$, or equivalently $f(\mathbf{x};\pm T)$ and $(\partial/\partial t)f(\mathbf{x};\pm T)$, calculates $G(\mathbf{p},\lambda)$ obtains $F_{e,o}(\mathbf{p})$ from Eq. (3.19) and finally $\rho_{e,o}(\mathbf{x})$ through the use of the second of Eq. (3.18).

C. An example

As an example we take $f_-(\mathbf{x};t)\equiv 0$, which is equivalent to taking $\psi(\mathbf{p},\lambda;-T)\equiv 0$. Moreover, we use $h_e(t)=\delta(t)$, $h_o(t)=\delta'(t)$. As before, $T\to 0$. We specify $f_+(\mathbf{x},t)$ by giving the Cauchy data as

$$f_+(\mathbf{x},0_+)\equiv 0\,,\quad \frac{\partial}{\partial t}f_+(\mathbf{x},0_+)=A\delta(\mathbf{x})\,. \tag{3.20}$$

We obtain for $\psi(\mathbf{p},\lambda;0_+)$

$$\psi(\mathbf{p},\lambda;0_+)=\frac{1}{(2\pi)^{3/2}}\frac{i\lambda}{p} \tag{3.21}$$

on using Eq. (3.7). The function $f_+(\mathbf{x};t)$ can now be found from Eq. (3.6).

$$f_+(\mathbf{x};t)=AG_1(\mathbf{x};t)\,, \tag{3.22}$$

where $G_1(\mathbf{x};t)$ is given by the first of Eqs. (3.9). We find from Eq. (3.14)

$$G(\mathbf{p},\lambda)=\frac{1}{\sqrt{2\pi}}\frac{A}{(2\pi)^{3/2}}\,. \tag{3.23}$$

Thus from Eq. (3.19) for all $g_o(k)$ and hence all $h_o(t)$

$$F_o(\mathbf{p})\equiv 0\quad\text{and hence}\quad \rho_o(\mathbf{x})\equiv 0\,. \tag{3.24}$$

Therefore, the present example illustrates the fact that while neither $h_e(t)$ or $h_o(t)$ can be, in general, taken to be identically zero, nevertheless, the choice of $f_\pm(\mathbf{x};t)$ may lead to either $\rho_e(\mathbf{x})$ or $\rho_o(\mathbf{x})$ being identically zero, in which case, of course, one does not have to specify $h_e(t)$ or $h_o(t)$. To complete the example, we shall now compute $\rho_e(\mathbf{x})$ which from Eq. (3.19) depends on the choice of $h_e(t)$. Let us first take $h_e(t)=\delta(t)$. Then

$$4\pi\rho_e(\mathbf{x})=A\delta(\mathbf{x})\,. \tag{3.25}$$

Thus

$$4\pi\rho(\mathbf{x};t)=A\delta(\mathbf{x})\delta(t)\,. \tag{3.26}$$

We could have anticipated the result Eqs. (3.25) and (3.26) for our choice of $h_e(t)$ because our function $f(\mathbf{x};t)$ is just the influence function with the initial condition (3.20) which must satisfy the three-dimensional wave equation (1.2) with the source given by Eq. (3.26). Nevertheless, the example is important in showing the inner consistency of the inverse method and in illustrating a case in which one of the functions $\rho_e(\mathbf{x})$ or $\rho_o(\mathbf{x})$ are identically zero.

This example can also be used to illustrate the effect of the choices of $h_e(t)$ and $h_o(t)$ on the localization of the source for given functions $f_\pm(\mathbf{x};t)$.

The source given in Eqs. (3.25) and (3.26) is strongly localized. Instead of choosing $h_e(t)$ to be $\delta(t)$, let us take $h_e(t)=\delta''(t)$.

In this case

$$4\pi\rho(\mathbf{x};t) = -\frac{A}{4\pi r}\delta''(t), \tag{3.27}$$

and the source, while still concentrated at $r=0$, now extends to infinity.

The example leading to Eq. (3.26) can be generalized to arbitrary final wave forms for $h_e(t)=\delta(t)$, $h_o(t)=\delta'(t)$. Let

$$f_+(\mathbf{x},0_+) = A(\mathbf{x}), \quad \frac{\partial}{\partial t} f_+(\mathbf{x},0_+) = B(\mathbf{x}). \tag{3.28}$$

Then

$$4\pi\rho(\mathbf{x};t) = B(\mathbf{x})\delta(t) - A(\mathbf{x})\delta'(t). \tag{3.29}$$

Equations (3.28) and (3.29) are, of course, analogous to Eqs. (2.44) and (2.45).

D. The inverse method in terms of spherical polar coordinates

Up to now we have discussed the direct and inverse source problems in what is a coordinate-independent formalism. For many purposes, however, spherical polar coordinates are very useful, especially when one wishes to study the radiated pulse at some distance from the source. We shall approach the problem by introducing spherical coordinates in **p**-space. For this purpose we use spherical harmonics $Y_{JM}(\theta,\phi)$ in the notation of Ref. 10, for example.

The completeness and orthonormality relations of these functions are

$$\int_0^{2\pi} d\phi \int_0^{\pi} d\theta \sin\theta\, Y^*_{JM}(\theta,\phi) Y_{J'M'}(\theta,\phi) = \delta_{JJ'}\delta_{MM'}, \tag{3.30}$$

$$\sum_{J=0}^{\infty}\sum_{M=-J}^{J} Y^*_{JM}(\theta,\phi) Y_{JM}(\theta',\phi')\sin\theta = \delta(\theta-\theta')\delta(\phi-\phi'). \tag{3.31}$$

The following relation will also prove useful:

$$Y^*_{JM}(\theta,\phi) = (-1)^M Y_{J,-M}(\theta,\phi). \tag{3.32}$$

Let us now consider the functions $\psi(p,J,M,\lambda;t)$, $F(p,J,M;k)$, $G(p,J,M,\lambda)$, $\gamma(p,J,M;t)$, $F_{e,o}(p,J,M)$ defined by

$$\psi(p,J,M,\lambda;t) = \int_0^{2\pi} d\phi \int_0^{\pi} d\theta \sin\theta\, Y^*_{JM}(\theta,\phi)\psi(\mathbf{p},\lambda;t), \tag{3.33}$$

and so on, where p,θ,ϕ are the usual spherical polar coordinates of **p**.

One can invert Eq. (3.33) as

$$\psi(\mathbf{p},\lambda;t) = \sum_{J=0}^{\infty}\sum_{M=-J}^{J} Y_{JM}(\theta,\phi)\psi(p,J,M,\lambda;t) \tag{3.33'}$$

and so on.

To get the relation between $f(\mathbf{x};t)$ and $\psi(p,J,M,\lambda;t)$, we have from Eqs. (2.4) and (3.33')

$$f(\mathbf{x};t) = \sum_{J=0}^{\infty}\sum_{M=-J}^{J}\int_0^{\infty} p^2\,dp\,\chi(\mathbf{x}|\,p,J,M)\psi(p,J,M,\lambda;t), \tag{3.34}$$

where

$$\chi(\mathbf{x}|\,p,J,M) = \int_0^{2\pi} d\phi \int_0^{\pi}\sin\theta\,d\theta\,\chi(\mathbf{x}|\mathbf{p})Y^*_{JM}(\theta,\phi). \tag{3.35}$$

The integral on the right of Eq. (3.33) is evaluated in many places. One obtains the result

$$\chi(\mathbf{x}|\,p,J,M) = (2/\pi)^{1/2}(i)^J j_J(pr)Y_{JM}(\theta,\phi). \tag{3.36}$$

In Eq. (3.36), r,θ,ϕ are the spherical polar coordinates of **x** (since the spherical coordinate angles of **p** do not appear, there should be no confusion about which vector the angles belong to), and $j_J(x)$ are the usual spherical Bessel functions of order J.

The functions $\chi(\mathbf{x}|\,p,J,M)$ satisfy the orthogonality and completeness relations

$$\int \chi^*(\mathbf{x}|\,p,J,M)\chi(\mathbf{x}|\,p',J',M')\,d\mathbf{x} = \frac{\delta(p-p')}{p^2}\delta_{JJ'}\delta_{MM'},$$

$$\sum_{J=0}^{\infty}\sum_{M=-J}^{J}\int_0^{\infty}\chi^*(\mathbf{x}|\,p,J,M)\chi(\mathbf{x}'|\,p,J,M)p^2\,dp = \delta(\mathbf{x}-\mathbf{x}'). \tag{3.37}$$

Equation (3.37) can be proved in several ways. Perhaps the most straightforward is to use the explicit expressions for $\chi(\mathbf{x}|\,p,J,M)$ of Eq. (3.36), the orthogonality and completeness relations (3.30) and (3.31) for the spherical harmonics, and the Fourier–Bessel transformation for spherical Bessel functions which takes the form

$$\frac{2}{\pi}\int_0^{\infty} j_J(tx')j_J(tx')t^2x^2\,dt = \delta(x-x'). \tag{3.38}$$

Thus Eq. (3.34) can be inverted to give

$$\psi(p,J,M,\lambda;t) = \int \chi^*(\mathbf{x}|\,p,J,M)f(\mathbf{x};t)\,d\mathbf{x}. \tag{3.34'}$$

Likewise,

$$4\pi\rho_{e,o}(\mathbf{x}) = \sum_{J=0}^{\infty}\sum_{M=-J}^{J}\int_0^{\infty} p^2\,dp\,\chi(\mathbf{x}|\,p,J,M)F_{e,o}(p,J,M),$$

$$F_{e,o}(p,J,M) = 4\pi\int \chi^*(\mathbf{x}|\,p,J,M)\rho_{e,o}(\mathbf{x})\,d\mathbf{x}, \tag{3.39}$$

$$4\pi\rho(\mathbf{x};t) = \sum_{J=0}^{\infty}\sum_{M=-J}^{J}\int_0^{\infty} p^2\,dp\,\chi(\mathbf{x}|\,p,J,M)\gamma(p,J,M;t),$$

$$\gamma(p,J,M;t) = 4\pi\int \chi^*(\mathbf{x}|\,p,J,M)\rho(\mathbf{x};t)\,d\mathbf{x}, \tag{3.40}$$

$$4\pi\rho(\mathbf{x};t) = \sum_{J=0}^{\infty}\sum_{M=-J}^{J}\int_0^{\infty} p^2\,dp\,\chi(\mathbf{x}|\,p,J,M)\frac{1}{\sqrt{2\pi}}\times\int_{-\infty}^{+\infty} dk\,e^{-ikt}F(p,J,M;k),$$

$$F(p,J,M;k) = 4\pi\int \chi^*(\mathbf{x}|\,p,J,M)\,d\mathbf{x}\,\frac{1}{\sqrt{2\pi}}\times\int_{-T}^{+T} e^{ikt}\,dt\,\rho(\mathbf{x};t), \tag{3.41}$$

$$F(p,J,M;k) = \frac{1}{\sqrt{2\pi}}\int_{-T}^{+T} e^{ikt}\gamma(p,J,M;t)\,dt,$$

$$\gamma(p,J,M;t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-ikt}F(p,J,M;k)\,dk. \tag{3.42}$$

The functions $\psi(p,J,M,\lambda;\pm T)$ are given in terms of $f(\mathbf{x};\pm T)$ and $(\partial/\partial t)f(\mathbf{x};\pm T)$ as follows [cf. Eq. (3.7)]:

$$\psi(p,J,M,\lambda;\pm T)=\frac{1}{2}\Big[\int \chi^*(\mathbf{x}|p,J,M)f_\pm(\mathbf{x};\pm T)\,d\mathbf{x} + i\frac{\lambda}{p}\int \chi^*(\mathbf{x}|p,J,M)\frac{\partial}{\partial t} \times f_\pm(\mathbf{x};\pm T)\,d\mathbf{x}\Big]. \tag{3.43}$$

Finally,

$$G(p,J,M,\lambda)\equiv F(p,J,M;\lambda p) = -\frac{2i\lambda p}{\sqrt{2\pi}}\,[e^{i\lambda pT}\psi(p,J,M,\lambda;T) - e^{-i\lambda pT}\psi(p,J,M,\lambda;-T)] \tag{3.44}$$

and

$$F_e(p,J,M)=\frac{1}{2g_e(p)}\sum_\lambda G(p,J,M,\lambda),$$
$$F_o(p,J,M)=\frac{1}{2g_o(p)}\sum_\lambda \lambda G(p,J,M,\lambda). \tag{3.45}$$

The requirement that $f(\mathbf{x};t)$ and $\rho(\mathbf{x};t)$ be real leads to [on using Eqs. (3.32) and (3.36)] requirements on $\psi(p,J,M,\lambda;t)$ and $F(p,J,M;k)$, namely,

$$\psi^*(p,J,M,\lambda;t)=(-1)^{J+M}\psi(p,J,-M,-\lambda;t),$$
$$F^*(p,J,M;k)=(-1)^{J+M}F(p,J,-M;-k). \tag{3.46}$$

Both of these conditions lead to [see Eq. (3.44)]

$$G^*(p,J,M,\lambda)=(-1)^{J+M}G(p,J,-M,-\lambda) \tag{3.47}$$

and hence from Eq. (3.5)

$$F_e^*(p,J,M)=(-1)^{J+M}F_e(p,J,-M),$$
$$F_o^*(p,J,M)=(-1)^{J+M}F_o(p,J,-M). \tag{3.48}$$

To summarize: knowledge of $f_\pm(\mathbf{x};t)$ is equivalent, through Eq. (3.43), to knowing $\psi(p,J,M,\lambda;\pm T)$, from which, in turn $G(p,J,M,\lambda)$ can be obtained from Eq. (3.44) and $F_{e,o}(p,J,M)$ from Eq. (3.45). Finally $\rho_{e,o}(\mathbf{x})$ are obtained through the use of Eq. (3.39). The reality conditions assure us that $\rho_{e,o}(\mathbf{x})$ will be real.

Let us consider the special case in which $\psi(p,J,M,\lambda;-T)\equiv 0$, and hence, equivalently, $f_-(\mathbf{x};t)\equiv 0$, and in which the wave function at $t=T$ is to consist of a single multipole, i.e.,

$$f_+(\mathbf{x};T)=R_{LN}(r)Y_{LN}(\theta,\phi)+R_{L,-N}(r)Y_{L,-N}(\theta,\phi),$$
$$\frac{\partial}{\partial t}f_+(\mathbf{x},T)=S_{LN}(r)Y_{LN}(\theta,\phi)+S_{L,-N}(r)Y_{L,-N}(\theta,\phi). \tag{3.49}$$

The requirement that $f_+(\mathbf{x},t)$ be real leads to

$$R^*_{LN}=(-1)^N R_{L,-N}(r),\quad S^*_{LN}(r)=(-1)^N S_{L,-N}(r). \tag{3.50}$$

In fact, it is to satisfy the reality requirement that, for a given L, one must have a term for $-N$ as well as for N. We assume Eq. (3.49) even for cases in which $N=0$, in order to unify the notation.

In the particular case $h_e(t)=\delta(t)$, $h_o(t)=\delta'(t)$, and $T\to 0$ we can find $\rho(\mathbf{x};t)$ immediately from Eqs. (3.28) and (3.29). We see that the source density consists of the same multipoles as appear in $f_+(\mathbf{x};T)$.

For the general form of $h_e(t)$ and $h_o(t)$ we have

$$4\pi\rho_e(\mathbf{x})=F_{LN}(r)Y_{LN}(\theta,\phi)+F_{L,-N}Y_{L,-N}(\theta,\phi), \tag{3.51}$$

where

$$F_{L,\pm N}(r)=\frac{1}{\pi}\int_0^\infty [\hat{S}_{L,\pm N}(p)\cos pT + p\hat{R}_{L,\pm N}(p)\sin pT]\,\frac{p^2}{g_e(p)}j_L(pr)\,dp \tag{3.51a}$$

with

$$\hat{R}_{LN}(p)=\Big(\frac{2}{\pi}\Big)^{1/2}\int_0^\infty R_{LN}(r)j_L(pr)r^2\,dr,$$
$$\hat{S}_{LN}(p)=\Big(\frac{2}{\pi}\Big)^{1/2}\int_0^\infty S_{LN}(r)j_L(pr)r^2\,dr. \tag{3.51b}$$

Similarly,

$$4\pi\rho_o(\mathbf{x})=G_{LN}(r)Y_{LN}(\theta,\phi)+G_{L,-N}Y_{L,-N}(\theta,\phi), \tag{3.52}$$

where

$$G_{L,\pm N}(r)=-\frac{i}{\pi}\int_0^\infty [\hat{S}_{L,\pm N}(p)\sin pT - p\hat{R}_{L,\pm N}(p)\cos pT]\,\frac{p^2}{g_o(p)}j_L(pr)\,dp. \tag{3.52'}$$

As a particularly simple example, let us take

$$f_+(\mathbf{x},T)=\frac{2R(r)}{\sqrt{4\pi}},\quad \frac{\partial}{\partial t}f_+(\mathbf{x},T)\equiv 0, \tag{3.53}$$

where $R(r)$ is a real function of the radius. Then

$$R_{00}(r)=R(r),\quad R_{LN}(r)\equiv 0\quad \text{for}\quad L\neq 0,$$
$$S_{LN}\equiv 0\quad \text{for all}\ L,N. \tag{3.54}$$

If now, in addition, we take $R(r)$ to be given by

$$R(r)=A\eta(a-r), \tag{3.55}$$

where A is any real constant and $\eta(x)$ is the Heaviside function $\eta(x)=1$ for $x\geqslant 1$, $\eta(x)=0$ for $x<0$, we can find $f_+(\mathbf{x},t)$ for all time t. One obtains

$$f_+(\mathbf{x},t) = \frac{a}{\sqrt{4\pi}r}[-(r+\tau)\eta(\tau-r-a)+(r-\tau)\eta(a-r-\tau) + (r+\tau)\eta(a-r+\tau)],\quad (\tau=t-T); \tag{3.56}$$

Eq. (3.56) represents the solution of the homogeneous three-dimensional wave equation when the function is initially constant in a sphere of radius a and such that the time derivative of the function is zero initially.

If we take $h_o(t)=\delta'(t)$ and $T\to 0$, then the source which gives rise to this wave would be [from Eq. (3.28)]

$$4\pi\rho(\mathbf{x};t) = -\left[2A\eta(a-r)/\sqrt{4\pi}\right]\delta'(t). \qquad (3.57)$$

We have thereby solved in a completely explicit manner an acoustic inverse source problem in three dimensions.

4. THE THREE-DIMENSIONAL ELECTROMAGNETIC INVERSE SOURCE PROBLEM

A culmination of sorts is the treatment of the direct and inverse source problems for Maxwell's equations (1.3). These equations are obviously much more complicated than that for the one-dimensional transmission line (1.1) or the three-dimensional acoustic equation (1.2). Historically, however, the present author treated the three-dimensional electromagnetic problem earlier in Ref. 5.

The vector character of Maxwell's equations would seem to require a much more complicated treatment than the scalar cases which we have thus far discussed. However, in Ref. 5 we showed that modes could be introduced which separated the transverse electromagnetic field and currents from the longitudinal components and that the transverse components themselves separated into two modes corresponding to the two possible circular polarizations of the field. These two transverse modes satisfied time-dependent equations no more complicated than that which appears in the expression of $\psi(p,\lambda;t)$ for the one-dimensional transmission line in Eq. (2.3). Thus the three-dimensional electromagnetic inverse problem can be treated in a way which is close to that for the much simpler one-dimensional equation.

In Ref. 5 we cast Maxwell's equations (1.3) into a form which resembled Dirac's equation for an electron. We used the eigenfunctions of the Dirac-like Hamiltonian to carry out the separation of the vector electromagnetic field into the longitudinal and transverse modes discussed above. In the present paper, however, we shall carry out the separation through the use of eigenfunctions of the curl operator which we introduced in Ref. 8.

We shall first review the properties of the eigenfunctions of the curl operator and then apply them to the direct and inverse source problems.

A. Eigenfunctions of the curl operator

The eigenfunctions of the curl operator were introduced in Ref. 8 to bring about a generalization of the Helmholtz decomposition theorem for vector fields. The Helmholtz theorem says that vectors can be expressed as the sum of a longitudinal and a transverse component such that the longitudinal component can be expressed as the gradient of a scalar potential and the transverse component can be obtained as the curl of a vector potential. The usual Helmholtz decomposition has proved particularly useful for fluid dynamics and electromagnetic theory. In the application to electromagnetic theory, however (also in a different way in fluid dynamics), this decomposition is seriously flawed. This flaw shows up in the need for rather complicated geometrical and algebraic arguments needed to solve the direct source problem, for example. Typically, one takes Fourier transforms of vector wave functions $\mathbf{V}(\mathbf{x})$. If $\mathbf{G}(\mathbf{p})$ and $\mathbf{H}(\mathbf{p})$ are the Fourier transforms of the longitudinal and transverse parts of $\mathbf{V}(\mathbf{x})$, they are characterized by the requirement that they satisfy the algebraic restrictions $\mathbf{p}\times\mathbf{G}(\mathbf{p}) = 0$ and $\mathbf{p}\cdot\mathbf{H}(\mathbf{p}) = 0$ which is the Fourier transform equivalent to requiring that the curl of the longitudinal part and the divergence of the transverse part $\mathbf{V}(\mathbf{x})$ vanish, respectively. The transform of Maxwell's equations lead to equations for $\mathbf{H}(\mathbf{p})$ and $\mathbf{G}(\mathbf{p})$ which are coupled in the same way as the vectors of the original untransformed equations. The sole benefit is to replace differentiation by multiplication. One then struggles with these equations in various ways to uncouple them. One way to uncouple the equations is to introduce vector and scalar potentials. However, though each component of the vector potential satisfies the usual second-order wave equation of the form (1.2), the components must satisfy a gauge condition. Moreover, some problems arise with setting the gauge.

All these complications are avoided by using the eigenfunctions of the curl operator. The curl operator $\nabla\times$ is treated as being an operator in a complex vector space with the obvious inner product. One finds the eigenfunctions of this operator and expands all vectors in terms of these eigenfunctions. The expansion is a generalization of the Fourier transform. Scalars, such as the charge density, are simply Fourier-transformed. The transformed Maxwell's equations appear as three uncoupled first-order differential equations in time, each of which is rotationally invariant and each of which is readily integrated. One of the equations is for the longitudinal component of the field in terms of the charge density, and the other two equations are for two transverse components corresponding to two circularly polarized fields. The direct and inverse source problems for the transverse components are very close to those of the one-dimensional problem. Vector and scalar potentials play no role whatever, though they can be introduced. (It might be mentioned that a further generalization of the Helmholtz theorem for tensor fields of arbitrary rank is given in Ref. 11.)

We now proceed to discuss the eigenfunctions of the curl operator. We got to them through a careful study of the properties of the rotation group. We shall simply summarize their properties. For more details we refer to Ref. 8.

There are two forms for the eigenfunctions of the curl operator. The first corresponds to the use of Cartesian coordinates and bears a close resemblance to the use of ordinary Fourier transformations in three dimensions, and the second is useful for treating problems in terms of spherical polar coordinates and uses vector spherical harmonics. The first form of the eigenfunctions will be treated now. The second will be treated in Sec. 4D of the present paper.

Let us consider an arbitrary unit vector $\boldsymbol{\eta}$ and a discrete variable $\lambda = \pm 1,0$. For each value of $\boldsymbol{\eta}$ we define a vector $\mathbf{Q}_\lambda(\boldsymbol{\eta})$ by

$$\mathbf{Q}_0(\boldsymbol{\eta}) = -\boldsymbol{\eta},$$

$$\mathbf{Q}_\lambda(\boldsymbol{\eta}) = -\frac{\lambda}{\sqrt{2}}\left[\frac{\eta_1(\eta_1+i\lambda\eta_2)}{1+\eta_3} - 1,\ \frac{\eta_2(\eta_1+i\lambda\eta_2)}{1+\eta_3} - i\lambda,\ \eta_1+i\lambda\eta_2\right]$$

$$(\lambda = \pm 1). \qquad (4.1)$$

It is readily seen that the vectors $\mathbf{Q}_\lambda(\boldsymbol{\eta})$ satisfy the following orthogonality and completeness relations:

$$\mathbf{Q}_\lambda^*(\boldsymbol{\eta})\cdot\mathbf{Q}_\mu(\boldsymbol{\eta})=\delta_{\lambda\mu}\,,\quad \sum_\lambda Q_{i\lambda}^*(\boldsymbol{\eta})Q_{j\lambda}(\boldsymbol{\eta})=\delta_{ij}\,. \tag{4.2}$$

Also

$$\boldsymbol{\eta}\cdot\mathbf{Q}_\lambda(\boldsymbol{\eta})=-\delta_{\lambda 0}\,,\quad \boldsymbol{\eta}\times\mathbf{Q}_\lambda(\boldsymbol{\eta})=-i\lambda\,\mathbf{Q}_\lambda(\boldsymbol{\eta})\,. \tag{4.3}$$

We now introduce the eigenfunctions of the curl operator $\boldsymbol{\chi}(\mathbf{x}|\mathbf{p},\lambda)$ by

$$\boldsymbol{\chi}(\mathbf{x}|\mathbf{p},\lambda)=[1/(2\pi)^{3/2}]e^{i\mathbf{p}\cdot\mathbf{x}}\mathbf{Q}_\lambda(\mathbf{p}/p)\,,\quad p=|\mathbf{p}|\,. \tag{4.4}$$

The eigenfunctions satisfy the following orthogonality and completeness relations:

$$\int\boldsymbol{\chi}^*(\mathbf{x}|\mathbf{p},\lambda)\cdot\boldsymbol{\chi}(\mathbf{x}|\mathbf{p}',\lambda')\,d\mathbf{x}=\delta(\mathbf{p}-\mathbf{p}')\delta_{\lambda,\lambda'}\,,$$

$$\sum_\lambda\int\chi_i^*(\mathbf{x}|\mathbf{p},\lambda)\chi_j(\mathbf{x}'|\mathbf{p},\lambda)\,d\mathbf{p}=\delta(\mathbf{x}-\mathbf{x}')\delta_{ij}\,. \tag{4.5}$$

Moreover, the eigenfunctions *are* eigenfunctions of the curl operator with eigenvalue λp, since

$$\nabla\times\boldsymbol{\chi}(\mathbf{x}|\mathbf{p},\lambda)=\lambda p\boldsymbol{\chi}(\mathbf{x}|\mathbf{p},\lambda)\,. \tag{4.6}$$

Another important property of the eigenfunctions is

$$\nabla\cdot\boldsymbol{\chi}(\mathbf{x}|\mathbf{p},\lambda)=-ip\frac{e^{i\mathbf{p}\cdot\mathbf{x}}}{(2\pi)^{3/2}}\delta_{\lambda,0}\,. \tag{4.7}$$

Using these properties of the eigenfunctions, we can now generalize the Helmholtz theorem as follows.

Any vector $\mathbf{v}(\mathbf{x})$ can be written in the form

$$\mathbf{v}(\mathbf{x})=\sum_\lambda\mathbf{v}_\lambda(\mathbf{x})\,,\quad\text{where}\quad \mathbf{v}_\lambda(\mathbf{x})=\int\boldsymbol{\chi}(\mathbf{x}|\mathbf{p},\lambda)V(\mathbf{p},\lambda)\,d\mathbf{p}\,, \tag{4.8}$$

where

$$V(\mathbf{p},\lambda)=\int\boldsymbol{\chi}^*(\mathbf{x}|\mathbf{p},\lambda)\cdot\mathbf{v}(\mathbf{x})\,d\mathbf{x}\,. \tag{4.9}$$

In this decomposition

$$\nabla\cdot\mathbf{v}_\lambda(\mathbf{x})=0\quad\text{for}\quad\lambda=\pm1\,,\quad\nabla\times\mathbf{v}_0(x)=0\,. \tag{4.10}$$

We have thereby sharpened the Helmholtz theorem by showing that there are *two*, rather than one, transverse components of a vector.

This decomposition is rotationally invariant in the following sense:

Let $\mathbf{v}'(\mathbf{x})$ be the vector obtained from $\mathbf{v}(\mathbf{x})$ by a rotation R of the coordinate system. That is, the components of the vector $\mathbf{v}(\mathbf{x})$ in the new coordinate system are $v_i'(\mathbf{x})$ so that

$$v_i'(\mathbf{x})=\sum_j R_{ij}v_j(R^{-1}\mathbf{x})\,, \tag{4.11}$$

where R_{ij} are the components of the rotation matrix R. Then

$$\mathbf{V}'(\mathbf{x})=\sum_\lambda\mathbf{v}_\lambda'(\mathbf{x})\,. \tag{4.12}$$

In words: the decomposition of the vector in the new frame of reference can be performed by first carrying out the decomposition in the old frame and then transforming each $\mathbf{v}_\lambda$ separately. The modes described by λ do not mix in the rotated frame. In a certain sense our decomposition is *irreducible*.

This result follows from the relation

$$\sum_j R_{ij}\chi_j(R^{-1}\mathbf{x}|\mathbf{p},\lambda)=\exp[2i\lambda\Phi(\boldsymbol{\theta},\boldsymbol{\eta})]\chi_i(\mathbf{x}|\mathbf{p},\lambda)\,, \tag{4.13}$$

where $\Phi(\boldsymbol{\theta},\boldsymbol{\eta})$ is defined as the principal branch of

$$\tan\Phi(\boldsymbol{\theta},\boldsymbol{\eta})=\frac{(\boldsymbol{\theta}\cdot\boldsymbol{\eta}+\theta_3)\tan(\theta/2)}{\theta(1+\eta_3)+(\boldsymbol{\theta}\times\boldsymbol{\eta})_3\tan(\theta/2)}\,, \tag{4.13'}$$

where $\boldsymbol{\eta}=(\mathbf{p}/p)$ and the unit vector along $\boldsymbol{\theta}$ specifies the direction of the axis of rotation and $\theta=|\boldsymbol{\theta}|$ is the angle of rotation in the axis-angle description of the rotation associated with the rotation matrix R. [The i in the exponential in Eq. (4.13) is the imaginary $\sqrt{-1}$, not the subscript i.] Thus the eigenfunctions of the curl operator are also eigenfunctions of the rotation operator.

We can now introduce vector potentials $\mathbf{A}_\lambda(\mathbf{x})$ defined by

$$\nabla\times\mathbf{A}_\lambda(x)=\mathbf{v}_\lambda(\mathbf{x})\,,\quad\lambda=\pm1\,, \tag{4.14}$$

and a scalar potential $V(x)$ defined by

$$\nabla V(\mathbf{x})=\mathbf{v}_0(\mathbf{x})\,. \tag{4.14'}$$

Equations (4.12) and (4.14) constitute our sharpening of the Helmholtz theorem.

We shall now give the general form of the vector potentials $\mathbf{A}_\lambda(\mathbf{x})$ and the scalar potential $V(\mathbf{x})$ which we regard as solutions of Eqs. (4.14) and (4.14'), respectively. To find the vector potentials, we write the vector potentials in the form of an expansion in terms of the eigenfunctions of the curl operator, as in Eq. (4.9),

$$\mathbf{A}_\lambda(\mathbf{x})=\sum_{\lambda'}\int\boldsymbol{\chi}(\mathbf{x}|\mathbf{p},\lambda')B(\mathbf{p},\lambda')\,d\mathbf{p}\,. \tag{4.15}$$

On substituting into Eq. (4.14) and using (4.6), the expansion Eq. (4.8), and the linear independence of the eigenfunctions, we obtain

$$B(\mathbf{p},\lambda)=\lambda V(\mathbf{p},\lambda)/p\,,\quad B(\mathbf{p},-\lambda)=0\,,$$
$$B(\mathbf{p},0)\ \text{arbitrary}\,. \tag{4.16}$$

Thus the vector potentials $\mathbf{A}_\lambda(\mathbf{x})$ can be written as

$$\mathbf{A}_\lambda(\mathbf{x})=\int\boldsymbol{\chi}(\mathbf{x}|\mathbf{p},\lambda)B(\mathbf{p},\lambda)\,d\mathbf{p}+\mathbf{W}(\mathbf{x})\,,$$
$$\mathbf{W}(\mathbf{x})=\int\boldsymbol{\chi}(\mathbf{x}|\mathbf{p},0)B(\mathbf{p},0)\,d\mathbf{p}\,. \tag{4.17}$$

The first of Eq. (4.17) show us that each of the two vector potentials (one for $\lambda=+1$ and one for $\lambda=-1$) consists of the sum of an essential, minimal part which depends uniquely on $\mathbf{v}_\lambda(\mathbf{x})$ and an arbitrary vector $\mathbf{W}(\mathbf{x})$ which can be written to have the form

$$\mathbf{W}(\mathbf{x})=\nabla Q(\mathbf{x})\,,\quad Q(\mathbf{x})=\frac{i}{(2\pi)^{3/2}}\int e^{i\mathbf{p}\cdot\mathbf{x}}\frac{B(\mathbf{p},0)}{p}\,d\mathbf{p}\,. \tag{4.17'}$$

Thus $\mathbf{W}(\mathbf{x})$ is the gradient of an arbitrary function $Q(\mathbf{x})$.

We recognize $\mathbf{W}(\mathbf{x})$ is the gauge. Our procedure has isolated the essential parts of the vector potential and has removed the mystery of the gauge.

We now want to find the function $V(\mathbf{x})$ from $\mathbf{v}_0(\mathbf{x})$. From Eq. (4.8) and the form of $\chi(\mathbf{x}|\mathbf{p},0)$ we have

$$V(\mathbf{x}) = -\frac{i}{(2\pi)^{3/2}}\int e^{i\mathbf{p}\cdot\mathbf{x}}\,\frac{V(\mathbf{p},0)}{p}\,d\mathbf{p}\,. \tag{4.18}$$

Two other properties of the vector decomposition into sums of eigenfunctions of the curl operator are of particular interest.

First we shall discuss the reality property. It is readily seen from the properties of the eigenfunctions that a necessary and sufficient condition for a vector $\mathbf{v}(\mathbf{x})$ to be real is that in the expansion equations (4.8) and (4.9)

$$V(\mathbf{p},\lambda) = -\frac{p_1 - i\lambda p_2}{p_1 + i\lambda p_2}\,V^*(-\mathbf{p},\lambda)\,. \tag{4.19}$$

The second property is an explicit formula by which the longitudinal portion of a vector can be removed from the vector, leaving thereby only a transverse part. Let $\mathbf{v}_T(\mathbf{x})$ be the transverse part of a vector $\mathbf{v}(\mathbf{x})$ and let $\mathbf{v}_L(\mathbf{x})$ be the longitudinal part.

Then

$$\mathbf{v}_L(\mathbf{x}) = \int \chi(\mathbf{x}|\mathbf{p},0)V(\mathbf{p},0)\,d\mathbf{p} = \int D_M(\mathbf{x}-\mathbf{x}')\mathbf{v}(\mathbf{x}')\,d\mathbf{x}'\,, \tag{4.20}$$

where $D_M(\mathbf{x})$ is a dyadic represented by a 3×3 matrix acting on the vector with elements

$$D_{Mij}(\mathbf{x}) = \int \chi_i(\mathbf{x}|\mathbf{p},0)\chi_j^*(0|\mathbf{p},0)\,d\mathbf{p}\,. \tag{4.20a}$$

By explicit calculation

$$D_{Mij}(\mathbf{x}-\mathbf{x}') = -\frac{1}{4\pi}\frac{\partial^2}{\partial x_i\,\partial x_j}\frac{1}{|\mathbf{x}-\mathbf{x}'|}\,. \tag{4.20b}$$

Thus

$$\mathbf{v}_L(\mathbf{x}) = -\frac{1}{4\pi}\nabla\left[\nabla\cdot\int\frac{\mathbf{v}(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\,d\mathbf{x}'\right]. \tag{4.21}$$

The above result indicates that if $\mathbf{v}(\mathbf{x})$ has compact support, its longitudinal component $\mathbf{v}_L(x)$ and hence its transverse component $\mathbf{v}_T(\mathbf{x}) = \mathbf{v}(\mathbf{x}) - \mathbf{v}_L(\mathbf{x})$ will only exceptionally also have compact support. This result is of the greatest importance for physical applications in both the direct and inverse problems. One would like to have the currents located in a finite portion of space, i.e., have compact support, for practical realizations. However, as we shall soon see, only the transverse component of the current contributes to radiation fields, and thus, even when the total current has compact support, the transverse portion may not.

Some of this difficulty can be overcome by working with spherical polar coordinates. Compactness can then be defined with respect to the radial variable alone and examples of compact transverse vectors can easily be given.

B. The direct problem for Maxwell's equations

It is convenient to rewrite Maxwell's equations in the Bateman form. Thus we introduce the complex vector

$$\boldsymbol{\psi}(\mathbf{x};t) = \mathbf{E}(\mathbf{x};t) - i\mathbf{H}(\mathbf{x};t)\,. \tag{4.22}$$

Maxwell's equations (1.3) become

$$\begin{aligned}\nabla\times\boldsymbol{\psi}(\mathbf{x};t) &= -i\frac{\partial}{\partial t}\boldsymbol{\psi}(\mathbf{x};t) - 4\pi i\mathbf{j}(\mathbf{x};t)\,,\\ \nabla\cdot\boldsymbol{\psi}(\mathbf{x};t) &= 4\pi\rho(\mathbf{x};t)\,.\end{aligned} \tag{4.23}$$

In contrast to the situation for the one-dimensional equation and the three-dimensional acoustic equation, the wave function $\boldsymbol{\psi}(\mathbf{x};t)$ is generally complex. However, the current $\mathbf{j}(\mathbf{x};t)$ and charge density $\rho(\mathbf{x};t)$ are real.

To solve Maxwell's equations, we expand vectors in terms of the eigenfunctions of the curl operator and Fourier expand the sole scalar of interest, the charge density.

Thus we write

$$\boldsymbol{\psi}(\mathbf{x};t) = \sum_\lambda\int \chi(\mathbf{x}|\mathbf{p},\lambda)\psi(\mathbf{p},\lambda;t)\,d\mathbf{p}\,, \tag{4.24}$$

$$4\pi\mathbf{j}(\mathbf{x};t) = \sum_\lambda\int \chi(\mathbf{x}|\mathbf{p},\lambda)\gamma(\mathbf{p},\lambda;t)\,d\mathbf{p}\,, \tag{4.25}$$

$$4\pi\rho(\mathbf{x};t) = \frac{1}{(2\pi)^{3/2}}\int e^{i\mathbf{p}\cdot\mathbf{x}}r(\mathbf{p};t)\,d\mathbf{p}\,. \tag{4.26}$$

On substituting into Eq. (4.23) and using the properties of the eigenfunctions (4.6) and (4.7), we have

$$\begin{aligned}\psi(\mathbf{p},0;t) &= (i/p)r(\mathbf{p};t)\,,\\ -\frac{\partial}{\partial t}\psi(\mathbf{p},0;t) &= \gamma(\mathbf{p},0;t)\,,\end{aligned} \tag{4.27}$$

which relate the longitudinal components of the electromagnetic field to the charge distribution and the longitudinal component of the current. The transverse part of the electromagnetic field is related to the transverse part of the current in the following way:

$$\frac{\partial}{\partial t}\psi(\mathbf{p},\lambda;t) - ip\lambda\psi(\mathbf{p},\lambda;t) = -\gamma(\mathbf{p},\lambda;t)\quad(\lambda=\pm 1)\,. \tag{4.28}$$

Our use of the eigenfunctions of the curl operator has enabled us to split the equations into a longitudinal part and two uncoupled transverse parts. The equations are of the form of inhomogeneous first order differential equations in time which are easily solved for.

Let us now complete the discussion for the longitudinal field.

In the direct problem it is clear from the first of Eq. (4.27) that the longitudinal part of the electromagnetic field is completely determined by the charge density. For

$$\begin{aligned}\boldsymbol{\psi}_L(\mathbf{x};t) &= \int \chi(\mathbf{x}|\,p,0)\frac{i}{p}r(\mathbf{p};t)\,dp\\ &= \frac{1}{(2\eta)^{3/2}}(-\nabla)\int e^{i\mathbf{p}\cdot\mathbf{x}}r(\mathbf{p};t)\frac{d\mathbf{p}}{p^2}\,.\end{aligned} \tag{4.29}$$

We now express $r(\mathbf{p};t)$ in terms of $4\pi\rho(\mathbf{x};t)$ by using the inverse Fourier transform in Eq. (4.26). One obtains from Eq. (4.29)

$$\boldsymbol{\psi}_L(\mathbf{x};t) = -\nabla\int G(\mathbf{x}-\mathbf{x}')4\pi\rho(\mathbf{x}';t)\,d\mathbf{x}'\,, \tag{4.30}$$

where

$$G(\mathbf{x}) = \frac{1}{(2\pi)^3}\int e^{i\mathbf{p}\cdot\mathbf{x}}\frac{d\mathbf{p}}{p^2} = \frac{1}{4\pi}\frac{1}{|\mathbf{x}|}. \tag{4.31}$$

Since the charge density is real and thus therefore $\boldsymbol{\psi}_L(\mathbf{x};t)$ is also real, we have from Eqs. (4.22), (4.30), and (4.31)

$$\mathbf{H}_L(\mathbf{x};t) \equiv 0, \quad \mathbf{E}_L(\mathbf{x};t) = -\nabla\int\frac{\rho(\mathbf{x}';t)}{|\mathbf{x}-\mathbf{x}'|}d\mathbf{x}', \tag{4.32}$$

which are old results in a new context.

In deriving Eq. (4.32) we have needed only the first of Eq. (4.27). The second of Eq. (4.27) may be regarded as giving the longitudinal part of the current in terms of the density. For, from Eq. (4.27) we have on eliminating $\psi(\mathbf{p},0;t)$ in an obvious manner

$$\gamma(\mathbf{p},0;t) = -\frac{i}{p}\frac{\partial}{\partial t}r(\mathbf{p};t). \tag{4.33}$$

We now evaluate $\mathbf{j}_L(\mathbf{x};t)$ in terms of $(\partial/\partial t)\rho(\mathbf{x};t)$ in much the same manner that was used to obtain Eq. (4.30). We find

$$\mathbf{j}_L(\mathbf{x};t) = \frac{1}{4\pi}\nabla\int\frac{\partial}{\partial t}\rho(\mathbf{x}';t)\frac{d\mathbf{x}'}{|\mathbf{x}-\mathbf{x}'|}. \tag{4.34}$$

Equation (4.34) is recognized as the integrated form of the equation of continuity. For

$$\nabla\cdot\mathbf{j}_L(\mathbf{x};t) = \frac{1}{4\pi}\int\frac{\partial}{\partial t}\rho(\mathbf{x}';t)\left[\nabla^2\frac{1}{|\mathbf{x}-\mathbf{x}'|}\right]d\mathbf{x}'. \tag{4.35}$$

But, since

$$\nabla^2\frac{1}{|\mathbf{x}-\mathbf{x}'|} = -4\pi\delta(\mathbf{x}-\mathbf{x}'), \tag{4.36}$$

we have, finally,

$$\nabla\cdot\mathbf{j}_L(\mathbf{x};t) + \frac{\partial}{\partial t}\rho(\mathbf{x};t) = 0. \tag{4.37}$$

The more usual form for the equation of continuity can be obtained by noting that $\nabla\cdot\mathbf{j}_T(\mathbf{x};t) = 0$ and writing $\mathbf{j}_L(\mathbf{x};t) = \mathbf{j}(\mathbf{x};t) - \mathbf{j}_T(\mathbf{x};t)$.

We have disposed completely of the direct problem for the longitudinal field. We can readily dispose of the inverse problem for the longitudinal field also, for, if the sources are off for $t < -T$ and $t > T$, the longitudinal field is also identically zero outside the time interval $-T \leqslant t \leqslant T$ and thus plays no role in the inverse problem.

We now go on to the direct problem for the transverse fields.

The differential equation (4.28) for the transverse fields are readily integrated to give

$$\psi(\mathbf{p},\lambda;t) = \psi(\mathbf{p},\lambda;-T)e^{i\lambda p(t+T)} - e^{i\lambda pt}\int_{-T}^{t}e^{-i\lambda pt'}\gamma(\mathbf{p},\lambda;t')\,dt'. \tag{4.38}$$

This result is an analog to Eq. (2.3) for the one-dimensional equation and to Eq. (3.2a) for the three-dimensional acoustic equation. The resemblance could be made closer by some changes in notation, but we forbear. It is the use of the eigenfunctions of the curl operator that leads to the simple result (4.38) and enables us to treat the direct and inverse problems of Maxwell's equations with no more real difficulty than the corresponding problems for the one-dimensional equation. From Eq. (4.38) we have

$$\begin{aligned}\boldsymbol{\psi}(\mathbf{x};t) &= \boldsymbol{\psi}_-(\mathbf{x},t) \quad \text{for } t < -T,\\ \boldsymbol{\psi}(\mathbf{x};t) &= \boldsymbol{\psi}_+(\mathbf{x},t) \quad \text{for } t > T,\end{aligned} \tag{4.39}$$

where

$$\boldsymbol{\psi}_\pm(\mathbf{x};t) = \sum_{\lambda=\pm1}\int\boldsymbol{\chi}(\mathbf{x}|\mathbf{p},\lambda)e^{i\lambda p(t\mp T)}\psi(\mathbf{p},\lambda;\pm T)\,d\mathbf{p}. \tag{4.39'}$$

It is not difficult to obtain $\boldsymbol{\psi}_\pm(\mathbf{x};t)$ in terms of $\boldsymbol{\psi}_\pm(\mathbf{x};\pm T)$. Using methods similar to those for the one-dimensional and three-dimensional acoustic cases, we introduce the matrix $G_M(\mathbf{x};t)$ by

$$\begin{aligned}G_{ii}(\mathbf{x};t) &= \left[-\frac{\partial^2}{\partial t^2} + \frac{\partial^2}{\partial x_i^2}\right]D(\mathbf{x};t),\\ G_{12}(\mathbf{x};t) &= \left[-i\frac{\partial^2}{\partial x_3\partial t} + \frac{\partial^2}{\partial x_1\partial x_2}\right]D(\mathbf{x};t) \quad \text{and cyclically},\\ G_{ij}(\mathbf{x};t) &= G^*_{ji}(\mathbf{x};t),\end{aligned} \tag{4.40}$$

where

$$D(\mathbf{x};t) = (1/4\pi r)\,\eta(r^2 - t^2), \quad r = |\mathbf{x}|. \tag{4.40'}$$

Then

$$\boldsymbol{\psi}_\pm(\mathbf{x};t) = \int G_M(\mathbf{x}-\mathbf{x}';t\mp T)\boldsymbol{\psi}_\pm(\mathbf{x}';\pm T)\,d\mathbf{x}'. \tag{4.41}$$

In Eq. (4.41) and later we shall use the notation $G_M(\mathbf{x};t)$ to be a dyadic as well as the matrix from whose elements the dyadic is constructed.

In the derivation of Eq. (4.41), which we shall not give here, it is shown that

$$\int G_M(\mathbf{x}-\mathbf{x}';t)\mathbf{v}_L(\mathbf{x}')\,d\mathbf{x}' = 0, \tag{4.42}$$

where $\mathbf{v}_L(\mathbf{x})$ is any longitudinal vector.

In Eq. (4.41) the vectors $\boldsymbol{\psi}_\pm(\mathbf{x})$ are transverse vectors, from Eq. (4.39'). However, any longitudinal vector can be added to them because of Eq. (4.42).

The solution for the transverse part of Maxwell's equations can be written

$$\boldsymbol{\psi}(\mathbf{x};t) = \int G_M(\mathbf{x}-\mathbf{x}';t+T)\boldsymbol{\psi}_-(\mathbf{x}';-T)\,d\mathbf{x}' - 4\pi\int_{-T}^{t}dt'\int G_M(\mathbf{x}-\mathbf{x}';t-t')\mathbf{j}(\mathbf{x}',t')\,d\mathbf{x}'. \tag{4.43}$$

Because of Eq. (4.42), the current in Eq. (4.43) may be replaced by its transverse part.

We can now solve the direct problem by noting that when $t \geqslant T$,

$$\boldsymbol{\psi}_+(\mathbf{x};t) = \boldsymbol{\psi}_-(\mathbf{x};t) - 4\pi\int_{-T}^{+T}dt'\int G_M(\mathbf{x}-\mathbf{x}';t-t')\mathbf{j}(\mathbf{x}';t')\,d\mathbf{x}'. \tag{4.44}$$

Hence given $\boldsymbol{\psi}_-(\mathbf{x};t)$ and the current $\mathbf{j}(\mathbf{x};t)$ or its transverse part, we can find the final electromagnetic field $\boldsymbol{\psi}_+(\mathbf{x};t)$.

Before we leave the direct problem, it seems worthwhile to discuss the significance of the discrete variable λ for

$\lambda = \pm 1$. It is readily seen that the general solution of Maxwell's equations without sources is of the form Eq. (4.39') in which $\psi(\mathbf{p},\lambda; \pm T)$ is replaced by $\psi(\mathbf{p},\lambda)$, where this latter function is an arbitrary function of $\mathbf{p},\lambda$. Moreover, it is convenient to take $T = 0$. Let us consider the case in which

$$\psi(\mathbf{p},1) = \delta(p_1)\delta(p_2)\delta(p_3 - k); \quad \text{also} \quad \psi(\mathbf{p},-1) \equiv 0. \tag{4.45}$$

Then

$$\begin{aligned} E_1(\mathbf{x};t) &= H_2(\mathbf{x};t) = \frac{1}{4(\pi)^{3/2}} \cos k(x_3 + t), \\ E_2(\mathbf{x};t) &= -H_1(\mathbf{x};t) = -\frac{1}{4(\pi)^{3/2}} \sin k(x_3 + t), \\ E_3(\mathbf{x};t) &= H_3(\mathbf{x};t) \equiv 0. \end{aligned} \tag{4.45'}$$

It is clear that this solution of Maxwell's equations is a positively polarized circularly polarized wave moving with wave number k in the negative z direction. Hence $\lambda = 1$ corresponds to positive circular polarization.

Likewise the choice of $\psi(\mathbf{p},-1) = \delta(p_1)\delta(p_2)\delta(p_3 - k)$, $\psi(\mathbf{p},+1) \equiv 0$ gives rise to a negative circularly polarized wave of wave number k moving in the positive z direction. Then $\lambda = -1$ means negative circular polarization, and, therefore, decomposition of $\boldsymbol{\psi}(\mathbf{x};t)$ into eigenvectors of curl operator is equivalent to expressing the electromagnetic field as a sum of circularly polarized radiation of various frequencies moving in various directions.

C. The inverse problem. First form

We shall now treat the inverse source problem for Maxwell's equations in a manner which parallels the coordinate-free treatment of the inverse source problem for the three-dimensional acoustic equation. Some differences are due to the fact that the time variable t appears differently in the acoustic and electromagnetic equations.

We are given $\boldsymbol{\psi}_-(\mathbf{x};t)$ and $\boldsymbol{\psi}_+(\mathbf{x};t)$, and we are required to find a current $\mathbf{j}(\mathbf{x};t)$ of the form

$$\mathbf{j}(\mathbf{x};t) = \mathbf{j}_e(\mathbf{x})h_e(t) + \mathbf{j}_o(\mathbf{x})h_o(t), \tag{4.46}$$

where $h_e(t)$ is a given real even function and $h_o(t)$ is a given real odd function of t. We shall show how *real* transverse vectors $\mathbf{j}_{e,o}(\mathbf{x})$ can be obtained uniquely from the *complex* vectors $\boldsymbol{\psi}_\pm(\mathbf{x};t)$.

Knowing $\boldsymbol{\psi}_\pm(\mathbf{x};t)$ is equivalent to knowing $\psi(\mathbf{p},\pm 1;\pm T)$ (all four of these amplitudes). Let us define $F(\mathbf{p},\lambda;k)$ for $\lambda = \pm 1$ by

$$\begin{aligned} F(\mathbf{p},\lambda;k) &= \frac{1}{\sqrt{2\pi}} \int_{-T}^{+T} e^{-ikt}\gamma(\mathbf{p},\lambda;t)\,dt \\ &= \frac{4\pi}{\sqrt{2\pi}} \int_{-T}^{+T} e^{-ikt}\,dt \int \boldsymbol{\chi}^*(\mathbf{x}|\mathbf{p},\lambda)\cdot\mathbf{j}(\mathbf{x};t)\,d\mathbf{x}. \end{aligned} \tag{4.47}$$

Then we define

$$\begin{aligned} G(\mathbf{p},\lambda) \equiv F(\mathbf{p},\lambda;\lambda p) = &-(1/\sqrt{2\pi})[e^{-i\lambda pT}\psi(\mathbf{p},\lambda;T) \\ &- e^{i\lambda pT}\psi(\mathbf{p},\lambda;-T)]. \end{aligned} \tag{4.48}$$

It is our intent to find $\mathbf{j}_{e,o}(\mathbf{x})$ of Eq. (4.46) from $G(\mathbf{p},\lambda)$.

We define, in analogy to Eq. (3.19),

$$\begin{aligned} F_e(\mathbf{p},\lambda) &= \frac{1}{2g_e(p)}\left[G(\mathbf{p},\lambda) - \frac{p_1 - i\lambda p_2}{p_1 + i\lambda p_2} G^*(-\mathbf{p},\lambda)\right], \\ F_o(\mathbf{p},\lambda) &= \frac{\lambda}{2g_o(p)}\left[G(\mathbf{p},\lambda) + \frac{p_1 - i\lambda p_2}{p_1 + i\lambda p_2} G^*(-\mathbf{p},\lambda)\right], \end{aligned} \tag{4.49}$$

where

$$g_{e,o}(k) = \frac{1}{\sqrt{2\pi}} \int_{-T}^{+T} h_{e,o}(t) e^{-ikt}\,dk. \tag{4.50}$$

Equation (4.50) differs from Eq. (2.30) in the sign of t, but, otherwise, the functions $g_{e,o}(k)$ play the same role in the present calculation as previously. As before, $g_e(k)$ is a *real*, even function of k and $g_o(k)$ is an *imaginary*, odd function of k.

Then

$$\begin{aligned} 4\pi\mathbf{j}_e(\mathbf{x}) &= \sum_\lambda \int \boldsymbol{\chi}(\mathbf{x}|\mathbf{p},\lambda)F_e(\mathbf{p},\lambda)\,d\mathbf{p}, \\ 4\pi\mathbf{j}_o(\mathbf{x}) &= \sum_\lambda \int \boldsymbol{\chi}(\mathbf{x}|\mathbf{p},\lambda)F_o(\mathbf{p},\lambda)\,d\mathbf{p}. \end{aligned} \tag{4.51}$$

The fact that

$$F_{e,o}(\mathbf{p},\lambda) = -\frac{p_1 - i\lambda p_2}{p_1 + i\lambda p_2} F^*_{e,o}(-\mathbf{p},\lambda) \tag{4.52}$$

assures us that $\mathbf{j}_{e,o}(\mathbf{x})$ are real vectors [see Eq. (4.19)].

Let us consider the special case in which $\boldsymbol{\psi}_-(\mathbf{x};t) \equiv 0$ and

$$h_e(t) = \delta(t), \quad h_o(t) = \delta'(t). \tag{4.53}$$

Moreover, we let $T \to +0$ and define

$$\boldsymbol{\psi}(\mathbf{x}) = \boldsymbol{\psi}_+(\mathbf{x};0_+) \equiv \boldsymbol{\psi}(\mathbf{x};0_+). \tag{4.54}$$

That is, we prescribe the radiation fields immediately after the current is turned off. Of course, the given vector $\boldsymbol{\psi}(\mathbf{x})$ must be a purely transverse field. We obtain, in complete analogy to Eqs. (2.45) and (3.39),

$$\begin{aligned} 4\pi\mathbf{j}_e(\mathbf{x}) &= -\operatorname{Re}\boldsymbol{\psi}(\mathbf{x}), \\ 4\pi\mathbf{j}_o(\mathbf{x}) &= \frac{1}{4\pi}\operatorname{Im}\nabla\times\int \frac{1}{|\mathbf{x}-\mathbf{x}'|}\boldsymbol{\psi}(\mathbf{x}')\,d\mathbf{x}'. \end{aligned} \tag{4.55}$$

However, since

$$\operatorname{Re}\boldsymbol{\psi}(\mathbf{x}) = \mathbf{E}(\mathbf{x};0_+), \quad \operatorname{Im}\boldsymbol{\psi}(\mathbf{x}) = -\mathbf{H}(\mathbf{x};0_+), \tag{4.55a}$$

Eq. (4.55) becomes

$$\begin{aligned} 4\pi\mathbf{j}_e(\mathbf{x}) &= -\mathbf{E}(\mathbf{x};0_+), \\ 4\pi\mathbf{j}_o(\mathbf{x}) &= -\frac{1}{4\pi}\nabla\times\int \frac{1}{|\mathbf{x}-\mathbf{x}'|}\mathbf{H}(\mathbf{x}';0_+)\,d\mathbf{x}'. \end{aligned} \tag{4.55b}$$

From Eq. (4.55b) it is clear that the current is a transverse vector as indeed are all currents as constructed by our inverse source method.

From Eq. (4.55) it is seen that a final field generally leads to currents which do not have compact support or, equivalently, are not identically zero outside a finite domain in space. Later we shall give cases of final fields for which the currents do have compact support. They will be characterized by final vectors $\mathbf{E}(\mathbf{x};0_+)$, which have compact support and $\mathbf{H}(\mathbf{x};0_+) \equiv 0$. These examples will be given in the next section when we consider the direct and inverse problems in terms of spherical polar coordinates.

We shall now give a very simple example of the application of Eq. (4.55).

Let $\psi(\mathbf{x})$ be the transverse part of $\mathbf{R}\delta(\mathbf{x})$, where the constant vector $\mathbf{R}$ is given by

$$\mathbf{R} = \mathbf{L} + i\mathbf{M}, \tag{4.56}$$

where $\mathbf{L}$ and $\mathbf{M}$ are real constant vectors. Then from Eq. (4.21) we have on subtracting the longitudinal part of $\mathbf{R}\delta(\mathbf{x})$

$$\psi(\mathbf{x}) = \mathbf{R}\delta(\mathbf{x}) + \frac{1}{4\pi}(\mathbf{R}\cdot\nabla)\nabla\frac{1}{r}. \tag{4.57}$$

We readily read off the result,

$$4\pi\mathbf{j}_e(\mathbf{x}) = \mathbf{L}\delta(\mathbf{x}) + \frac{1}{4\pi}(\mathbf{L}\cdot\nabla)\nabla\frac{1}{r},$$

$$4\pi\mathbf{j}_o(\mathbf{x}) = \frac{1}{4\pi}(\mathbf{M}\times\nabla)\frac{1}{r}. \tag{4.58}$$

It also follows from Eqs. (4.41) and (4.42) that

$$\psi_+(\mathbf{x};t) = G_M(\mathbf{x};t)\mathbf{R}. \tag{4.59}$$

This last result then says that the Riemann dyadic is the solution of Maxwell's equations with the sources given by Eq. (4.58).

D. The inverse problem in terms of spherical polar coordinates

We want to treat Maxwell's equations, both for the direct and inverse source problems, using spherical polar coordinates. Such coordinates are natural because we want to see how the radiating field behaves at large distances from the origin, and, to discuss this matter, one need consider only the radial variable. Moreover, vector multipoles of the kind used in antenna theory, which would be needed to construct the current distribution for the direct and inverse problems, appear in a very natural way.

As a first and decisive step we want to introduce complete sets of eigenfunctions of the curl operator in spherical polar coordinates, and, since these involve vector spherical harmonics, we shall discuss them briefly.

Vector spherical harmonics have been introduced in Ref. 12 for describing currents which lead to certain radiation patterns. Let J be a nonnegative integer ($J = 0,1,2,\cdots$). For each J we introduce a number M which takes on the values $-J, -J+1, \dots, J-1, J$. For each J we also introduce the nonnegative integer L which, for $J = 0$, takes on only one value $L = 1$, but for all other values of J takes on the three values $J-1, J, J+1$. For the triplet of values J, L, M vector spherical harmonics are introduced which are denoted by $\mathbf{Y}_{JLM}(\theta,\phi)$, where θ and ϕ are the usual polar angles. The properties of the vector spherical harmonics are discussed in Ref. 10 in a form to which we shall adhere. The vector spherical harmonics form a complete, orthogonal set in the θ,ϕ space.

It is natural to describe the vector spherical harmonics by giving their components in terms of the triad of unit vectors in the radial, θ, and ϕ directions. The unit vectors are denoted by $\mathbf{a}_r$, $\mathbf{a}_\theta$, $\mathbf{a}_\phi$, respectively. The components of the vector spherical harmonics are given by Eqs. (109) — (111) of Ref. 8.

We designate the eigenfunctions of the curl operator in spherical polar coordinates by $\chi(\mathbf{x}|\,p, J,M,\lambda)$. The role of the discrete variable $\lambda = 0, \pm 1$ and the continuous variable $\mathrm{P}(0 \leqslant \mathrm{p} < \infty)$ is the same as before, i.e., λp is the eigenvalue of the curl operator:

$$\nabla\times\chi(\mathbf{x}|\,p,J,M,\lambda) = \lambda p\chi(\mathbf{x}|\,p,J,M,\lambda). \tag{4.60}$$

For $\lambda = \pm 1$, J takes on the values $1,2,\cdots$ and M takes on the values $-J, -J+1,\dots, J-1, J$. For $\lambda = 0$, J takes on the values $0,1,2,\cdots$ and M takes on the same values as before.

We shall now give the functions $\chi(\mathbf{x}|\,p, J,M,\lambda)$ explicitly:

$$\begin{aligned}\chi(\mathbf{x}|\,p,J,M,\lambda) = {} & -(1/\sqrt{\pi})(i)^J\{-\lambda\,\mathbf{Y}_{JJM}(\theta,\phi)j_J(pr)\\ & + i[J/(2J+1)]^{1/2}\mathbf{Y}_{J,J+1,M}(\theta,\phi)j_{J+1}(pr)\\ & - i[(J+1)/(2J+1)]^{1/2}\\ & \times\mathbf{Y}_{J,J-1,M}(\theta,\phi)j_{J-1}(pr)\}\\ & \text{for } \lambda = \pm 1,\\ \chi(\mathbf{x}|\,p,J,M,0) = {} & [2/\pi]^{1/2}i^{J+1}\{[J+1)/(2J+1)]^{1/2}\\ & \times\mathbf{Y}_{J,J+1,M}(\theta,\phi)j_{J+1}(pr)\\ & + [J/(2J+1)]^{1/2}\mathbf{Y}_{J,J-1,M}(\theta,\phi)\\ & \times j_{J-1}(pr)\}.\end{aligned} \tag{4.61}$$

In Eq. (4.61), r,θ,ϕ are the spherical polar coordinates of $\mathbf{x}$.

On using identities for the vector spherical harmonics given in Ref. 10 and properties of the spherical Bessel functions given, for example, in Ref. 13, the eigenfunctions can also be put in the form

$$\begin{aligned}&\chi(\mathbf{x}|\,p,J,M,\lambda) = (1/\sqrt{\pi})(i^J/p)[\lambda p + \nabla\times]\mathbf{Y}_{JJM}(\theta,\phi)j_J(pr)\\ &\quad \text{for } \lambda = \pm 1,\\ &\chi(\mathbf{x}|\,p,J,M,0) = (2/\pi)^{1/2}(i^{J+1}/p)\nabla[Y_{JM}(\theta,\phi)j_J(pr)].\end{aligned} \tag{4.62}$$

The first of Eqs. (4.62) can be used to split vectors into their poloidal and toroidal parts (see, e.g., Ref. 14).

The eigenfunctions satisfy the orthogonality and completeness relationships

$$\begin{aligned}&\int \chi^*(\mathbf{x}|\,p,J,M,\lambda)\cdot\chi(\mathbf{x}|\,p',J',M',\lambda')\,d\mathbf{x}\\ &\quad = \frac{\delta(p-p')}{p^2}\delta_{JJ'}\delta_{MM'}\delta_{\lambda\lambda'},\\ &\sum_{J,M,\lambda}\int \chi_i^*(\mathbf{x}|\,p,J,M,\lambda)\chi_j(\mathbf{x}'|\,p,J,M,\lambda)p^2\,dp\\ &\quad = \delta(\mathbf{x}-\mathbf{x}')\delta_{ij}.\end{aligned} \tag{4.63}$$

The decomposition (4.8) of a vector into its transverse and longitudinal components is now written

$$\mathbf{v}(\mathbf{x}) = \sum_\lambda \mathbf{v}_\lambda(\mathbf{x}),$$

$$\mathbf{v}_\lambda(\mathbf{x}) = \sum_{J,M} \int \chi(\mathbf{x}|p,J,M,\lambda)V(p,J,M,\lambda)p^2\,dp, \tag{4.64}$$

$$V(p,J,M,\lambda) = \int \chi^*(\mathbf{x}|p,J,M,\lambda)\cdot\mathbf{v}(\mathbf{x})\,d\mathbf{x}.$$

The amplitudes $V(p,J,M,\lambda)$ have particularly simple properties under rotation of coordinates. Let $R(\boldsymbol{\theta})$ be a rotation parametrized according to the discussion of Eqs. (4.11)–(4.13'). Then if $V'(p,J,M,\lambda)$ is the amplitude corresponding to $\mathbf{v}'_\lambda(\mathbf{x})$, we have

$$V'(p,J,M,\lambda) = \sum_{M'} [\exp(i\boldsymbol{\theta}\cdot\mathbf{S}^{(J)})]_{M,M'}V(p,J,M',\lambda), \tag{4.65}$$

where $[\exp(i\boldsymbol{\theta}\cdot\mathbf{S}^{(J)})]_{M,M'}$ is a matrix element of the matrix $e^{i\boldsymbol{\theta}\cdot\mathbf{S}^{(J)}}$, in which $S_i^{(J)}$ are the spin operators in the Jth irreducible representation of the rotation group in the form given in Ref. 10, for example. The explicit expression of the matrix elements of $e^{i\boldsymbol{\theta}\cdot\mathbf{S}^{(J)}}$ is given in Ref. 15. From the viewpoint of group theory, the use of this representation represents a further uncoupling of vectors. To put the matter succinctly, the curl operator leaves the helicity and irreducible representations of the rotation group invariant. Though group theory has motivated much of the discussion, we shall not pursue the matter further in the present paper.

A relation satisfied by the eigenfunctions which will prove useful for use in obtaining real currents is

$$\chi^*(\mathbf{x}|p,J,M,\lambda) = -(-1)^{J+M}\chi(\mathbf{x}|p,J,-M,\lambda). \tag{4.66}$$

If the vector $\mathbf{v}(\mathbf{x})$ in the expansion Eq. (4.65) is to be real, then we must have

$$V^*(p,J,M,\lambda) = -(-1)^{J+M}V(p,J,-M,\lambda). \tag{4.67}$$

We now expand $\boldsymbol{\psi}(\mathbf{x};t)$ and the current $\mathbf{j}(\mathbf{x};t)$ in terms of the eigenfunctions $\chi(\mathbf{x}|p,J,M,\lambda)$,

$$\boldsymbol{\psi}(x;t) = \sum_{J,M,\lambda}\int \chi(\mathbf{x}|p,J,M,\lambda)\psi(p,J,M,\lambda;t)p^2\,dp,$$
$$4\pi j(\mathbf{x};t) = \sum_{J,M,\lambda}\int \chi(\mathbf{x}|p,J,M,\lambda)\gamma(p,J,M,\lambda;t)p^2\,dp. \tag{4.68}$$

The reality of $\mathbf{j}(\mathbf{x};t)$ leads to the condition Eq. (4.69) on $\gamma(p,J,M,\lambda;t)$.

The analog of Eq. (4.38) is

$$\psi(p,J,M,\lambda;t) = \psi(p,J,M,\lambda;-T)e^{i\lambda p(t+T)} - e^{i\lambda pt}\int_{-T}^{t} e^{-i\lambda pt'}\gamma(p,J,M,\lambda;t')\,dt',$$
$$\lambda = \pm 1. \tag{4.69}$$

We shall not consider the component for $\lambda = 0$, since we have already disposed of the longitudinal components of the fields and currents.

Let us define

$$F(p,J,M,\lambda;k) = \frac{1}{\sqrt{2\pi}}\int_{-T}^{+T}\gamma(p,J,M,\lambda;t)e^{-ikt}\,dt \tag{4.70}$$

as the analog of Eq. (4.47). Also we define [see Eqs. (4.47) and (4.49)]

$$G(p,J,M,\lambda) \equiv F(p,J,M,\lambda;\lambda p) = \frac{1}{\sqrt{2\pi}}[e^{-i\lambda pT}\psi(p,J,M,\lambda;T) - e^{i\lambda pT}\psi(p,J,M,\lambda;-T)], \tag{4.71}$$

$$F_e(p,J,M,\lambda) = \frac{1}{2g_e(p)}[G(p,J,M,\lambda) - (-1)^{J+M}G^*(p,J,-M,\lambda)],$$
$$F_o(p,J,M,\lambda) = \frac{\lambda}{2g_o(p)}[G(p,J,M,\lambda) + (-1)^{J+M}G^*(p,J,-M,\lambda)]. \tag{4.72}$$

Both $F_e(p,J,M,\lambda)$ and $F_o(p,J,M,\lambda)$ satisfy the condition Eq. (4.69).

One then sees that

$$4\pi\mathbf{j}_e(\mathbf{x}) = \sum_{J,M,\lambda}\int \chi(\mathbf{x}|p,J,M,\lambda)F_e(p,J,M,\lambda)p^2\,dp,$$
$$4\pi\mathbf{j}_o(\mathbf{x}) = \sum_{J,M,\lambda}\int \chi(\mathbf{x}|p,J,M,\lambda)F_o(p,J,M,\lambda)p^2\,dp. \tag{4.73}$$

Let us now again consider the case that $\boldsymbol{\psi}_-(\mathbf{x};t)\equiv 0$. One may consider simple modes for $\boldsymbol{\psi}_+(\mathbf{x}|t)$ by choosing

$$\psi(p,J,M,\lambda,T) = W(p,\lambda)\delta_{JK}\delta_{MN}. \tag{4.74}$$

Then

$$\boldsymbol{\psi}(\mathbf{x};t) = \sum_\lambda \int \chi(\mathbf{x}|p,K,N,\lambda)W(p,\lambda)e^{i\lambda p(t-T)}p^2\,dp \equiv \boldsymbol{\psi}_+(\mathbf{x};t),$$
$$t > T. \tag{4.75}$$

The currents will have the same transverse modes as $\boldsymbol{\psi}(\mathbf{x};t)$:

$$4\pi\mathbf{j}_e(\mathbf{x}) = -\frac{1}{\sqrt{2\pi}}\,\mathrm{Re}\sum_\lambda\int\frac{p^2dp}{g_e(p)} \times \chi(\mathbf{x}|p,K,N,\lambda)W(p,\lambda)e^{-i\lambda pT},$$
$$4\pi\mathbf{j}_o(\mathbf{x}) = \frac{1}{\sqrt{2\pi}}\,\mathrm{Im}\sum_\lambda\int\frac{p^2dp}{ig_o(p)} \times \chi(\mathbf{x}|p,K,N,\lambda)W(p,\lambda)e^{-i\lambda pT}. \tag{4.76}$$

Equations (4.75) and (4.76) assure us that the fields and currents are transverse.

We shall now consider a particularly simple case in which the current has compact support. Generalizations of this case will be obvious. Let $h_e(t) = \delta(t)$, $h_o(t) = \delta'(t)$, $T = 0_+$. Also, $K = 1, N = 0$, $W(p,\lambda) = \lambda W(p)$, where $W(p)$ is real. Then on using the explicit form of the eigenfunctions Eq. (4.61) and noting that the resulting $\boldsymbol{\psi}(\mathbf{x};0_+)$ is real

$$\mathbf{E}(\mathbf{x};0_+) = \frac{2i}{\sqrt{\pi}}\mathbf{Y}_{110}(\theta,\phi)\Gamma(r),$$
$$\Gamma(r) \equiv \int_0^\infty j_1(pr)p^2W(p)\,dp, \tag{4.77}$$
$$\mathbf{H}(\mathbf{x};0_+) \equiv 0.$$

From Eq. (4.55b)

$$4\pi \mathbf{j}_e(\mathbf{x}) = -\mathbf{E}(\mathbf{x};0_+) = -\frac{2i}{\sqrt{\pi}}\mathbf{Y}_{110}(\theta,\phi)\Gamma(r), \quad \mathbf{j}_o(\mathbf{x}) \equiv 0. \tag{4.77a}$$

Thus, if we pick $\Gamma(r) \equiv 0$ for $r > a > 0$, the current will have compact support. Using the Fourier–Bessel theorem, it is always possible to find $W(p)$ from $\Gamma(r)$ and thus $\boldsymbol{\psi}(\mathbf{x};t)$ for $t > 0$. Hence we have given a class of currents with compact support with an initial field which also has compact support for which the fields can be found at all later times.

That $\mathbf{Y}_{110}(\theta,\phi)$ is purely imaginary so that $\mathbf{E}(\mathbf{x},0_+)$ and $\mathbf{j}_e(\mathbf{x})$ are real is seen by noting that

$$\begin{aligned} &\mathbf{a}_r \cdot \mathbf{Y}_{110}(\theta,\phi) = \mathbf{a}_\theta \cdot \mathbf{Y}_{110}(\theta,\phi) = 0, \\ &\mathbf{a}_\phi \cdot \mathbf{Y}_{110}(\theta,\phi) = (3/8\pi)^{1/2} i \sin\theta. \end{aligned} \tag{4.78}$$

From (4.78) it is seen that $\mathbf{j}_e(\mathbf{x})$ may be considered to be due to charged spheres moving about the z axis, where the sphere at radius r rotates with an angular velocity proportional to $\Gamma(r)/r$, and the charge density is $\delta(t)$. Alternatively, we may consider the charge density to be unity and consider the angular velocity of the spheres at time t and radius r to be given by

$$\Omega(r,t) = (3/2)^{1/2}\pi^{-1}[\Gamma(r)/r]\delta(t). \tag{4.79}$$

We shall adopt this latter interpretation. Denoting the vector angular velocity by $\boldsymbol{\Omega}(r,t)$ with $\Omega(r,t) = |\boldsymbol{\Omega}(r,t)|$, the current which gives rise to the field

$$\mathbf{E}(\mathbf{x};0_+) = -\boldsymbol{\omega}(r), \quad \mathbf{H}(\mathbf{x};0_+) \equiv 0, \tag{4.80}$$

where $\boldsymbol{\omega}(r)$ is defined by

$$\boldsymbol{\Omega}(r;t) = \boldsymbol{\omega}(r)\delta(t) \tag{4.81}$$

[i.e., is the coefficient of $\delta(t)$ in Eq. (4.79)], is

$$4\pi \mathbf{j}(\mathbf{x};t) = \boldsymbol{\Omega}(r;t) \times \mathbf{x}. \tag{4.82}$$

[Of course, strictly speaking, the angular velocity from Eq. (4.82) is actually $\boldsymbol{\Omega}/4\pi$, but there is some convenience in using our definition (4.79).]

One could now find the fields $\mathbf{E}(\mathbf{x},t)$ and $\mathbf{H}(\mathbf{x},t)$ for $t > 0$ by using Eq. (4.41), which gives the solution in terms of an influence function. In practice this is often hard to do because the use of Eq. (4.41) requires a certain virtuosity in visualizing integrations in various domains of three-dimensional space. Actually it is sometimes easier to use Eq. (4.75) if the function $W(p,\lambda)$ is sufficiently simple. We shall now give an important case for which this is so. We shall consider the problem in which the current is due to the rotation of a sphere of radius a inside of which the charge density is unity and the time variation is essentially arbitrary.

That is, the current is given by

$$\begin{aligned} 4\pi \mathbf{j}(\mathbf{x};t) &= \boldsymbol{\omega} \times \mathbf{x} F(t) \quad \text{for } r < a \\ &\equiv 0 \quad \text{for } r > a. \end{aligned} \tag{4.83}$$

In Eq. (4.83), $\boldsymbol{\omega}$ is a constant vector and $F(t)$ is essentially an arbitrary function of time but may conveniently be thought to have finite support for the arguments below to be given in a simple way.

Thus we shall now consider the *direct* source problem for which the current is given by Eq. (4.83) and the initial field with

$$\omega(r) = |\boldsymbol{\omega}(r)| = \omega\eta(a-r). \tag{4.84}$$

Initially we shall take for the function $F(t)$ in Eq. (4.83)

$$F(t) = \delta(t), \tag{4.85}$$

and the current given by Eq. (4.83) with the initial condition that the electromagnetic field be zero. We know from our arguments that the fields at $t = 0_+$ are given by Eq. (4.80). We shall obtain $W(p)$ from Eq. (4.77) and thus $W(p,\lambda) = \lambda W(p)$. Finally, we shall obtain $\mathbf{E}(\mathbf{x},t)$ and $\mathbf{H}(\mathbf{x},t)$ for $t > 0$ from Eq. (4.75). Formally, for all time $\mathbf{E}(\mathbf{x},t)$ and $\mathbf{H}(\mathbf{x},t)$ can be written as products of an $\eta(t)$ function and the fields at time $t > 0$.

Having found the electromagnetic fields for the current given by Eqs. (4.83) and (4.85) one easily sees that the fields $\mathbf{E}_F(\mathbf{x};t)$ and $\mathbf{H}_F(\mathbf{x};t)$ defined by

$$\begin{aligned} \mathbf{E}_F(\mathbf{x};t) &= \int_{-\infty}^{+\infty} \mathbf{E}(\mathbf{x};t-t')F(t')\,dt', \\ \mathbf{E}_F(\mathbf{x};t) &= \int_{-\infty}^{+\infty} H(\mathbf{x};t-t')F(t')\,dt' \end{aligned} \tag{4.86}$$

satisfy Maxwell's equations with the current Eq. (4.83) with initial fields zero if the functions $F(t)$ are picked appropriately. [It might be mentioned that arguments of this sort can also be applied to more general situations for finding explicit solutions for the acoustic and electromagnetic equations with given sources and currents.]

We now proceed to sketch the way the fields are found for the case $F(t)$ given by Eq. (4.85).

For the problem we are considering we have

$$\Gamma(r)/r = (2/3)^{1/2}\pi\omega\eta(a-r). \tag{4.87}$$

We substitute into the second of the equations on the first line of Eq. (4.77) and use the Fourier–Bessel transform Eq. (3.38) to find $W(p)$:

$$W(p) = (8/3)^{1/2}a^4(\omega/pa)j_2(pa). \tag{4.88}$$

On substituting into Eq. (4.75) and using the first of Eq. (4.62) for the eigenfunctions of the curl operator, we find that the fields can be expressed in terms of a vector potential $\mathbf{A}(\mathbf{x};t)$:

$$\mathbf{E}(\mathbf{x};t) = \frac{\partial}{\partial t}\mathbf{A}(\mathbf{x};t), \quad \mathbf{H}(\mathbf{x};t) = \boldsymbol{\nabla} \times \mathbf{A}(\mathbf{x};t). \tag{4.89}$$

In Eq. (4.89)

$$\mathbf{A}(\mathbf{x};t) = -\mathbf{a}_\phi \omega a^2 \sin\theta \times \left[\frac{1}{\pi}\int_{-\infty}^{+\infty} \sin\left(w\frac{t}{a}\right) j_1\left(w\frac{r}{a}\right) j_2(w)\,dw\right]. \tag{4.90}$$

In obtaining Eq. (4.90) we have used Eq. (4.78). Our principal effort in the calculation is to evaluate the integral in Eq. (4.91). We employed tedious but simple methods to evaluate the integral. Our final results for the fields are

$$
\begin{aligned}
\mathbf{E}(\mathbf{x};t) = & -\boldsymbol{\omega}\times\mathbf{x}(\ \eta(a-r)\eta(a-r-t) \\
& -\{\ \eta(t-|r-a|)-\eta(t-a-r)\} \\
& \times\{(t/4r^3)[a^2-(t-r)^2]-(1/2r^2)(t-r)^2\}), \\
\mathbf{H}(\mathbf{x};t) = & -[(\boldsymbol{\omega}\cdot\mathbf{x})\mathbf{x}/r^3](2tr\eta(a-r)\eta(a-r-t) \\
& +\{\ \eta(t-|r-a|)-\eta(t-a-r)\}\{(1/8r^2) \\
& \times[a^2-(t-r)^2]^2-[(t-r)/2r][a^2-(t-r)^2]\}) \\
& +\frac{\mathbf{x}\times(\boldsymbol{\omega}\times\mathbf{x})}{r^3}(2tr\eta(a-r)\eta(a-r-t) \\
& +\{\ \eta(t-|r-a|)-\eta(t-a-r)\}\{-(1/16r^2) \\
& \times[a^2-(t-r)^2]^2+(t/4r)[a^2-(t-r)^2] \\
& -(t-r)^2/2\}).
\end{aligned} \tag{4.91}
$$

One can similarly treat the case of concentric spherical shells rotating about the same fixed axis with each shell having a different angular velocity or constant charge. In this connection it should be mentioned that the problem of determining fields due to rotating charged shells has been treated for the case of *constant* angular velocity and in a general relativistic context in Refs. 16 and 17.

[1]C. Müller, J. Ratl. Mech. Anal. **4**, 235 (1955).
[2]C. Müller, "Symposium on Electromagnetic Wave Theory," IRE Trans. Antennas Propagation **AP-4**, 224 (1956).
[3]B. J. Hoenders, *Inverse Source Problems in Optics*, edited by H. P. Baltes (Springer, New York, 1978), p. 41ff.
[4]H. E. Moses, Phys. Rev. **102**, 559 (1956), reprinted in *Inverse Scattering Papers: 1955–1963*, by I. Kay and H. E. Moses (Math. Sci. Press, Brookline, MA, 1982).
[5]H. E. Moses, Phys. Rev. **113**, 1970 (1959).
[6]N. Bleistein and J. K. Cohen, J. Math. Phys. **18**, 194 (1977).
[7]R. Courant and D. Hilbert, *Methods of Mathematical Physics, II* (Interscience, New York, 1962).
[8]H. E. Moses, SIAM J. Appl. Math. **21**, 114 (1971).
[9]H. E. Moses, Nuovo Cimento **1**, 103 (1955), reprinted in I. Kay and H. E. Moses, Ref. 4.
[10]A. R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton U. P., Princeton, NJ, 1957).
[11]H. E. Moses and A. F. Quesada, Arch. Ratl. Mech. **50**, 194 (1973).
[12]J. M. Blatt and V. Weisskopf, *Theoretical Nuclear Physics* (Wiley, New York, 1952).
[13]*Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, edited by M. Abramowitz and I. A. Stegun, Natl. Bur. Stand. Appl. Math. Ser. 55 (Natl. Bur. Stand., Washington, DC, 1964).
[14]H. E. Moses, J. Math. Phys. **17**, 1821 (1976).
[15]H. E. Moses, Ann. Phys. **37**, 224 (1966); **42**, 343 (1967).
[16]J. M. Cohen, Phys. Rev. **148**, 1264 (1966).
[17]C. Briggs, J. M. Cohen, C. L. DeWolfson, and L. S. Kegeles, Phys. Rev. D **23**, 1235 (1981).

On the solutions of the Gel'fand–Levitan equation

Reese T. Prosser
Department of Mathematics, Dartmouth College, Hanover, New Hampshire 03755

(Received 8 June 1983; accepted for publication 17 October 1983)

The solutions of the Gel'fand–Levitan equation are obtained in a closed form for general rational reflection coefficients, and as a convergent limit of closed forms for reflection coefficients sufficiently close to being rational.

PACS numbers: 03.80. + r, 03.65.Nk, 02.30.Rz

1. INTRODUCTION

In their study of one-dimensional inverse scattering problems, Gel'fand and Levitan introduced a certain linear integral equation which plays a central role.[1] In the case we consider here, the problem involves a *scattering potential* $V(x)$, defined for $-\infty < x < \infty$, which can be recovered from the resulting *reflection coefficient* $r(k)$, defined for $-\infty < k < \infty$, as follows: Set

$$R(x,y) = \hat{r}(x+y) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-ikx} r(k) e^{-iky}\, dk, \quad (1)$$

and then solve for $K(x,y)$ the *Gel'fand–Levitan equation*

$$K(x,y) + R(x,y) + \int_{-\infty}^{x} K(x,z) R(z,y)\,dz = 0. \quad (2)$$

Then $V(x)$ is given by

$$V(x) = 2\frac{dK(x,x)}{dx}. \quad (3)$$

The reflection coefficient $r(k)$ is subject to the requirements

$$r(-k) = \overline{r(k)}, \quad (4)$$

so that $R(x,y)$ is real, and

$$|r(k)| \leqslant 1, \quad (5)$$

together with some mild regularity conditions, so that the potential $V(x)$ so obtained is physically reasonable (cf. Refs. 2 and 3).

We consider here the general problem of solving the Gel'fand–Levitan equation (2) under the assumptions Eqs. (4) and (5). We know that if $r(k)$ is a *rational* function of k which is analytic in the upper half-plane, then the solutions of Eq. (2) are available in closed form. These solutions were first obtained by Kay,[4] and more recently by Pechenick and Cohen[5] in somewhat different form. (See also Ref. 6.) In Sec. 2 of this paper, we obtain the solutions of Eq. (2) in closed form in the more general case that $r(k)$ is rational, but not necessarily analytic in either half-plane. Our results include the formulas of both Kay and Pechenick–Cohen.

In Sec. 3 we consider the problem of nonrational reflection coefficients. We show here that if $r(k)$ can be approximated by a rational function $r_0(k)$ of k in a suitably chosen norm, then the solution of Eq. (2) obtained from $r(k)$ can be uniformly approximated by the solution obtained from $r_0(k)$. In the process, we obtain quantitative estimates of the effect on the solution of small errors in the measurement of $r(k)$.

2. RATIONAL REFLECTION COEFFICIENTS

Here we assume that $r(k)$ is a rational function of k, subject to the restrictions Eqs. (4) and (5). We write

$$r(k) = r_-(k) + r_+(k), \quad (6)$$

where $r_-(k)$ ($r_+(k)$) is analytic in the lower (upper) half k-plane, and assume

$$r_-(k) = \sum_{i=1}^{p} a_i/(k-b_i),$$
$$r_+(k) = \sum_{i=p+1}^{p+q} a_i/(k-b_i), \quad (7)$$

where the a_i and b_i are complex numbers chosen so that Eqs. (4) and (5) are satisfied, and so that

$$\operatorname{Im} b_i > 0 \quad \text{for } 1 \leqslant i \leqslant p,$$
$$\operatorname{Im} b_i < 0 \quad \text{for } p+1 \leqslant i \leqslant p+q. \quad (8)$$

We shall assume, moreover, that the b_i are all distinct. It follows from Eq. (1) that

$$R(x,y) = R_-(x,y) + R_+(x,y), \quad (9)$$

where

$$R_-(x,y) = \Theta(-x-y)\sum_{i=1}^{p}(ia_i)e^{-ib_i(x+y)} \quad (10)$$

$$R_+(x,y) = -\Theta(x+y)\sum_{i=p+1}^{p+q}(ia_i)e^{-ib_i(x+y)}. \quad (11)$$

Here $\Theta(z)$ is the Heaviside function:

$$\Theta(z) = \begin{cases} 0 & \text{if } z<0, \\ \frac{1}{2} & \text{if } z=0, \\ 1 & \text{if } z>0. \end{cases} \quad (12)$$

The function $R(x,y)$ satisfies a linear ordinary differential equation in y with constant coefficients, and so do $R_-(x,y)$ and $R_+(x,y)$. To see this, we rewrite $r(k)$ as

$$r(k) = p(k)/q(k), \quad (13)$$

where $p(k)$ and $q(k)$ are polynomials in k, and then use Eq. (1) to verify that

$$q(D)R(x,y) = p(D)\delta(x+y) \quad (14)$$

with $D = i\partial/\partial y$. A similar argument holds for $R_\pm(x,y)$.

As a consequence, we find that the solution $K(x,y)$ of Eq. (2) also satisfies a linear ordinary differential equation with constant coefficients. We assume $y < x$, and distinguish

 0022-2488/84/061924-06$02.50

two cases:

Case I: $x+y<0$. In this case $z+y<0$ if $z<x$, and $R(x,y)=R_-(x,y)$. Then Eq. (2) becomes

$$K(x,y)+R_-(x,y)+\int_{-\infty}^{x} K(x,z)R_-(z,y)dz=0. \quad (15)$$

Hence if we write $r_-(k)=p_-(k)/q_-(k)$, then Eq. (14) yields

$$q_-(D)K(x,y)+0+0=0. \quad (16)$$

It follows that in this case $K(x,y)$ must have the form

$$K_-(x,y)=\sum_{j=1}^{p} C_j(x)e^{-ib_jy}, \quad (17)$$

where the b_j are the zeros of $q_-(k)$, i.e., the poles or $r_-(k)$ [cf. Eq. (7)].

Case II: $x+y>0$. In this case Eq. (2) involves both R_+ and R_-, and instead of Eq. (16) we obtain

$$q(D)K(x,y)+0+p(D)K(x,-y)=0. \quad (18)$$

Hence

$$q(-D)q(D)K(x,y)=-p(D)q(-D)K(x,y)$$
$$=p(D)p(-D)K(x,y). \quad (19)$$

It follows that in this case $K(x,y)$ must have the form

$$K_+(x,y)=\sum_{j=1}^{p+q}(A_j(x)e^{ik_jy}+B_j(x)e^{-ik_jy}), \quad (20)$$

where the $\pm k_j$ are the $2(p+q)$ solutions of the equation

$$r(k)r(-k)=|r(k)|^2=1. \quad (21)$$

We shall assume here that these solutions are all distinct.

Now we put the forms Eqs. (17) and (20) back into Eq. (2), do the integration, and equate the coefficients of the resulting exponential functions of y. We assume $y<x$, and now distinguish three cases:

Case I: $x<0$. In this case Eq. (2) reduces to Eq. (15), and $K(x,y)$ to $K_-(x,y)$. We obtain, as the coefficient of e^{-ib_iy},

$$\sum_{j=1}^{p} C_j(x)+ia_ie^{-ib_ix}-a_i \times\sum_{j=1}^{p} C_j(x)e^{-i(b_j+b_i)x}/(b_j+b_i)=0. \quad (22)$$

Case II: $x>0$, *but* $x+y<0$. In this case Eq. (2) becomes

$$K_-(x,y)+R_-(x,y)+\int_{-\infty}^{-x}K_-(x,z)R_-(z,y)dz + \int_{-x}^{x}K_+(x,z)R_-(z,y)dz=0, \quad (23)$$

and we obtain, as the coefficient of e^{-ib_iy},

$$\sum_{j=1}^{p} C_j(x)+ia_ie^{-ib_ix}-a_i\sum_{j=1}^{p}C_j(x)\frac{e^{+i(b_i+b_j)x}}{(b_i+b_j)} +a_i\sum_{j=1}^{p+q}\left(A_j(x)\frac{e^{i(k_j-b_i)x}}{(k_j-b_i)}-B_j(x)\frac{e^{-i(k_j+b_i)x}}{(k_j+b_i)}\right) -a_i\sum_{j=1}^{p+q}\left(A_j(x)\frac{e^{-i(k_j-b_i)x}}{(k_j-b_i)}-B_j(x)\frac{e^{i(k_j+b_i)x}}{(k_j+b_i)}\right)=0. \quad (24)$$

Case III: $x>0$, $x+y>0$. In this case Eq. (2) becomes

$$K_+(x,y)+R_+(x,y)+\int_{-\infty}^{-x}K_-(x,z)R_-(z,y)dy + \int_{-x}^{-y}K_+(x,z)R_-(z,y)dy + \int_{-y}^{x}K_+(x,z)R_+(z,y)dy=0. \quad (25)$$

and we obtain as the coefficient of e^{-ib_iy}, $1\leqslant i\leqslant p$,

$$-a_i\sum_{j=1}^{p}C_j(x)\frac{e^{+i(b_j+b_i)x}}{(b_j+b_i)} -a_i\sum_{j=1}^{p+q}\left(A_j(x)\frac{e^{-i(k_j-b_i)x}}{(k_j-b_i)}-B_j(x)\frac{e^{i(k_j+b_i)x}}{(k_j+b_i)}\right)=0, \quad (26)$$

and as the coefficient of e^{-ib_iy}, $p+1\leqslant i\leqslant p+q$,

$$-ia_ie^{ib_ix}-a_i \times\sum_{j=1}^{p+q}\left(A_j(x)\frac{e^{i(k_j-b_i)x}}{(k_j-b_i)}-B_j(x)\frac{e^{-i(k_j+b_i)x}}{(k_j+b_i)}\right)=0. \quad (27)$$

In addition, we find in Eq. (25) a term independent of y:

$$\sum_{i=1}^{p+q}\sum_{j=1}^{p+q}(A_j(x)e^{i(k_j-b_i)x}+B_j(x)e^{-i(k_j+b_i)x}) +\sum_{i=1}^{p+q}(a_i)\sum_{j=1}^{p+q}\left(A_j(x)\frac{e^{-i(k_j-b_i)x}}{(k_j-b_i)} -B_j(x)\frac{e^{i(k_j+b_i)x}}{(k_j+b_i)}\right)=0. \quad (28)$$

It is clear that any choice of $A_j(x)$, $B_j(x)$, and $C_j(x)$ which satisfies Eqs. (22)–(28), when inserted into Eqs. (17) and (20), yields a solution of the Gel'fand–Levitan equation, Eq. (2).

When $x<0$, only Eq. (22) is involved, which may be rewritten

$$\sum_{j=1}^{p}(\delta_{ij}+B_{ij}(x))C_j(x)=-ia_ie^{-ib_ix}, \quad (29)$$

where $B_{ij}(x)$ is the matrix

$$B_{ij}(x)=-a_ie^{-i(b_j+b_i)x}/(b_j+b_i). \quad (30)$$

We observe now that this matrix must be nonsingular, because any nontrivial solution of the homogeneous equation $\Sigma B_{ij}C_j=0$ would yield, via Eq. (17), a nontrivial solution of the homogeneous form of Eq. (2), and we know that there are none.[1,2] Hence we may write

$$D_{jk}(x)=(I+B(x))^{-1}_{jk}. \quad (31)$$

Then

$$C_j(x)=\sum_k D_{jk}(x)(-ia_k)e^{-ib_kx}, \quad (32)$$

and

$$K_-(x,y)=\sum_{k,j}D_{jk}(x)(-ia_k)e^{-ib_kx}e^{-ib_jy}. \quad (33)$$

If we observe that

$$B'_{kj}(y)=+ia_ke^{-i(b_j+b_k)y}, \quad (34)$$

where $B'_{kj}(x)=dB_{kj}(x)/dx$, then we may write

$$K_-(x,y)=-\sum D_{jk}(x)B'_{kj}(x)e^{-ib_k(x-y)}. \quad (35)$$

In particular, when $x=y$,

$$K_-(x,x) = -\sum D_{jk}(x)B'_{kj}(x) = -d\,[\operatorname{tr}\log(I+B(x))]/dx. \tag{36}$$

This formula for $K_-(x,x)$ has already been given by Kay.[3]

When $x>0$, then Eq. (22) is replaced by Eqs. (24)–(28). In this case, we first note that Eq. (28) may be rewritten, using Eq. (7), as

$$\sum_{j=p+1}^{p+q} (A_j(x)e^{ik_jx} + B_j(x)e^{-ik_jx}) + \sum_{j=p+1}^{p+q} (r(k_j)A_j(x)e^{-ik_jx} + r(-k_j)B_j(x)e^{+ik_jx}) = 0. \tag{37}$$

In order to satisfy Eq. (37), we note that $|r(k_j)| = 1$ [cf. Eq. (21)] and write

$$r(k_j) = e^{2i\delta_j}. \tag{38}$$

We set, for $p+1\leqslant j\leqslant 2p+q$,

$$A_j(x) = C_j(x)e^{-i\delta_j},\quad B_j(x) = C_j(x)e^{i\delta_j}, \tag{39}$$

with the $C_j(x)$, $p+1\leqslant j\leqslant 2p+q$, arbitrary. The remaining equations, Eqs. (24), (26), and (27), may now be rewritten as

$$\sum_{j=1}^{2p+q} M_{ij}C_j(x) = D_i(x), \tag{40}$$

where the $(2p+q)\times(2p+q)$ matrix $M_{ij}(x)$ is given by

$$M_{ij}(x) = \begin{pmatrix} B_{ij}(-x) & A_{i,j-p}(-x) \\ \dfrac{1}{a_{i-p}} & A_{i-p,j-p}(x) \\ 0 & A_{i-p,j-p}(x) \end{pmatrix}. \tag{41}$$

Here $B_{ij}(x)$, for $1\leqslant i,j\leqslant p$, is given by Eq. (30), and A_{ij}, for $1\leqslant i,j\leqslant p+q$, is given by

$$A_{ij}(x) = \frac{a_i e^{i(k_j-b_i)x-i\delta_j}}{k_j-b_i} - \frac{a_i e^{-i(k_j+b_i)x+i\delta_j}}{k_j+b_i}. \tag{42}$$

The column vector $D_i(x)$ in Eq. (40) is given by

$$D_i(x) = \begin{cases} 0, & 1\leqslant i\leqslant p, \\ -ia_{i-p}e^{-ib_{i-p}x}, & p+1\leqslant i\leqslant 2p+q\,. \end{cases} \tag{43}$$

Here again we argue that the matrix $M_{ij}(x)$ must be invertible, since a nontrivial solution of Eq. (40) would lead to a nontrivial solution of Eq. (2). If we set

$$N_{jk}(x) = (M(x)^{-1})_{jk}, \tag{44}$$

then Eq. (40) leads to

$$C_j(x) = \sum_{k=1}^{2p+q} N_{jk}(x)D_k(x). \tag{45}$$

From Eq. (45) we obtain

$$K_+(x,y) = \sum_{j,k=1}^{p+q} N_{j+p,k+p}(-ia_k) \times e^{-ib_kx}(e^{ik_jy-i\delta_j} - e^{-ik_jy+i\delta_j}). \tag{46}$$

If we observe that, for $1\leqslant j,k\leqslant p+q$,

$$M'_{k+p,j+p}(y) = A'_{kj}(y) = ia_k(e^{i(k_j-b_k)y-i\delta_j} - e^{-i(k_j+b_k)y+i\delta_j}), \tag{47}$$

where $M'_{kj}(x) = dM_{kj}(x)/dx$, then we may write Eq. (46) as

$$K_+(x,y) = -\sum_{j,k=1}^{p+q} N_{j+p,k+p}(x)M'_{k+p,j+p}(y)e^{-ib_k(x-y)}. \tag{48}$$

When $r_-(k)=0$, so $r(k)=r_+(k)$ is analytic in the upper half-plane, then $p=0$, and Eq. (48) reduces to

$$K_+(x,y) = -\sum_{j,k=1}^{q} A_{jk}^{-1}(x)A'_{kj}(y)e^{-ib_k(x-y)}, \tag{49}$$

and when $x=y$,

$$K_+(x,x) = -\sum_{j,k=1}^{q} A_{jk}^{-1}(x)A'_{jk}(x) = -d\,(\operatorname{tr}\log A(x))/dx. \tag{50}$$

This formula for $K_+(x,x)$ has already been given in Ref. 6.

From Eq. (45) we also obtain, for $x>0$,

$$K_-(x,y) = \sum_{j=1}^{p}\sum_{k=1}^{p+q} N_{j,k+p}(x)(-ia_k)e^{-ib_kx-ib_jy}. \tag{51}$$

Finally, we note that the desired solution $K(x,y)$ of Eq. (2) is given by

$$K(x,y) = \begin{cases} K_-(x,y) & \text{if } x+y<0, \\ K_+(x,y) & \text{if } x+y>0. \end{cases} \tag{52}$$

Note that

$$K(x,x) = \begin{cases} K_-(x,x) & \text{if } x<0, \\ K_+(x,x) & \text{if } x>0, \end{cases} \tag{53}$$

so that Eq. (51) plays no role in recovering $V(x)$ [cf. Eq. (3)].

3. NONRATIONAL REFLECTION COEFFICIENTS

We now turn to the case where the reflection coefficient $r(k)$ is not rational in k, and seek to approximate it by another reflection coefficient $r_0(k)$ which is rational in k, in such a way that the solution $K_0(x,y)$ of Eq. (2) corresponding to $r_0(k)$ approximates the solution $K(x,y)$ corresponding to $r(k)$.

To this end, we first replace Eq. (2) by a more general equation for the kernel $K(x,y,w)$ (cf. Ref. 6):

$$K(x,y,w) + R(x,y) + \int_{-\infty}^{w} K(x,z,w)R(z,y)dz = 0. \tag{54}$$

Here w is a real parameter, and $x, y\leqslant w$. Note that if $w=x$, then $K(x,y,w) = K(x,y)$. This more general equation may be expressed in operator form, with w as a parameter, as

$$K(w) + R(w) + K(w)R(w) = 0. \tag{55}$$

Here $K(w)$ and $R(w)$ are integral operators with kernels $K(x,y,w)$ and $R(x,y,w)$, where

$$R(x,y,w) = \begin{cases} R(x,y) & \text{if } x\leqslant w \text{ and } y\leqslant w, \\ 0 & \text{otherwise.} \end{cases} \tag{56}$$

Note that because of Eq. (4), R is a symmetric operator, and hence so are $R(w)$ and $K(w)$. By taking adjoints, we get also

$$K(w) + R(w) + R(w)K(w) = 0, \tag{57}$$

and in either case we have

$$(I+K(w))(I+R(w)) = I. \tag{58}$$

Now suppose $r_i(k)$, $i=0,1$, are two admissible reflection coefficients, $R_i(x,y)$ the corresponding integral kernels obtained from Eq. (1), and $R_i(x,y,w)$ and $K_i(x,y,w)$ the corresponding integral kernels of Eq. (54). Then from Eqs. (55) and (57) we obtain

$$\begin{aligned} K_1(w) - K_0(w) + R_1(w) - R_0(w) + K_1(w)(R_1(w) - R_0(w)) \\ + (R_1(w) - R_0(w))K_0(w) + K_1(w)(R_1(w) \\ - R_0(w))K_0(w) = 0. \end{aligned} \tag{59}$$

Equation (59) gives us the difference $K_1(w) - K_0(w)$ in terms of the difference $R_1(w) - R_0(w)$.

Now if $A(x,y)$ is an arbitrary integral kernel, then we may define the following norms:

$$\|A\|_\infty = \sup_{x,y}|A(x,y)|, \tag{60}$$

$$\|A\|_H = \max\left\{\sup_x \int |A(x,y)|dy,\ \sup_y \int |A(x,y)|dx\right\}. \tag{61}$$

Then $\|A\|_\infty$ gives a uniform bound for $|A(x,y)|$, while $\|A\|_H$ gives the so-called Holmgren norm of A, which majorizes the operator norm of A (cf. Ref. 7). Note that if $B(x,y)$ is another such integral kernel, then we always have

$$\|AB\|_H \leqslant \|A\|_H \|B\|_H, \tag{62}$$

$$\begin{aligned} \|AB\|_\infty &\leqslant \|A\|_\infty \|B\|_H, \\ \|AB\|_\infty &\leqslant \|A\|_H \|B\|_\infty. \end{aligned} \tag{63}$$

Now from Eq. (55) we get, for $i=0,1$,

$$K_i(w) = -R_i(w) - K_i(w)R_i(w), \tag{64}$$

and so from Eqs. (62) and (64),

$$\|K_i(w)\|_H \leqslant \|R_i(w)\|_H + \|R_i(w)\|_H \|K_i(w)\|_H.$$

Hence, if $\|R_i(w)\|_H < 1$,

$$\|K_i(w)\|_H \leqslant \frac{\|R_i(w)\|_H}{1 - \|R_i(w)\|_H}. \tag{65}$$

Similarly, from Eqs. (63) and (64),

$$\|K_i(w)\|_\infty \leqslant \|R_i(w)\|_\infty + \|R_i(w)\|_\infty \|K_i(w)\|_H. \tag{66}$$

Hence, if $\|R_i(w)\|_H < 1$,

$$\|K_i(w)\|_\infty \leqslant \frac{\|R_i(w)\|_\infty}{1 - \|R_i(w)\|_H}.$$

Finally, from Eq. (59) we get

$$\begin{aligned} \|K_1(w) - K_0(w)\|_H \\ \leqslant \|R_1(w) - R_0(w)\|_H (1 + \|K_1(w)\|_H + \|K_0(w)\|_H \\ + \|K_1(w)\|_H \|K_0(w)\|_H), \end{aligned} \tag{67}$$

so

$$\|K_1(w) - K_0(w)\|_H \leqslant \frac{\|R_1(w) - R_0(w)\|_H}{(1 - \|R_1(w)\|_H)(1 - \|R_0(w)\|_H)}. \tag{68}$$

Similarly, we find

$$\|K_1(w) - K_0(w)\|_\infty \leqslant \frac{\|R_1(w) - R_0(w)\|_\infty}{(1 - \|R_1(w)\|_H)(1 - \|R_0(w)\|_H)}. \tag{69}$$

In particular, Eq. (69) gives a uniform bound for the difference $|K_1(x,y,w) - K_0(x,y,w)|$ in terms of a similar bound for the difference $|R_1(x,y,w) - R_0(x,y,w)|$.

To calculate the right-hand side of Eq. (69), we need only note that

$$\begin{aligned} \|R_1(w)\|_\infty &= \sup_{x,y<w}|R_1(x,y)| \leqslant \sup_{x,y}|R_i(x,y)| \\ &= \sup_z|\hat r_i(z)| = \|\hat r_i\|_\infty \leqslant \|r_i\|_1, \end{aligned} \tag{70}$$

where $\hat r(z)$ is the Fourier transform of $r(k)$ [cf. Eq. (1)], while $\|R_i(w)\|_H$

$$\begin{aligned} &= \max\left\{\sup_{x<w}\int_{-\infty}^{w}|R_i(x,y)|dy,\ \sup_{y<w}\int_{-\infty}^{w}|R_i(x,y)|dx\right\} \\ &\leqslant \int_{-\infty}^{2w}|\hat r_i(z)|dz \leqslant \|\hat r_i\|_1. \end{aligned} \tag{71}$$

Since these estimates are uniform in w, we have

Theorem 1: If $r_i(k)$, $i=0,1$, are reflection coefficients such that $\|\hat r_i\|_1 \leqslant M < 1$, and $\|r_1 - r_0\|_1 < \epsilon$, then the corresponding solutions $K_i(x,y)$ of Eq. (2) must satisfy

$$|K_1(x,y) - K_2(x,y)| \leqslant \epsilon/(1-M)^2 \tag{72}$$

uniformly in x and y.

Now suppose $V_i(x)$ are the potentials obtained from $K_i(x,x) = K_i(x,x,x)$ according to Eq. (3). It is also of interest to obtain uniform estimates for the difference $|V_1(x) - V_0(x)|$. In view of Eq. (3), this requires uniform estimates for the difference

$$|dK_1(x,x,x)/dx - dK_0(x,x,x)/dx|.$$

First we note that

$$dK(x,x,x)/dx = \partial_x K(x,y,w) + \partial_y K(x,y,w) + \partial_w K(x,y,w)|_{x=y=w}, \tag{73}$$

where ∂_x denotes $\partial/\partial x$, etc. Now from Eq. (64) we get

$$-\partial_x K = +\partial_x R + \partial_x KR \tag{74}$$

so that

$$\|\partial_x K\|_H \leqslant \|\partial_x R\|_H + \|\partial_x K\|_H \|R\|_H \tag{75}$$

and, if $\|R\|_H < 1$,

$$\|\partial_x K\|_H \leqslant \|\partial_x R\|_H/(1 - \|R\|_H). \tag{76}$$

Similarly,

$$\|\partial_y K\|_H \leqslant \|\partial_y R\|_H/(1 - \|R\|_H). \tag{77}$$

Finally, from Eq. (54),

$$\begin{aligned} -\partial_w K(x,y,w) = +0 + R(x,w)K(w,y,w) \\ + \int_{-\infty}^{w} \partial_w K(x,z,w)R(z,y)dz, \end{aligned} \tag{78}$$

so

$$\|\partial_w K\|_H \leqslant \|R \circ K\|_H + \|\partial_w K\|_H \|R\|_H, \tag{79}$$

where $(R\circ K)(x,y) = R(x,w)K(w,y,w)$. Now

$$\begin{aligned} \int |R(x,w)K(w,y,w)|dx &\leqslant \|R\|_H \|K\|_\infty \\ &\leqslant \frac{\|R\|_H \|R\|_\infty}{1 - \|R\|_H}, \end{aligned} \tag{80}$$

and

$$\int |R(x,w)K(w,y,w)|dy \leqslant \|R\|_\infty \|K\|_H \leqslant \frac{\|R\|_\infty \|R\|_H}{1-\|R\|_H}. \tag{81}$$

Hence,

$$\|\partial_w K\|_H \leqslant \|R\|_\infty \|R\|_H/(1-\|R\|_H)^2. \tag{82}$$

If $K_i(x,y,w)$, $i=0,1$, are solutions of Eq. (54) correspond ing to reflection coefficients $r_i(k)$, then from Eq. (59),

$$-\partial_x(K_1-K_0) = +\partial_x(R_1-R_0) + \partial_x K_1(R_1-R_0) + \partial_x(R_1-R_0)K_0 + \partial_x K_1(R_1-R_0)K_0. \tag{83}$$

Hence,

$$\begin{aligned}\|\partial_x(K_1-K_0)\|_\infty &\leqslant \|\partial_x(R_1-R_0)\|_\infty + \|\partial_x K_1\|_H \|R_1-R_0\|_\infty + \|\partial_x(R_1-R_0)\|_\infty \|K_0\|_H + \|\partial_x K_1\|_H \|\mathrm{R}_1-\mathrm{R}_0\|_\infty \|\mathrm{K}_0\|_H \\ &\leqslant \|\partial_x(R_1-R_0)\|_\infty (1+\|K_0\|_H) + \|\partial_x K_1\|_H \|R_1-R_0\|_\infty (1+\|K_0\|_H).\end{aligned} \tag{84}$$

Since by Eq. (65),

$$1+\|K_0\|_H \leqslant 1/(1-\|R_0\|_H), \tag{85}$$

and by Eq. (76),

$$\|\partial_x K_1\|_H \leqslant \|\partial_x R_1\|_H/(1-\|R_1\|_H), \tag{86}$$

we get

$$\begin{aligned}\|\partial_x(K_1-K_0)\|_\infty &\leqslant \frac{\|R_1-R_0\|_\infty \|\partial_x R_1\|_H}{(1-\|R_0\|_H)(1-\|R_1\|_H)} + \frac{\|\partial_x(\mathrm{R}_1-\mathrm{R}_0)\|_\infty}{1-\|\mathrm{R}_0\|_H} \\ &\leqslant \frac{\|R_1-R_0\|_\infty \|\partial_x R_1\|_H + \|\partial_x(R_1-R_0)\|_\infty}{(1-\|R_0\|_H)(1-\|R_1\|_H)}.\end{aligned} \tag{87}$$

In the same way we get

$$\|\partial_y(K_1-K_0)\|_\infty \leqslant \frac{\|R_1-R_0\|_\infty \|\partial_y R_0\|_H + \|\partial_y(R_1-R_0)\|_\infty}{(1-\|R_0\|_H)(1-\|R_1\|_H)}. \tag{88}$$

For $\partial_w(K_1-K_0)$ we use

$$\begin{aligned}-\partial_w(K_1-K_0) = {} & +K_1\circ(R_1-R_0) + \partial_w K_1(R_1-R_0) + (R_1-R_0)\circ K_0 + (R_1-R_0)\partial_w K_0 \\ & + K_1\circ(R_1-R_0)K_0 + \partial_w K_1(R_1-R_0)K_0 + K_1(R_1-R_0)\circ K_0 + K_1(R_1-R_0)\partial_w K_0.\end{aligned} \tag{89}$$

Then

$$\begin{aligned}\|\partial_w(K_1-K_0)\|_\infty &\leqslant \|R_1-R_0\|_\infty (\|K_1\|_\infty + \|\partial_w K_1\|_H + \|K_0\|_\infty + \|\partial_w K_0\|_H \\ &\quad + \|K_1\|_\infty \|K_0\|_H + \|\partial_w K_1\|_H \|K_0\|_H + \|K_1\|_H \|K_0\|_\infty + \|K_1\|_H \|\partial_w K_0\|_H) \\ &\leqslant \|R_1-R_0\|_\infty \left(\frac{\|\partial_w K_1\|_H + \|K_1\|_\infty}{1-\|R_0\|_H} + \frac{\|\partial_w K_0\|_H + \|K_0\|_\infty}{(1-\|R_1\|_H}\right) \\ &\leqslant \|R_1-R_0\|_\infty \left(\frac{\|R_1\|_\infty(1+\|R_1\|_H)}{(1-\|R_0\|_H)(1-\|R_1\|_H)^2} + \frac{\|R_0\|_\infty(1+\|R_0\|_H)}{(1-\|R_1\|_H)(1-\|R_0\|_H)^2}\right) \\ &\leqslant \|R_1-R_0\|_\infty \left(\frac{\|R_1\|_\infty(1+\|R_1\|_H) + \|R_0\|_\infty(1+\|R_0\|_H)}{(1-\|R_1\|_H)^2(1-\|R_0\|_H)^2}\right).\end{aligned} \tag{90}$$

Putting all this together, we obtain finally

$$\begin{aligned}&\|(\partial_x+\partial_y+\partial_w)(K_1-K_0)\|_\infty \\ &\leqslant \|R_1-R_0\|_\infty \left(\frac{\|\partial_x R_1\|_H + \|\partial_y\|R_0\|_H}{(1-\|R_1\|_H)(1-\|R_0\|_H)} + \frac{\|R_1\|_\infty(1+\|R_1\|_H) + \|R_0\|_\infty(1+\|R_0\|_H)}{(1-\|R_1\|_H)^2(1-\|R_0\|_H)^2}\right) \\ &\quad + \|\partial_x(R_1-R_0)\|_\infty \left(\frac{2}{(1-\|R_1\|_H)(1-\|R_0\|_H)}\right).\end{aligned} \tag{91}$$

Here we have used the fact that $\partial_x R = \partial_y R$, since

$$\partial_x R(x,y) = \partial_y R(x,y) = \hat{r}'(x+y).$$

This gives us

Theorem 2: *If the reflection coefficients* $r_i(k)$, $i=0,1$, *satisfy*: $\|\hat{r}_i\|_1 \leqslant M < 1$, $\|\hat{r}_i'\|_1 \leqslant N$, $\|r_i\|_1 \leqslant P$, *and* $\|r_1-r_0\|_1 < \epsilon$, $\|k(r_1-r_0)\| < \delta$, *then we have*

$$\|V_1(x)-V_0(x)\|_\infty \leqslant 4\epsilon\left(\frac{N+P(1+M)}{(1-M)^4}\right) + 4\delta\left(\frac{1}{(1-M)^2}\right). \tag{92}$$

Finally, we introduce the norm

$$\|r\| = \max\{\|r\|_1, \|kr\|_1, \|\hat{r}\|_1\}. \tag{93}$$

Then Theorem 2 leads to

Corollary 3: Let S be the set of reflection coefficients $r(k)$ satisfying Eqs. (4), (5), and

$$\|r\| \leqslant M < 1.$$

If $r \in S$, then r may be approximated in the norm Eq. (93) by a rational reflection coefficient r_0, so that the potential $V(x)$ corresponding to $r(k)$ is uniformly approximated by the potential $V_0(x)$ corresponding to $r_0(k)$.

4. COMMENTS

We have assumed throughout Sec. 1, where the reflection coefficient $r(k)$ is rational, that the poles of $r(k)$ are all simple [cf. Eq. (7) ff.] and that the zeroes of $|r(k)|^2 - 1$ are all distinct [cf. Eq. (21) ff.]. The case of multiple poles has been considered recently by Pechenick and Cohen[5]; they arrive at formulas for $K(x, y)$ somewhat more complicated than Eq. (50). But now it is clear that, in the sense of the norm Eq. (93), the rational coefficient $r(k)$ is a continuous function of the position of its poles, so that a slight displacement of any pole results in a slight change in $r(k)$ in the sense of Eq. (93). Similarly for the zeros of $|r(k)|^2 - 1$; a slight change in the coefficients in $r(k)$ results again in a slight change in $r(k)$, and hence in a slight change in the corresponding $V(x)$. Thus our assumptions cause no essential loss of generality. This argument has already been used by Kay.[4]

In most practical applications of this theory, the reflection coefficient is constructed from a finite set of measurements. Of the many construction procedures available, the construction of a rational reflection coefficient via Padé approximates is often a particularly attractive choice. Our results may then be used to give quantitative error bounds on the outcome.

It is unfortunate that our results in Sec. 2 are restricted by the condition $\|\hat{r}\|_1 < 1$. This seems to be a consequence of our methods, and further analysis will probably show that it can be replaced by $\|\hat{r}\|_1 < \infty$. In the meantime, we note that in fact it suffices that [cf. Eq. (71)]

$$\int_{-\infty}^{2w} |\hat{r}(k)| dk < 1. \tag{94}$$

Hence our results hold at least for all w, $-\infty < w < w_0$, for which Eq. (94) holds.

Note added in proof. The referee points out that the method used here to solve the Gel'fand–Levitan equation for rational coefficients is not new. It was first used by Bargmann in his study of the inverse problem for the half-line (unpublished), and was then extended to the tensor-force case by Fulton and Newton[8] in the earliest published version of Bargmann's procedure. It has also been used extensively by Kay and Moses.[2]

At the recent conference on inverse scattering problems held in Tulsa, Oklahoma, May, 1983, Sabatier presented a paper[9] which solves the Gel'fand–Levitan equation for rational reflection coefficients by a quite different procedure, using a sequence of Bäcklund transformations to reduce the reflection coefficient, one pole at a time, to zero. His procedure offers certain computational advantages over ours.

[1] I. M. Gel'fand and B. M. Levitan, "On The Determination of a Differential Equation by its Spectral Function," Amer. Math. Soc. Transl. Ser. 2 **1**, 253 (1955).

[2] I. Kay and H.E. Moses, *Inverse Scattering Papers: 1955–1963* (Math. Sci. Press, Brookline, MA, 1982).

[3] P. Deift and E. Trubowitz, "Inverse Scattering on the Line," Commun. Pure Appl. Math. **32**, 121 (1979).

[4] I. Kay, "The Inverse Scattering Problem When the Reflection Coefficient is a Rational Function," Commun. Pure Appl. Math. **13**, 371 (1960).

[5] K. R. Pechenick and J. Cohen, "Inverse scattering—exact solution of the Gel'fand–Levitan equation," J. Math. Phys. **22**, 1513 (1981).

[6] H. E. Moses and R. T. Prosser, "Eigenvalues and eigenfunctions associated with the Gel'fand—Levitan equation," J. Math. Phys. **25**, 108 (1984).

[7] N. Dunford and J. T. Schwartz, *Linear Operators* I (Interscience, New York, 1957), p. 518.

[8] T. Fulton and R. G. Newton, Nuovo Cimento **3**, 677 (1956).

[9] P. Sabatier, in *Proceedings of the Conference on Inverse Scattering: Theory and Application*, University of Tulsa, May, 1983 (SIAM, Philedelphia, 1983).

The optics of null strings

Jerzy F. Plebański
Centro de Estudios Avanzados y Investigación del IPN, México D.F. 14, Ap. Post. 14-740, Mexico [a]
Krzysztof Rózga
Program in Mathematical Sciences, The University of Texas at Dallas, Richardson, Texas 75080 [b]

(Received 6 August 1982; accepted for publication 9 November 1983)

The general "optical theory" of congruences of totally null strings is studied within complexified general relativity, employing the spinorial formalism, with the emphasis on the role of the Sommers vector. The results obtained yield a deeper geometric interpretation of the "complexified anatomy" of the fundamental Goldberg–Sachs theorem.

PACS numbers: 04.20.Cv

1. INTRODUCTION

The concept of a congruence of null strings is instrumental in the theory of $\mathscr{H}\mathscr{H}$ spaces ("higher heavens"), i.e., the complexified solutions to the Einstein empty-space equations, with the conformal curvature degenerated on at least one side. A theorem given in Ref. 1 states that the existence of such a congruence is a necessary and sufficient condition—modulo the empty space Einstein equations—for the conformal curvature to be degenerate at least on one side. This theorem was a starting point of an integration process of Robinson and one of us[2] (for a complete computation see Ref. 3; see also Ref. 4, and for a spinorial transcription of the results see Refs. 5 and 6), which has led to the result that a complexified solution to the Einstein equations with the conformal curvature degenerated at least on one side is entirely determined by one scalar key function, fulfilling a differential constraint of the second order, with a quadratic nonlinearity only. The program of our group aims—on the basis of this result—to derive all algebraically degenerate real spacetimes of signature $(+\ +\ +\ -)$, as real cross sections in the sense of Ref. 7, of the analytic solutions within the formalism of $\mathscr{H}\mathscr{H}$ spaces. As a side result, demonstrating the versatility of the formalism of $\mathscr{H}\mathscr{H}$ spaces in applications, it has been recently shown[8] that, employing it, one can obtain a synthetized form of all algebraically degenerated metrics of $\mathscr{H}$-spaces, derived using the N–P formalism in Ref. 9. (Compare also some preliminary results in that line in Ref. 10—important from the point of view of the program and results of our group; for a resumé of further results see Refs. 11 and 12). It should be mentioned also that already in Ref. 13, published about the same time as Refs. 1 and 10, E. Flaherty, in the final section of his book, appreciated the important role of "twistor surfaces," called null strings in our terminology.

This paper is thought of as a synthetic outline of the geometry of congruences of null strings and their intersections, postulated *a priori*, with the emphasis on the role played by the left (and right) Sommers vectors.[14] The theory is developed along similar lines as the standard theory of the optical congruences of null vector lines. A special emphasis is given to the transformational properties of the results under the conformal gauge.

This paper relies on the use of spinorial techniques, amply described in Refs. 13 and 15; for the convenience of the reader, Sec. 2 outlines the basic formulas of this formalism. A paper by I. Robinson[16] is essential in understanding the theory of null strings from the more familiar point of view of tensorial techniques.

2. THE SPINORIAL FORMALISM

We classify the Riemannian structures in four dimensions, $V=(M,g)$ [M is a differential manifold of $\dim M=4$, $g\in\Lambda^1\otimes_s\Lambda^1$ is nonsingular metric] according to the scheme:

$$\text{CR: } M \text{ complex analytic; signature meaningless}$$

$$\text{RR: } M \text{ real} \begin{cases} \text{HR: signature } (+\ +\ +\ -) \\ \text{UR: signature } (+\ +\ -\ -) \\ \text{ER: signature } (+\ +\ +\ +), \end{cases} \tag{2.1}$$

referring to these structures as to the "complex relativity" (CR) and "real relativities" (RR), subdivised into "hyperbolic relativity" (HR), "ultrahyperbolic relativity" (UR), and "elliptic (or Euclidean) relativity" (ER).

In all cases, with $A=1,2$, $\dot{B}=\dot{1},\dot{2}$, the metric is postulated in the form of

$$g=-\tfrac{1}{2}g_{A\dot{B}}\underset{s}{\otimes}g^{A\dot{B}}, \tag{2.2}$$

where the 1-forms $g^{A\dot{B}}$ constitute a base of Λ^1. Spinorial indices are to be manipulated according to $\Psi_1=\Psi^2$, $\Psi_2=-\Psi^1$, and similarly for the dotted indices. The forms $g^{A\dot{B}}$ are in CR complex analytic, in RR in general complex valued-smooth over M, and specifically, endowed with the properties that (bar denotes complex conjugation)

$$\begin{aligned} &\text{HR:} \overline{g^{A\dot{B}}}=g^{B\dot{A}}, \\ &\text{UR:} \overline{g^{A\dot{B}}}=g^{A\dot{B}}, \\ &\text{ER:} \overline{g^{A\dot{B}}}=-g_{A\dot{B}}. \end{aligned} \tag{2.3}$$

We understand then the indices of $g^{A\dot{B}}$ as the objects of the tensorial transformations from the groups

$$l^{A'}{}_{A}\times l^{\dot{B}'}{}_{\dot{B}} \tag{2.4}$$

---

[a] On leave of absence from University of Warsaw, Poland.
[b] On leave of absence from Technical University of Kielce, Poland.

 0022-2488/84/061930-11$02.50

$$\left.\begin{array}{r}\mathrm{CR{:}SL(2,C)\times S\dot{L}(2,C)}\\ \mathrm{HR{:}SL(2,C)\times \overline{SL(2,C)}}\\ \mathrm{UR{:}SL(2,}R\mathrm{)\times S\dot{L}(2,}R\mathrm{)}\\ \mathrm{ER{:}SU(2)\times S\dot{U}(2)}\end{array}\right\}\begin{array}{c}\text{dot distinguishes}\\ \text{an independent copy}\\ \text{of the same group.}\end{array}$$

These transformations leave (2.2) invariant, compatibly with (2.3); of course, the matrices $\|l^{A'}{}_A\|$ and $\|l^{\dot A'}{}_{\dot A}\|$ depend (smoothly) on a point in M.

Then, the Cartran covariant differential D-founded on the symmetric connection 1-forms,

$$\Gamma_{AB} = \Gamma_{(AB)},\ \Gamma_{\dot A\dot B} = \Gamma_{(\dot A\dot B)} \tag{2.5}$$

is defined for any Λ^p-valued tensor with respect to the corresponding group from the list (2.4) by

$$DT^{A\ldots\dot B\ldots} = dT^{\cdots} + \Gamma^A{}_{S\wedge}T^{S\ldots} + \Gamma^{\dot B}{}_{\dot S\wedge}T^{\ldots\dot S\ldots} + \cdots, \tag{2.6}$$

(d meaning the external differential). This definition leads to the (generalized) Ricci formulas

$$DDT^{A\ldots\dot B\ldots} = R^A{}_{S\wedge}T^{S\ldots} + R^{\dot B}{}_{\dot S\wedge}T^{\ldots\dot S\ldots} + \cdots, \tag{2.7}$$

where the curvature 2-forms are tensors defined by

$$\begin{aligned}R^A{}_B &= d\Gamma^A{}_B + \Gamma^A{}_{S\wedge}\Gamma^S{}_B,\\ R^{\dot A}{}_{\dot B} &= d\Gamma^{\dot A}{}_{\dot B} + \Gamma^{\dot A}{}_{\dot S\wedge}\Gamma^{\dot S}{}_{\dot B}.\end{aligned} \tag{2.8}$$

Using these notions, one then states the first and the second (Cartan) structure equations in the form of

$$\begin{aligned}&S_{\mathrm{I}}: Dg^{A\dot B} = \text{torsion} = 0,\\ &S_{\mathrm{II}}: R^A{}_B = -\frac{1}{2}C^A{}_{BCD}S^{CD} + \frac{R}{24}S^A{}_B + \frac{1}{2}C^A{}_{B\dot C\dot D}S^{\dot C\dot D},\\ &S_{\dot{\mathrm{I}}\mathrm{I}}: R^{\dot A}{}_{\dot B} = -\frac{1}{2}C^{\dot A}{}_{\dot B\dot C\dot D}S^{\dot C\dot D} + \frac{R}{24}S^{\dot A}{}_{\dot B} + \frac{1}{2}S^{CD}C_{CD}{}^{\dot A}{}_{\dot B}.\end{aligned} \tag{2.9}$$

Of course, assuming torsion equal to zero, we work with the Levi-Cività connections. The 2-forms $S^{AB} = S^{(AB)}$ and $S^{\dot A\dot B} = S^{(\dot A\dot B)}$ appearing in (2.9) are defined by

$$g^{A\dot B}\wedge g^{C\dot D} = \epsilon^{AC}S^{\dot B\dot D} + \epsilon^{\dot B\dot D}S^{AC}, \tag{2.10}$$

where $\|\epsilon^{AB}\| = \|{}^{\ 0}_{-1}\,{}^{1}_{0}\| = \|\epsilon^{\dot A\dot B}\|$ are the Levi-Cività objects in 2 dimensions.

The Λ^0 valued coefficients in (2.9) have the interpretation: (i)R = scalar curvature, (ii) $C_{ABCD} = C_{(ABCD)}$ and $C_{\dot A\dot B\dot C\dot D} = C_{(\dot A\dot B\dot C\dot D)}$ are the self-dual and anti-self-dual parts of the conformal curvature (Weyl tensor), and (iii) $C_{AB\dot C\dot D} = C_{(AB)(\dot C\dot D)}$ is the spinorial image of the Ricci tensor with its trace extracted. All curvature objects are tensors with respect to the corresponding group from the list (2.1) and, more specifically, are endowed with the properties:

$$\begin{aligned}&\text{CR:}C_{ABCD}, C_{\dot A\dot B\dot C\dot D}, C_{AB\dot C\dot D}, R\text{—are all complex, unrelated,}\\ &\text{UR: —"— —"— —"— —are all real, unrelated,}\\ &\text{HR: }\overline{C_{ABCD}} = C_{\dot A\dot B\dot C\dot D},\ \overline{C_{AB\dot C\dot D}} = C_{CD\dot A\dot B},\ \bar R = R,\\ &\text{ER: }\overline{C^{ABCD}} = C_{ABCD},\\ &\overline{C^{\dot A\dot B\dot C\dot D}} = C_{\dot A\dot B\dot C\dot D},\\ &\overline{C^{AB\dot C\dot D}} = C_{AB\dot C\dot D},\ \bar R = R.\end{aligned} \tag{2.11}$$

One introduces then the spinorial covariant gradient $\nabla_{A\dot B}$, acting on Λ^0 spinorial tensors:

$$DT^{\cdots} = -\frac{1}{2}g^{A\dot B}\nabla_{A\dot B}T^{\cdots};\ T^{\cdots}\in\Lambda^0. \tag{2.12}$$

We can now comment that the algebraic form of the right members of the S_{II} and $S_{\dot{\mathrm{I}}\mathrm{I}}$ equations, which have entered the same objects R and $C_{AB\dot C\dot D}$, assures automatically the integrability conditions of the S_{I} equations, i.e., $DDg^{A\dot B} = 0$. Then, the integrability conditions of the S_{II} and $S_{\dot{\mathrm{I}}\mathrm{I}}$ equations, the Bianchi identities ($DR_{AB} = 0 = DR_{\dot A\dot B}$), have the effective scalar form of

$$\begin{aligned}&\text{(i)}\nabla^{R\dot S}C_{AR\dot B\dot S} + \tfrac{1}{8}\nabla_{A\dot B}R = 0,\\ &\text{(ii)}\nabla^S{}_{\dot A}C_{BCDS} + \nabla_{(B}{}^{\dot S}C_{CD)\dot A\dot S} = 0,\\ &\text{(iii)}\nabla_A{}^{\dot S}C_{\dot B\dot C\dot D\dot S} + \nabla^S{}_{(\dot B}C_{|SA|\dot C\dot D)} = 0.\end{aligned} \tag{2.13}$$

It is also useful to note that the effective scalar form of the Ricci identities for 1-index spinors amounts to the following:

$$\begin{aligned}&\frac{1}{2}\nabla^S{}_{(\dot C}\nabla_{|S|\dot D)}\Psi^A = \Psi^S C^A{}_{S\dot C\dot D},\\ &\frac{1}{2}\nabla_{(C}{}^{\dot S}\nabla_{D)\dot S}\Psi^A = \Psi^S\Big(-C^A{}_{SCD} + \frac{R}{12}\epsilon_{S(C}\delta^A{}_{D)}\Big)\end{aligned} \tag{2.14}$$

and

$$\begin{aligned}&\frac{1}{2}\nabla_{(C}{}^{\dot S}\nabla_{D)\dot S}\Psi^{\dot A} = C_{CD}{}^{\dot A}{}_{\dot S}\Psi^{\dot S},\\ &\frac{1}{2}\nabla^S{}_{(\dot C}\nabla_{|S|\dot D)}\Psi^{\dot A} = \Big(-C_{\dot C\dot D}{}^{\dot A}{}_{\dot S} + \frac{R}{12}\epsilon_{\dot S(\dot C}\delta^{\dot A}{}_{\dot D)}\Big)\Psi^{\dot S},\end{aligned} \tag{2.15}$$

from which the corresponding rules for the tensorial objects of the more general type, $\Psi^{A\ldots\dot B\ldots}\in\Lambda^0$, easily follow.

Within the spinorial formalism, it is convenient to employ as the bases of Λ^p, correspondingly, the objects

$$\begin{aligned}&\Lambda^0{:}1,\ \Lambda^1{:}g^{A\dot B},\ \Lambda^2{:}\\ &\quad S^{AB}{:}= \tfrac{1}{2}\epsilon_{\dot R\dot S}g^{A\dot R}\wedge g^{B\dot S},\ S^{\dot A\dot B}{:}= \tfrac{1}{2}\epsilon_{RS}g^{R\dot A}\wedge g^{S\dot B},\\ &\Lambda^3{:}\check g^{A\dot B}{:}= \tfrac{1}{3}g^A{}_{\dot S}\wedge S^{\dot S\dot B} \equiv \tfrac{1}{3}S^A{}_S\wedge g^{S\dot B},\\ &\Lambda^4{:}v{:}= -\tfrac{1}{12}S^{AB}\wedge S_{AB} \equiv \tfrac{1}{12}S^{\dot A\dot B}\wedge S_{\dot A\dot B} \equiv \tfrac{1}{8}\check g^{A\dot B}\wedge g_{A\dot B} \neq 0.\end{aligned} \tag{2.16}$$

The objects listed above are all tensors with respect to the corresponding groups from the list (2.4). Any in general complex-valued $\omega\in\Lambda = \oplus_{p=0}^{4}\Lambda^p$ can be now written as spanned by these bases:

$$\begin{aligned}\omega &= \omega_0\cdot 1 - \tfrac{1}{2}\omega_{A\dot B}g^{A\dot B} + \tfrac{1}{4}\omega_{AB}S^{AB}\\ &\quad + \tfrac{1}{4}\omega_{\dot A\dot B}S^{\dot A\dot B} - \tfrac{1}{2}\check\omega_{A\dot B}\check g^{A\dot B} + \omega_4\cdot v,\\ &\quad \omega_0, \omega_{A\dot B}, \omega_{AB}, \omega_{\dot A\dot B}, \check\omega_{A\dot B}, \omega_4\in\Lambda^0.\end{aligned} \tag{2.17}$$

It is then convenient to introduce a somewhat unconventional definition of duality—the Hodge star—defined as the mapping $\Lambda^p\to\Lambda^{4-p}$ via the action of $*$ on the bases:

$$\begin{aligned}&*1 = v,\ *g^{A\dot B} = \check g^{A\dot B},\ *S^{AB} = S^{AB},\ *S^{\dot A\dot B} = -S^{\dot A\dot B},\\ &*\check g^{A\dot B} = g^{A\dot B},\ *v = 1,\end{aligned} \tag{2.18}$$

with $*\omega$ given thus by

$$*\omega = \omega_0 v - \tfrac{1}{2}\omega_{A\dot{B}}\check{g}^{A\dot{B}} + \tfrac{1}{4}\omega_{AB}S^{AB} - \tfrac{1}{4}\omega_{\dot{A}\dot{B}}S^{\dot{A}\dot{B}} - \tfrac{1}{2}\check{\omega}_{A\dot{B}}g^{A\dot{B}} + \omega_4\cdot 1, \tag{2.19}$$

hence$** = 1$ on Λ.

With the star so defined, we can understand as the inner product (sometimes called step product)$\lrcorner:\Lambda \times \Lambda \to \Lambda$, the operation

$$\alpha, \beta \in \Lambda:\ \alpha \lrcorner \beta := *(\alpha \wedge *\beta). \tag{2.20}$$

One then easily shows that this $\lrcorner$, in the cotangent language (within a metrical structure), has the standard properties. If, in a local coordinate patch $\{x^\alpha\}$, the metric is $g = g_{\mu\nu}\, dx^\mu \otimes_s dx^\nu$, then $dx^\mu \lrcorner dx^\nu = g^{\mu\nu}$, $\alpha_\mu\, dx^\mu \lrcorner \beta_\nu\, dx^\nu = \alpha^\mu \beta_\mu$.

Moreover, within our conventions the following also holds:

$$\text{if } \alpha,\beta,\gamma \in \Lambda^1 \text{ then } \alpha \lrcorner(\beta \wedge \gamma) = (\alpha \lrcorner \beta)\gamma - (\alpha \lrcorner \gamma)\beta. \tag{2.21}$$

We close this section by acknowledging the useful identities

$$\begin{aligned} &\tfrac{1}{4}S^{AB} \wedge S_{CD} = -\delta^A{}_{(C}\delta^B{}_{D)}\cdot v,\\ &\tfrac{1}{4}S^{\dot{A}\dot{B}} \wedge S_{\dot{C}\dot{D}} = \delta^{\dot{A}}{}_{(\dot{C}}\delta^{\dot{B}}{}_{\dot{D})}\cdot v,\\ &S_{AB} \wedge S_{\dot{C}\dot{D}} = 0, \end{aligned} \tag{2.22}$$

and

$$\begin{aligned} g^{A\dot{B}} \wedge S^{\dot{C}\dot{D}} &= -\check{g}^{A\dot{C}}\epsilon^{\dot{B}\dot{D}} - \check{g}^{A\dot{D}}\epsilon^{\dot{B}\dot{C}},\\ S^{AB} \wedge g^{C\dot{D}} &= -\epsilon^{AC}\check{g}^{B\dot{D}} - \epsilon^{BC}\check{g}^{A\dot{D}}. \end{aligned} \tag{2.23}$$

The collection of formulas given in this section is reasonably complete; knowing them plus some trivial properties of the operations on Λ like $dd = 0$, the Liebnitz properties of d and D, and the like, one has, in principle, all necessary tools needed in order to execute any relativistic computation in language explicitly covariant with respect to the groups listed in (2.4). In the effective applications, within the work of our group, and the work initiated and continued by R. P. Kerr and co-workers in the last decade, it turned out that the formalism outlined above is somewhat more advantageous than the N–P formalism.[17] The last has the obvious merit of working with the Λ^0 objects only; our treatment, on the other hand, maintains all advantages of spinorial tricks, and at the same time conserves a sort of "phonetic" notation; by working directly in the language of external forms, a more straightforward access to the results of modern differential geometry, which employs this language, is facilitated. It is to be stressed, however, that our formalism, summarized concisely by the 23 formulas of this section, is isomorphic to the NP treatment; it is rather a matter of taste and pragmatic choice for any practicioner of relativity which version of the spinorial formalism gets more effectively at the practical results. (For a comparison of various conventions used in practice see Ref. 18).

3. CONGRUENCES OF NULL STRINGS—BASIC CONCEPTS

We say that a Riemannian structure $V = (M,g)$ admits a left congruence of null strings, in a connected (and singly connected) domain $\mathscr{D} \subset M$, if there exist over $\mathscr{D}$ smooth (in CR analytic) 2- and 1-forms, $\Sigma \neq 0$ and θ, such that

$$\text{(i)}\Sigma \wedge \Sigma = 0, \quad \text{(ii)}*\Sigma = \Sigma, \quad \text{(iii)}d\Sigma = \theta \wedge \Sigma. \tag{3.1}$$

The algebraic conditions (i) and (ii) assure that the vector space $\mathfrak{X} \subset \Lambda^1$ defined by

$$\mathfrak{X} \ni x \leftrightarrow x \lrcorner \Sigma = 0, \tag{3.2}$$

has $\dim(\mathfrak{X}) = 2$ and is totally null in the sense of

$$x, y \in \mathfrak{X} \to x \lrcorner y = 0; \tag{3.3}$$

in particular $x \lrcorner x = 0$.

Proof: Indeed, according to (3.1) (i), Σ is simple, and thus there are $\sigma^A \in \Lambda^1$ such that

$$\Sigma = \sigma^1 \wedge \sigma^2 \neq 0. \tag{3.4}$$

Then, a straightforward calculation employing (2.21) implies that

$$\sigma^A \lrcorner \sigma^B = 0. \tag{3.5}$$

On the other hand, from (3.2) one infers that $\{\sigma^1,\sigma^2\}$ span $\mathfrak{X}$. Thus, dim $\mathfrak{X} = 2$ and the condition (3.3) holds. ■

Then, taking into account (3.4), it follows that (3.1) (iii) has the form of the thesis of the Frobenius theorem: $d(\sigma^1 \wedge \sigma^2) = \theta \wedge (\sigma^1 \wedge \sigma^2)$. Hence there exist, over $D,(q^A, l^A{}_B) \in \Lambda^0$ such that

$$\sigma^A = l^A{}_B\, dq^B,\ \Delta := \det(l^A{}_B) \neq 0, \tag{3.6}$$

and consequently

$$\Sigma = \Delta\, dq^1 \wedge dq^2 \neq 0. \tag{3.7}$$

This, used in (3.1) (iii), implies that

$$\theta = d \ln\Delta + \theta_A\, dq^A;\ \theta_A \in \Lambda^0. \tag{3.8}$$

Obviously $dq^A \in \mathfrak{X}$, and therefore

$$dq^A \lrcorner dq^B = 0. \tag{3.9}$$

The equations $q^A = \text{const}$ (over $\mathscr{D}$), define then some two-surfaces (a two-dimensional foliation), with the tangent spaces to these surfaces totally null, in the sense of orthogonality of any 2-vectors contained in the tangent space. Notice that the conditions (3.1) determine the q^A's modulo the transformations

$$q^{A'} = q^{A'}(q^B),\ \frac{\partial(q^{A'})}{\partial(q^B)} \neq 0. \tag{3.10}$$

The 2-surfaces from the congruence Σ are automatically geodesic.

Proof: Indeed, in a local chart $\{x^\alpha\}$, (3.9) amounts to

$$q^{A;\mu}q^B{}_{;\mu} = 0. \tag{3.11}$$

Define then

$$\mathfrak{X}^{ABC} := q^{A;\mu}q^{B;\nu}q^C{}_{;\mu\nu}. \tag{3.12}$$

From (3.11) it follows that $:\mathfrak{X}^{ABC} = \mathfrak{X}^{A[BC]}$; on the other hand, the definition (3,12) implies $\mathfrak{X}^{ABC} = \mathfrak{X}^{(AB)C}$. From these two symmetries one can easily prove that $\mathfrak{X}^{ABC} = 0$. This, however, means that, for B and C fixed, the covariant derivative of the vector field $q^{C;\nu}$ along $q^{B;\mu}$ is tangent to $q^A = \text{const}$ ($A = 1,2$) surfaces. These surfaces are geodesic since $q^{A;\mu}$ for $A = 1,2$ span their tangent spaces. ■

The property (3.9), accompanied by $dq^1 \wedge dq^2 \neq 0$, was used in Ref. 1 as the starting point in defining a congruence of "totally null strings," which have, along leaves of the foliation, an induced "metric" of rank zero. This makes them a different concept from the "null strings" studied by A. Schild (within HR),[19] which are also geodesic surfaces, and null, in the sense $\Sigma \rfloor \Sigma = 0$, but which have, along the leaves of the congruence, an induced "metric" of rank one.

In comprehending the structure of our definition (3.1) of a left congruence of (totally) null strings, it is useful to recall that any 2-form $\Sigma \neq 0$ endowed, within a metrical structure, with the properties

$$\text{(i)} \Sigma \wedge \Sigma = 0, \quad \text{(ii)} \Sigma \rfloor d * \Sigma = 0, \tag{3.13}$$

determines a congruence of 2-surfaces.

Proof: A simple way of seeing this, consists of representing Σ in a local chart $\{x^\alpha\}$ in the form of

$$\Sigma = \tfrac{1}{2}\Sigma_{\mu\nu} dx^\mu \wedge dx^\nu; \tag{3.14}$$

then, from (3.13) (i), the object $\Sigma^{\mu\nu}$, well defined within a metrical structure, has the form of a simple bivector

$$\Sigma^{\mu\nu} = a^{[\mu} b^{\nu]}. \tag{3.15}$$

The condition that the vector fields a^μ and b^μ are surface forming, i.e., their commutator is spanned by themselves, is equivalent to

$$\Sigma^{[\alpha\beta}{}_{,\delta} \Sigma^{\gamma]\delta} = 0, \tag{3.16}$$

where, if convenient, "$,\delta$" can be replaced by "$;\delta$". The invariant form of (3.16) amounts precisely to (3.13) (ii). ■

With (3.1) determining a congruence of 2-surfaces, Σ must, of course, fulfill condition (3.13) (ii). Indeed, we have that

$$\begin{aligned}\Sigma \rfloor d * \Sigma = \Sigma \rfloor d\Sigma &= \Sigma \rfloor (\theta \wedge \Sigma) = *(\Sigma \wedge *(\theta \wedge \Sigma)) \\ &= *(*(\theta \wedge *\Sigma) \wedge *\Sigma) = (\theta \rfloor \Sigma) \rfloor \Sigma \\ &= [\theta \rfloor (\sigma^1 \wedge \sigma^2)] \rfloor (\sigma^1 \wedge \sigma^2) = 0 \end{aligned} \tag{3.17}$$

(because of 2.21 and 3.5).

Having now completed an outline of the basic implications of our definition (3.1), we should like to observe that a transformation

$$\Sigma = \kappa \Sigma', \quad \theta = \theta' + d \ln \kappa; \quad 0 \neq \kappa \in \Lambda^0, \tag{3.18}$$

maintains all the conditions (3.1) invariant. It follows, that we can always, without any lost generality, remembering (3.8), selecting $\kappa = \Delta$, choose to work with the congruence of strings given by $\Sigma \in \Lambda^2$ so normalized that (after dropping primes),

$$\text{(i)} \Sigma \wedge \Sigma = 0, \quad \text{(ii)} *\Sigma = \Sigma, \quad \text{(iii)} d\Sigma = 0. \tag{3.19}$$

We will consider (3.19) as the definition of a congruence of left (totally) null strings via a 2-form given in the canonical normalization. Notice that the conditions (3.19) are still invariant with respect to a transformation (3.18), provided it is constrained by

$$d\kappa \rfloor \Sigma = 0. \tag{3.20}$$

With the congruence given via Σ in canonical normalization, one can now associate the deviation 1-form. Suppose that in a local chart $\{x^\mu\}$ (over $\mathscr{D}$), the congruence is given in the form of

$$x^\mu = x^\mu(\tau^A, q^{\dot B}); \quad \frac{\partial(x^1,x^2,x^3,x^4)}{\partial(\tau^1,\tau^2,q^{\dot 1},q^{\dot 2})} \neq 0, \tag{3.21}$$

with τ^1, τ^2 parametrizing the leaves $q^{\dot A} = \text{const}$.

The geodesic nature of our congruence implies the existence of $C^R_{AB} = C^R_{(AB)}$ such that

$$\frac{D^2 x^\mu}{\partial\tau^A \partial\tau^B} = C^R_{AB} \frac{\partial x^\mu}{\partial\tau^R} \rightarrow \frac{D}{\partial\tau^A} \Sigma^{\mu\nu} = C^S_{AS} \Sigma^{\mu\nu}, \tag{3.22}$$

where $\Sigma^{\mu\nu}$ represents contravariant components of Σ in canonical normalization. The last relation amounts to $\Sigma^{\mu\nu}{}_{;\rho}\, \partial x^\rho / \partial\tau^A = C_A{}^S{}_S \Sigma^{\mu\nu}$, which is then readily seen to be equivalent to

$$\Sigma^{\alpha\beta}{}_{;\delta} \Sigma^{\gamma\delta} = \Sigma^{\alpha\beta} \Theta^\gamma, \tag{3.23}$$

where $\Theta^\gamma = \tfrac{1}{2} C_A{}^S{}_S \epsilon^{AB} \partial x^\gamma / \partial\tau^B$ can now be interpreted as a vector field. Consistency with (3.16) necessitates then

$$\Sigma^{[\alpha\beta} \Theta^{\gamma]} = 0. \tag{3.24}$$

The cotangent object

$$\Theta := \Theta_\alpha\, dx^\alpha = -\tfrac{1}{2} \Theta_{A\dot B} g^{A\dot B} \tag{3.25}$$

is now the deviation 1-form of the congruence; in terms of it, (3.24) is equivalent to

$$\Theta \rfloor \Sigma = 0. \tag{3.26}$$

With the congruence defined by (3.19), under the still-permitted rescaling $\Sigma = \kappa \Sigma'$, κ constrained by(3.20), Θ transforms according to $\Theta = \kappa \Theta'$. Therefore, $\Theta \neq 0$, or $\Theta = 0$ forms an invariant characteristic of the congruence. We shall call the congruence with $\Theta \neq 0$ *deviating*, and with $\Theta = 0$, *plane*.

We conclude this section by stating that left congruences of null strings can exist only in the metrical structures of CR and UR; as we shall see they are particularly interesting from the point of view of CR interpreted as complexified HR.

Parallel to left congruences of null strings, we can also consider right congruences of null strings; a general right congruence is defined by $0 \neq \dot\Sigma \in \Lambda^2$ and $\dot\theta \in \Lambda^1$ such that

$$\text{(i)}\, \dot\Sigma \wedge \dot\Sigma = 0, \quad \text{(ii)}\, *\dot\Sigma = -\dot\Sigma, \quad \text{(iii)}\, d\dot\Sigma = \dot\theta \wedge \dot\Sigma, \tag{3.27}$$

or, with $\dot\Sigma$ in canonical normalization,

$$\text{(i)}\, \dot\Sigma \wedge \dot\Sigma = 0, \quad \text{(ii)}\, *\dot\Sigma = -\dot\Sigma, \quad \text{(iii)}\, d\dot\Sigma = 0. \tag{3.28}$$

All arguments and properties established for left congruences apply *mutatis mutandis*, for the right congruences; the right deviation shall be denoted by $\dot\Theta$.

4. CONGRUENCES OF NULL STRINGS IN SPINORIAL FORMALISM

The objective of this text—to explore systematically the structures $V = (M,g)$ which admit a congruence of null strings, and to develop the theory of the "optical properties" of these congruences along similar lines as is done (within HR) for the geodesic null congruences of vector lines—can be conveniently approached by employing the spinorial formalism. The existence of the left congruence of null strings, as determined by Σ given in the canonical normalization, can be readily seen to be equivalent to the existence of a

$D(\frac{1}{2},0)$ spinor such that

$$\text{(i) } k_A \neq 0, \quad \text{(ii)} \nabla^S{}_{\dot{B}}(k_S k_A) = 0, \tag{4.1}$$

i.e., to the left "Maxwell equations" for a null $D(1,0)$ field.

Indeed, a simple and self-dual Σ must be of the form

$$\Sigma = k_A k_B S^{AB}. \tag{4.2}$$

Having fulfilled with this (3.19) (i) and (ii), one then easily finds that (3.19) (iii) amounts to

$$d\Sigma = -*g^{A\dot{B}}\nabla^S{}_{\dot{B}}(k_S k_A) = 0. \tag{4.3}$$

Now, in HR, the optical properties of a geodesic congruence of null vectors $(k_\mu : k^\mu k_\mu = 0, k^\nu k^\mu{}_{;\nu} = \lambda k^\mu)$ are characterized by the algebraic and differential properties of the matrix $\|k_{\mu;\nu}\|$. It is thus to be expected that the optical properties of Σ shall be characterized by the algebraic and differential properties of the covariant derivatives, $\Sigma_{\mu\nu;\lambda}$, equivalently $\nabla_{A\dot{B}}k_C k_D$. Therefore, the first thing to do is to establish the algebraic structure of these derivatives, compatibly with (4.1). It is easily seen that (4.1) is equivalent to the existence of a $D(\frac{1}{2},\frac{1}{2})$ vectorial object, the Sommers vector[14]

$$Z := -\tfrac{1}{2}Z_{A\dot{B}}g^{A\dot{B}} = Z_\mu \, dx^\mu, \tag{4.4}$$

such that

$$\nabla_{A\dot{B}}k_C = 3Z_{A\dot{B}}k_C + 2\epsilon_{AC}k^S Z_{S\dot{B}}. \tag{4.5}$$

If the structure $V = (M,g)$ admits a left congruence of null strings Σ, then all conformally equivalent structures $V' = (M,g')$, $g' = \Phi^{-2}g$, $\Phi \neq 0$ also admit the same congruence. This is a consequence of the fact that the Hodge star applied with respect to 2-forms in the sense of V ("$*$") and in the sense V'("$*'$") do overlap : $\Lambda^2 \ni \omega \rightarrow *\omega = *'\omega$. Therefore (3.1), and in particular (3.19), are conformally invariant formulas. (Our definition of $*$ is so arranged that in a local chart $\{x^\alpha\}$,

$$*\left(\frac{1}{2}f_{\mu\nu}dx^\mu \wedge dx^\nu\right) = \frac{1}{2}\check{f}_{\mu\nu}dx^\mu \wedge dx^\nu,$$

$$\check{f}_{\mu\nu} := \frac{i}{2\sqrt{-g}}\, g_{\mu\alpha}g_{\nu\beta}\epsilon^{\alpha\beta\gamma\delta}f_{\gamma\delta}, \tag{4.6}$$

so that $*$'s in the sense of $g_{\mu\nu}$ and $g'_{\mu\nu} = \Phi^{-2}g_{\mu\nu}$ do coincide for every $f_{\mu\nu} = f_{[\mu\nu]}$.)

Rewriting now the basic (4.5) (by contracting it with $-\frac{1}{2}g^{A\dot{B}}$) as a 1-form equation,

$$Dk^A = 3Zk^A + k^S Z_{S\dot{B}}g^{A\dot{B}}, \tag{4.7}$$

we shall now examine its covariance with respect to conformal transformations of the metric.

We acknowledge first that, by taking as the spinorial tetrad of $g' = \Phi^{-2}g$ simply

$$g'^{A\dot{B}} = \Phi^{-1}g^{A\dot{B}}, \tag{4.8}$$

it follows from $D'g'^{A\dot{B}} = 0$ that

$$\Gamma'_{AB} = \Gamma_{AB} - \tfrac{1}{2}d\ln\Phi \,\lrcorner\, S_{AB}, \quad \Gamma'_{\dot{A}\dot{B}} = \Gamma_{\dot{A}\dot{B}} - \tfrac{1}{2}d\ln\Phi \,\lrcorner\, S_{\dot{A}\dot{B}}. \tag{4.9}$$

The corresponding transformation law of the curvature coefficients is then

$$\begin{aligned} &C'_{ABCD} = \Phi^2 C_{ABCD}, \quad C'_{\dot{A}\dot{B}\dot{C}\dot{D}} = \Phi^2 C_{\dot{A}\dot{B}\dot{C}\dot{D}}, \\ &C'^{AB}{}_{\dot{C}\dot{D}} = \Phi^2[C^{AB}{}_{\dot{C}\dot{D}} - \tfrac{1}{2}\Phi^{-1}\nabla^{(A}{}_{(\dot{C}}\nabla^{B)}{}_{\dot{D})}\Phi], \\ &R' = \Phi^2[R + 6\Phi\Box\Phi^{-1}], \end{aligned} \tag{4.10}$$

where $\Box$ is defined by

$$\Box : \Lambda^0 \rightarrow \Lambda^0; \Box := -\tfrac{1}{2}\nabla^{A\dot{B}}\nabla_{A\dot{B}}. \tag{4.11}$$

[In these formulas $\nabla_{A\dot{B}}$ refers to the g-metric; the first line of (4.10) is the spinorial counterpart of "$C'^\alpha{}_{\beta\gamma\delta} = C^\alpha{}_{\beta\gamma\delta}$" of the classical Riemannian geometry.]

Now, if we understand (3.19) as conformally invariant formulas, it is natural to require $\Sigma = \Sigma'$; i.e.,

$$k_A k_B S^{AB} = k'_A k'_B S'^{AB} \tag{4.12}$$

which, with $S_{AB} = \Phi^{-2}S_{AB}$ suggests as a natural transformation law under conformal transformations

$$\mathscr{C} : k'_A = \Phi k_A. \tag{4.13}$$

One then easily finds that, assigning to Z the transformation law

$$\mathscr{C} : Z' := Z + \tfrac{1}{2}d\ln\Phi \leftrightarrow Z'_{A\dot{B}} = \Phi(Z_{A\dot{B}} + \tfrac{1}{2}\nabla_{A\dot{B}}\ln\Phi), \tag{4.14}$$

(4.7) can be equivalently stated as

$$D'k'^A = 3Z'k'^A + k'^S Z'_{S\dot{B}}g'^{A\dot{B}}, \tag{4.15}$$

with D' in the sense of Γ'_{AB}. We can therefore interpret (4.13) and (4.14) as the transformation law of $(k_A, Z_{A\dot{B}})$ under the conformal gauge $\mathscr{C}$.

It should be remarked that Sommers[14] assigns a different transformation law for k_A and $Z_{A\dot{B}}$.

There is an important relation between the Sommers vector Z, with respect to a left null string Σ in canonical normalization, and the deviation form of the string Θ. With $S^{AB} = \frac{1}{2}S^{AB}{}_{\mu\nu}dx^\mu \wedge dx^\nu$ and, therefore, $\Sigma^{\mu\nu} = k_A k_B S^{AB\mu\nu}$ used in (3.23), and employing (4.5), one easily finds out that

$$\Theta_{A\dot{B}} = 4k_A k^S Z_{S\dot{B}}, \tag{4.16}$$

which is equivalent to

$$\Theta = -2Z \,\lrcorner\, \Sigma. \tag{4.17}$$

From the last formula and (4.14) it easily follows that Θ transforms under the conformal gauge according to

$$\mathscr{C} : \Theta' = \Phi^2(\Theta - d\ln\Phi \,\lrcorner\, \Sigma), \tag{4.18}$$

compatibly with $\Theta \,\lrcorner\, \Sigma = 0 \rightarrow \Theta' \,\lrcorner'\, \Sigma' = 0$.

With respect to a fixed metrical structure (M,g), the objects $(k_A, Z_{A\dot{B}})$ related by (4.7) contain essentially all "optical" information concerning the corresponding congruence of left null strings given in canonical normalization, determining the covariant derivatives $\Sigma_{\alpha\beta;\gamma}$. One can thus attempt to subclassify our congruences according to the algebraic properties of $Z_{A\dot{B}}$.

The only invariant which one can construct from $Z_{A\dot{B}}$ is

$$\Delta \doteq Z \,\lrcorner\, Z = -\tfrac{1}{2}Z_{A\dot{B}}Z^{A\dot{B}} = Z_\mu Z^\mu. \tag{4.19}$$

Therefore, it makes sense to distinguish the general and isotropic cases

$$G : \Delta \neq 0 \quad I : \Delta = 0. \tag{4.20}$$

Notice that in the general case we have

$$G : k^S Z_{S\dot{B}} \neq 0 \tag{4.21}$$

and so $\Theta \neq 0$.

In the isotropic case $Z_{A\dot{B}}$ has the algebriac form of

$$\text{I} : Z_{A\dot{B}} = \alpha_A \beta_{\dot{B}}; \tag{4.22}$$

therefore, there is a subcase, where

$$I_0{:}Z_{A\dot{B}} = k_A\beta_{\dot{B}} \leftrightarrow k^S Z_{S\dot{B}} = 0 \leftrightarrow \Theta = 0. \tag{4.23}$$

This subclassification into the cases G, I, I_0 makes sense with respect to a fixed Riemannian structure, i.e., with the $\mathscr{C}$-gauge frozen. On the other hand, if the conformal factor is available,

$$\begin{aligned}\Delta' &= Z' \lrcorner Z' = \Phi^2(Z' \lrcorner Z') \\ &= -\tfrac{1}{2}\Phi^2(Z_{A\dot{B}} + \tfrac{1}{2}\nabla_{A\dot{B}} \ln \Phi)(Z^{A\dot{B}} + \tfrac{1}{2}\nabla^{A\dot{B}} \ln \Phi)\end{aligned} \tag{4.24}$$

is strongly dependent on the choice for that factor. In particular, it is always possible, when convenient, to select a point along the orbit of the $\mathscr{C}$-group so that

$$\mathscr{C}_0{:}\Delta = 0, \tag{4.25}$$

making Z isotropic.

The conformally invariant geometric properties of Σ are clearly characterized by the algebraic and differential properties of the $\mathscr{C}$-invariant 2-form

$$\omega{:} = dZ = {:}\tfrac{1}{2}\rho_{AB}S^{AB} + \tfrac{1}{2}\rho_{\dot{A}\dot{B}}S^{\dot{A}\dot{B}}. \tag{4.26}$$

Defining the quantities

$$r = \tfrac{1}{2}\rho_{AB}\rho^{AB}, \quad \dot{r} = \tfrac{1}{2}\rho_{\dot{A}\dot{B}}\rho^{\dot{A}\dot{B}}, \tag{4.27}$$

we have then, under $\mathscr{C}$-gauge,

$$\mathscr{C}{:}r' = \Phi^4 r, \quad \dot{r}' = \Phi^4 \dot{r}; \tag{4.28}$$

therefore, whether the "rotation coefficients" ρ_{AB} and $\rho_{\dot{A}\dot{B}}$ are, respectively, algebraically general or null, are $\mathscr{C}$-invariant properties. One also easily sees that

$$J{:} = dZ \lrcorner \Sigma = *(dZ \wedge \Sigma) = 2\rho_{AB}k^A k^B \tag{4.29}$$

is a $\mathscr{C}$-covariant, $J' = \Phi^4 J$, and therefore, $J \neq 0$, or

$$J = 0 \leftrightarrow \rho_{AB} = k_{(A}\rho_{B)} \tag{4.30}$$

are $\mathscr{C}$-invariant properties.

5. CONGRUENCES OF NULL STRINGS ON BOTH SIDES

Working, in principle on the level of CR, we shall study in this section the case where the structure $V = (M,g)$ admits simultaneously left and right congruences of null strings, i.e., we postulate the existence of (nontrivial) $\Sigma, \dot{\Sigma} \in \Lambda^2$ such that (3.19) and (3.28) are satisfied. The congruences can be then characterized by the spinor fields $k_A, k_{\dot{A}}$ and the Sommers vectors Z and $\dot{Z}$, such that (4.5) and a similar relation between $k_{\dot{A}}$ and $\dot{Z}_{A\dot{B}}$ hold:

$$\nabla_{A\dot{B}}k_{\dot{C}} = 3\dot{Z}_{A\dot{B}}k_{\dot{C}} + 2\epsilon_{\dot{B}\dot{C}}k^{\dot{S}}\dot{Z}_{A\dot{S}}. \tag{5.1}$$

In CR and UR $k_A, k_{\dot{A}}$ and $Z_{A\dot{B}}, \dot{Z}_{A\dot{B}}$ remain unrelated. In HR, with $k_{\dot{A}} = \overline{k_A}$ and $g^{A\dot{B}} = g^{B\dot{A}}, \dot{Z}$ ought to be interpreted as the complex conjugate of Z, i.e., we have there

$$\text{HR}{:}\ \overline{Z_{B\dot{A}}} = Z_{A\dot{B}}. \tag{5.2}$$

The corresponding deviations of both congruences are

$$\Theta = -2Z \lrcorner \Sigma, \quad \dot{\Theta} = -2\dot{Z} \lrcorner \dot{\Sigma}. \tag{5.3}$$

Consider now the null vector field

$$k{:} = -\tfrac{1}{2}k_A k_{\dot{B}} g^{A\dot{B}} = k_\mu\, dx^\mu. \tag{5.4}$$

This vector field is simultaneously contained in tangent spaces to both congruences; indeed we have

$$k \lrcorner \Sigma = 0, \quad k \lrcorner \dot{\Sigma} = 0. \tag{5.5}$$

The vector field k can be, therefore, thought of as generating a congruence of (null) vector lines which are the intersections of the 2-surfaces from the congruences Σ and $\dot{\Sigma}$. Employing (4.5) and (5.1), one easily shows that those lines are geodesic.

The "optical theory" of null congruences of geodesic lines, elaborated by R. Sachs[20] and others in HR, is of basic importance, because of the Goldberg–Sachs theorem,[21] in the theory of algebraically degenerate solutions of Einstein's equations (see, e.g., Ref. 21). Our congruence k, consisting of the intersection of the left and right congruences of null strings, can now be interpreted along the standard lines of this optical theory. In doing so, it is convenient to employ the vector fields

$$\begin{aligned}\mathfrak{X}{:} &= \tfrac{1}{2}(Z + \dot{Z}) = -\tfrac{1}{2}\mathfrak{X}_{A\dot{B}}g^{A\dot{B}}, \\ Y{:} &= \tfrac{1}{2}(Z - \dot{Z}) = -\tfrac{1}{2}Y_{A\dot{B}}g^{A\dot{B}}.\end{aligned} \tag{5.6}$$

Now, using (4.5) and (5.1), one easily finds for the magnification parameter of the congruence k

$$\vartheta{:} = k^\mu{}_{;\mu} = *d*k = -\tfrac{1}{2}\nabla^{A\dot{B}}(k_A k_{\dot{B}}) = \mathfrak{X}_{A\dot{B}}k^A k^{\dot{B}}. \tag{5.7}$$

Next, one finds the differential of k in the form of

$$dk = 2\mathfrak{X} \wedge k - 4*(Y \wedge k), \tag{5.8}$$

which implies, among other things, that

$$\tfrac{1}{2}*(k \wedge dk) = i\tau k, \quad i\tau{:} = Y_{A\dot{B}}k^A k^{\dot{B}}. \tag{5.9}$$

Of course, τ has the interpretation of the twist parameter of the congruence k and is real in HR.

Having established this, we can now seek the correspondence with the conventional approach to the theory of the null geodesic congruences, via the properties of connections in the corresponding tetrad gauge, as is done in Ref. 21.

The null tetrad formalism of Ref. 21, which works with

$$g = 2e^1 \underset{s}{\otimes} e^2 + 2e^3 \underset{s}{\otimes} e^4 \tag{5.10}$$

and the connections $\Gamma_{ab} = \Gamma_{[ab]}$, overlaps with the spinorial treatment when we identify

$$\begin{aligned}(g^{A\dot{B}}) &= \sqrt{2}\begin{pmatrix} e^4, & e^2 \\ e^1, & -e^3 \end{pmatrix} \\ (\Gamma_{AB}) &= -\begin{pmatrix} \Gamma_{42}, \tfrac{1}{2}(\Gamma_{12} + \Gamma_{34}) \\ \tfrac{1}{2}(\Gamma_{12} + \Gamma_{34}), \Gamma_{31} \end{pmatrix}, \\ (\Gamma_{\dot{A}\dot{B}}) &= -\begin{pmatrix} \Gamma_{41}, \tfrac{1}{2}(-\Gamma_{12} + \Gamma_{34}) \\ \tfrac{1}{2}(-\Gamma_{12} + \Gamma_{34}), \Gamma_{32} \end{pmatrix}.\end{aligned} \tag{5.11}$$

Choosing now the $SL(2,\mathbb{C}) \times \dot{SL}(2,\mathbb{C})$ gauge so that

$$k^A = 2^{1/4}\delta^A_1, \quad k^{\dot{A}} = 2^{1/4}\delta^{\dot{A}}_{\dot{1}}, \tag{5.12}$$

we have

$$k = -\tfrac{1}{2}k^A k^{\dot{B}} g_{A\dot{B}} = -(1/\sqrt{2})g^{2\dot{2}} = e^3, \tag{5.13}$$

and from (4.5) and (5.1) spelled out in that gauge we obtain

$$\begin{aligned}\Gamma_{11} &= (\mathfrak{X}_{1\dot{1}} + Y_{1\dot{1}})g^{2\dot{1}} + (\mathfrak{X}_{1\dot{2}} + Y_{1\dot{2}})g^{2\dot{2}}, \\ \Gamma_{\dot{1}\dot{1}} &= (\mathfrak{X}_{1\dot{1}} - Y_{1\dot{1}})g^{1\dot{2}} + (\mathfrak{X}_{2\dot{1}} - Y_{2\dot{1}})g^{2\dot{2}}.\end{aligned} \tag{5.14}$$

This means, however, that the congruence $k = e^3$ is geodesic ($\Gamma_{424} = 0 = \Gamma_{414}$) and shear-free ($\Gamma_{422} = 0 = \Gamma_{411}$).

The Sachs complex expansion is given by

$$z := -\Gamma_{421} = \sqrt{2}(\mathfrak{X}_{1\dot{1}} + Y_{1\dot{1}})$$
$$= (\mathfrak{X}_{A\dot{B}} + Y_{A\dot{B}})k^A k^{\dot{B}} = \vartheta + i\tau = Z_{A\dot{B}} k^A k^{\dot{B}}, \tag{5.15}$$
$$\dot{z} := -\Gamma_{412} = \sqrt{2}(\mathfrak{X}_{1\dot{1}} - Y_{1\dot{1}})$$
$$= (\mathfrak{X}_{A\dot{B}} - Y_{A\dot{B}})k^A k^{\dot{B}} = \vartheta - i\tau = \dot{Z}_{A\dot{B}} k^A k^{\dot{B}},$$

as was expected. A more compact form of these formulas is of course

$$z = \vartheta + i\tau = -2k \lrcorner Z, \quad \dot{z} = \vartheta - i\tau = -2k \lrcorner \dot{Z}. \tag{5.16}$$

Under the conformal gauge the basic objects $k_A, k_{\dot{A}}, Z$, and $\dot{Z}$ transform according to

$$\mathscr{C}: k'_A = \Phi k_A, \ Z' = Z + \tfrac{1}{2} d \ln\Phi,$$
$$k'_{\dot{A}} = \Phi k_{\dot{A}}, \ \dot{Z}' = \dot{Z} + \tfrac{1}{2} d \ln\Phi, \tag{5.17}$$

which implies $k'_\mu = \Phi k_\mu$. The congruence of vector lines determined by $k' = \Phi k$ is then a geodesic shearless congruence of null lines with respect to $g' = \Phi^{-2} g$, with the corresponding Sachs parameters

$$z' = \Phi^3 (z + \tfrac{1}{2} k^A k^{\dot{B}} \nabla_{A\dot{B}} \ln \Phi),$$
$$\mathscr{C}: \tag{5.18}$$
$$\dot{z}' = \Phi^3 (\dot{z} + \tfrac{1}{2} k^A k^{\dot{B}} \nabla_{A\dot{B}} \ln \Phi).$$

In particular, the twist, transforming according to

$$\mathscr{C}: \tau' = \Phi^3 \tau, \tag{5.19}$$

is a $\mathscr{C}$-gauge covariant.

We can now approach conveniently the important question of this section: What is the mechanism, according to which the deviation properties of the congruences of strings determined by Σ and $\dot{\Sigma}$, define the optical scalars of the congruence of lines being their intersections?

With our definitions of Θ, $\dot{\Theta}$, k, z, and $\dot{z}$, one easily finds that

$$k \wedge \Theta = z\Sigma, \quad k \wedge \dot{\Theta} = \dot{z}\dot{\Sigma}. \tag{5.20}$$

Therefore, if one wants, in HR, the congruence k to be endowed with a nontrivial Sachs parameter, $\bar{z} = z \neq 0$, then both congruences Σ and $\dot{\Sigma}$ must necessarily be deviating ($\dot{\Theta} = \Theta \neq 0$). Of course, with both congruences Σ and $\dot{\Sigma}$ plane ($\Theta = 0 = \dot{\Theta}$), they necessarily intersect along a nondiverging congruence k ($z = 0 = \dot{z}$). There is, however, a subpossibility that with $Z_{A\dot{B}} = \alpha_A k_{\dot{B}}$ and $\dot{Z}_{A\dot{B}} = k_A \alpha_{\dot{B}}$ and $\alpha^A k_A \neq 0 \neq \alpha^{\dot{A}} k_{\dot{A}}, z = \dot{z} = 0$, while $\Theta \neq 0 \neq \dot{\Theta}$, i.e., the deviating strings can, in particular, intersect along a nondiverging congruence k.

6. A CONGRUENCE OF LEFT NULL STRINGS AND LEFT CURVATURE

Given a left congruence of null strings induced by Σ (in canonical normalization), we can consider $(k_A, Z_{A\dot{B}})$ as the basic objects describing the strings' structure, which are constrained by

$$Dk^A = 3Zk^A + k^S Z_{S\dot{B}} g^{A\dot{B}}. \tag{6.1}$$

The objective of this section is to study the integrability conditions of this relation. We define

$$\gamma^{AB} := -Z \lrcorner S^{AB} \text{ and } \gamma^{\dot{A}\dot{B}} := -Z \lrcorner S^{\dot{A}\dot{B}}. \tag{6.2}$$

Then the objects

$$\tilde{\Gamma}_{AB} := \Gamma_{AB} - \gamma_{AB}, \ \tilde{\Gamma}_{\dot{A}\dot{B}} := \Gamma_{\dot{A}\dot{B}} - \gamma_{\dot{A}\dot{B}} \tag{6.3}$$

endowed with the natural transformation law of $SL(2,\mathbb{C}) \times \dot{SL}(2,\mathbb{C})$ connections, provide a new affine connection on M. Because of (4.9) and (4.14), $\tilde{\Gamma}_{AB}$ and $\tilde{\Gamma}_{\dot{A}\dot{B}}$ are conformally invariant.

Denoting now by $\tilde{D}$ the Cartan covariant differential in the sense of twiddled connections, one easily finds that

$$\tilde{D} g^{A\dot{B}} = -2Z \wedge g^{A\dot{B}} =: \mathscr{T}, \tag{6.4}$$

while (6.1) can be simply stated in the form

$$\tilde{D} k^A = 2k^A Z. \tag{6.5}$$

Investigating the integrability conditions of (6.1), it is now technically advantageous to work with its equivalent form (6.5) with connection $\tilde{\Gamma}$, and $\mathscr{T}$ interpreted as the torsion with respect to this connection.

The integrability condition of (6.5), because of $\tilde{D}\tilde{D}k^A = \tilde{R}^A{}_B k^B$, amounts now to

$$\tilde{R}^A{}_B k^B = 2k^A dZ, \tag{6.6}$$

and is, therefore, equivalent to the existence of $\chi^A \in \Lambda^2$, such that

$$\text{(i) } \tilde{R}_{AB} = k_{(A}\chi_{B)}, \ \text{(ii) } 4dZ = k^S \chi_S. \tag{6.7}$$

Notice that under the conformal gauge $\mathscr{C}$, $\tilde{R}'_{AB} = \tilde{R}_{AB}$ while $k'_A = \Phi k_A$, $dZ' = dZ$, and consequently χ^A transforms according to

$$\mathscr{C}: \chi'^A = \Phi^{-1} \chi^A. \tag{6.8}$$

The results obtained up to this point can now be expressed in terms of the original connection Γ and the coefficients of the left curvature $R^A{}_B$.

For this purpose, we first decompose the covariant gradient of $Z_{A\dot{B}}$ into its irreducible parts

$$\nabla^A{}_{\dot{C}} Z^B{}_{\dot{D}} = \sigma^{AB}{}_{\dot{C}\dot{D}} + \epsilon_{\dot{C}\dot{D}} \rho^{AB} + \epsilon^{AB} \rho_{\dot{C}\dot{D}} - \epsilon^{AB} \epsilon_{\dot{C}\dot{D}} \eta,$$
$$\sigma^{AB}{}_{\dot{C}\dot{D}} = \sigma^{(AB)}{}_{(\dot{C}\dot{D})}, \tag{6.9}$$
$$\rho_{AB} = \rho_{(AB)}, \ \rho_{\dot{A}\dot{B}} = \rho_{(\dot{A}\dot{B})}.$$

The objects

$$\rho_{AB} = \tfrac{1}{2} \nabla_{(A}{}^{\dot{S}} Z_{B)\dot{S}}, \ \rho_{\dot{A}\dot{B}} = \tfrac{1}{2} \nabla^S{}_{(\dot{A}} Z_{|S|\dot{B})} \tag{6.10}$$

coincide with the rotation coefficients in (4.26), and η given by

$$\eta = -\tfrac{1}{4} \nabla^{A\dot{B}} Z_{A\dot{B}} = \tfrac{1}{2} Z^\mu{}_{;\mu} \tag{6.11}$$

has the interpretation of the expansion of the Sommers vector. Consequently, $\sigma_{AB\dot{C}\dot{D}}$ ought to be interpreted as the shear of that vector.

The object χ^A which enters in $\tilde{R}_{AB} = k_{(A}\chi_{B)}$, we represent as spanned by S^{CD} and $S^{\dot{C}\dot{D}}$:

$$\chi^A = \chi^A{}_{\dot{C}\dot{D}} S^{\dot{C}\dot{D}} + \Psi^A{}_{CD} S^{CD} + S^{AC} \Psi_C,$$
$$\Psi_{ABC} = \Psi_{(ABC)}. \tag{6.12}$$

After some work, one can show that the conditions (6.7) (i) and (ii) reduce to the statement that $\Psi_A = 0$, and

$$\text{(i)} R/24 = -(\Delta + \eta), \ \text{(ii)} \Delta = -\tfrac{1}{2} Z_{A\dot{B}} Z^{A\dot{B}}, \tag{6.13}$$

$$\text{(i)} -\tfrac{1}{2} C_{ABCD} = k_{(A} \Psi_{BCD)}, \ \text{(ii)} 2\rho_{AB} = k^S \Psi_{SAB}, \tag{6.14}$$

(i) $\frac{1}{2}C^{AB}{}_{\dot{C}\dot{D}} = k^{(A}\chi^{B)}{}_{\dot{C}\dot{D}} - \frac{1}{2}\sigma^{AB}{}_{\dot{C}\dot{D}} + Z^{A}{}_{(\dot{C}}Z^{B}{}_{\dot{D})}$, (6.15)

(ii) $2\rho_{\dot{A}\dot{B}} = k^{S}\chi_{S\dot{A}\dot{B}}$, $(\chi_{A\dot{B}\dot{C}} = \chi_{A(\dot{B}\dot{C})})$.

The formulas (6.13), (6.14), and (6.15) can be considered as the effective form of the integrability conditions of (6.1).

Notice that, under the $\mathscr{C}$-gauge, with $k'_A = \Phi k_A$,

$$\mathscr{C}: \quad \begin{aligned} &\rho'_{AB} = \Phi^2\rho_{AB},\ \rho'_{\dot{A}\dot{B}} = \Phi^2\rho_{\dot{A}\dot{B}}, \\ &\chi'_{A\dot{B}\dot{C}} = \Phi\chi_{A\dot{B}\dot{C}},\ \Psi'_{ABC} = \Phi\Psi_{ABC}. \end{aligned} \tag{6.16}$$

At the next step one should like to investigate integrability conditions of (6.9). We present only a part of them, that which can be obtained also from Bianchi identities $\widetilde{D}\widetilde{R}_{AB} = 0$:

$$(\nabla^{R\dot{S}} + 5Z^{R\dot{S}})\chi_{R\dot{S}\dot{A}} = 0$$

and (6.17)

$$(\nabla^{S}{}_{\dot{C}} + 3Z^{S}{}_{\dot{C}})\Psi_{ABS} - (\nabla_{(A}{}^{\dot{S}} + Z_{(A}{}^{\dot{S}})\chi_{B)\dot{C}\dot{S}} = 0.$$

7. NULL STRINGS AND ALGEBRAIC DEGENERATION

The aim of this section is to examine our previous results in the case of Einstein spaces, with the traceless part of the Ricci tensor vanishing:

$$\text{ES}: C_{AB\dot{C}\dot{D}} = 0 \rightarrow R = -4\lambda = \text{const.} \tag{7.1}$$

Many of the results here are variations on Goldberg–Sachs theorem, to be found in the paper of I. Robinson and A. Schild,[22] which carries over to the complex case.[1]

Lemma I: If ES has nontrivial, algebraically degenerated, left conformal curvature, then it admits a left congruence of null strings.

Proof: Indeed, from assumption we have to consider the possibilities of C_{ABCD} having the shapes

$$\begin{aligned} &\text{Type II or } D: -\tfrac{1}{2}C_{ABCD} = k_{(A}k_B\Psi_{CD)}, && \Psi_{(1)} := \Psi_{AB}k^Ak^B \neq 0, \\ &\text{Type III}: -\tfrac{1}{2}C_{ABCD} = k_{(A}k_Bk_C\Psi_{D)}, && \Psi_{(2)} := \Psi_A k^A \neq 0, \\ &\text{Type } N: -\tfrac{1}{2}C_{ABCD} = k_Ak_Bk_Ck_D\Psi_{(3)}, && \Psi_{(3)} \neq 0, \end{aligned} \tag{7.2}$$

where $k_A \neq 0$ is a multiple Penrose spinor. But with $C_{AB\dot{C}\dot{D}} = 0$, from (2.13) (ii) we have

$$\nabla^{S}{}_{\dot{A}}C_{BCDS} = 0. \tag{7.3}$$

Then, substituting here from (7.2) and contracting the result for the first line with $k^Bk^Ck^D$, for the second with k^Ck^D, and for the third with k^D, we infer for all the cases the necessity of the condition

$$k^Ak^B\nabla_{A\dot{C}}k_B = 0. \tag{7.4}$$

This is exactly the condition for a geodesic shear-free spinor field studied in Ref. 14, which has been shown in Sec. 4 to be equivalent to the existence of a left congruence of null strings, induced by $\Sigma = k_Ak_BS^{AB}$, given in arbitrary normalization. It is then obvious for all subcases from (7.2), one can always rescale the multiple Penrose spinor so that $\Sigma = k_Ak_BS^{AB}$ is given in canonical normalization. ∎

The case when the left conformal curvature spinor vanishes is independent of interest[23] (see also Ref. 24). Then there exists a complete set of left null strings. (The terminology used will be made clear below.) A construction based on the geodesic deviation equation leads to

Lemma I': Let $C_{ABCD} \equiv 0$ and p be an arbitrary point of M. Then any undotted nonzero spinor $\mathring{k}^A$ at the point p defines a left null string passing through p. As a consequence of that, there exist, at least locally, left congruences of null strings.

Proof: Indeed, $\mathring{k}^A \neq 0$ defines a null two-dimensional subspace W_p of the tangent space at p, consisting of the vectors of the form $\mathring{k}^A l^{\dot{B}}\partial_{A\dot{B}}(p)$, where $l^{\dot{B}}$ is an arbitrary dotted spinor and $\{\partial_{A\dot{B}}\}$ is a null tetrad dual to $\{g^{A\dot{B}}\}$. Next, one can construct locally a geodesic two-dimensional surface S generated by the geodesic lines passing through p in directions of W_p. That surface is automatically a null string. To prove this, it suffices to prove that W_p propagated parallel along generators of S remains tangent to it. For this purpose consider a null geodesic line $\gamma(\tau)$ on S passing through p for $\tau = 0$, and introduce a null tetrad $\{e_a\}$ propagated parallel along it, such that $e_4(0)$ and $e_2(0)$ span W_p, and $e_4(\tau)$ is tangent to $\gamma(\tau)$; τ is an affine parameter.

Let $x^\mu(\tau,\epsilon)$ denote the family of geodesic lines forming S: $\epsilon = 0$ corresponds to the line $\gamma(\tau)$ and $x^\mu(0,\epsilon) \equiv x^\mu(p)$. Then the components of the geodesic deviation vector $\delta x^\mu(\tau) := (\partial x^\mu/\partial\epsilon)(\tau,0)$, which is tangent to S, in the null tetrad $\{e_a(\tau)\}$ satisfy the following conditions:

(i) $\hat{\delta}x^a(0) = 0$,

(ii) $\dfrac{d\hat{\delta}x^3}{d\tau}(0) = 0 = \dfrac{d\hat{\delta}x^1}{d\tau}(0)$.

Indeed, (i) follows from the fact that all geodesics forming that family intersect at p. To check (ii) we notice at first that

$$\frac{d}{d\tau}(\delta x^\mu)(0) = \frac{\partial^2 x^\mu}{\partial\tau\partial\epsilon}(0,0) = \frac{\partial^2 x^\mu}{\partial\epsilon\partial\tau}(0,0) = \frac{\partial}{\partial\epsilon}\left(\frac{\partial x^\mu}{\partial\tau}(0,\epsilon)\right)\bigg|_{\epsilon=0}.$$

Therefore, since the vectors $(\partial x^\mu/\partial\tau)(0,\epsilon)$ are contained in W_p, the same is true for their derivatives with respect to ϵ. Next, one easily checks that

$$\frac{d}{d\tau}(\hat{\delta}x^a)(0) = \frac{d}{d\tau}(e^a_\mu\delta x^\mu)(0) = e^a_\mu(0)\frac{D}{d\tau}(\delta x^\mu)(0) = e^a_\mu(0)\frac{d}{d\tau}(\delta x^\mu)(0).$$

(Here e^a_μ are the components of the dual tetrad $:e^a = e^a_\mu dx^\mu$.) Hence property (ii) follows directly.

On the other hand, from the geodesic deviation equation, if $C_{ABCD} \equiv 0$ is assumed, one infers that

$$\frac{d^2\hat{\delta}x^1}{d\tau^2} = \frac{1}{2}R_{44}\hat{\delta}x^1 - \frac{1}{2}R_{24}\hat{\delta}x^3,\ R_{44} := e^\mu_4 e^\nu_4 R_{\mu\nu}, \tag{7.5}$$

$$\frac{d^2\hat{\delta}x^3}{d\tau^2} = 0, \qquad R_{24} := e^\mu_2 e^\nu_4 R_{\mu\nu}.$$

(For the correspondence between the components of curvature tensor and its irreducible spinorial images see Ref. 15.)

The system of differential equations (7.5) with the initial conditions (i) and (ii) has only the trivial solution. This shows that $\hat{\delta}x^3 \equiv 0 \equiv \hat{\delta}x^1$; therefore, the tangent spaces to S along $\gamma(\tau)$ are spanned by $e_4(\tau)$ *and* $e_2(\tau)$ only. The same argument can be repeated for an arbitrary geodesic generator of S passing through p.

To prove the existence of a congruence of null strings in a neighborhood of any point $p \in M$, take two different left-null strings S and S' passing through that point. Let k^A and l^A denote the corresponding spinor fields along S and S', respectively, and $k^A(p)l_A(p) \neq 0$. Extend $k^A(p)$ to a spinor field, in some neighborhood U' of p on S', such that $k^A(q)l_A(q) \neq 0$ for $q \in U'$. Then construct the unique null strings through the points at U' (defined by k^A on U'). This construction thus provides a left congruence of null strings in some neighborhood U of the point p in M, assuming $C_{ABCD} = 0$. ■

Lemma II: If an ES admits a left congruence of null strings, then its left conformal curvature is algebraically degenerate.

Proof: Indeed, taking the congruence as determined by Σ given in canonical normalization, there are $(k_A, Z_{A\dot{B}})$ which satisfy (6.1) and its integrability conditions, in particular (6.14), from which it follows that

$$C_{ABCD}k^Bk^Ck^D = -\tfrac{1}{2}k_A\Psi,$$

$$\text{where } \Psi = \Psi_{ABC}k^Ak^Bk^C = 2\rho_{AB}k^Ak^B. \tag{7.6}$$

Then using (2.13) (ii) and (4.5) one easily shows that $\Psi \neq 0$ implies

$$Z_{A\dot{B}} = k_A\alpha_{\dot{B}} + \tfrac{1}{10}\nabla_{A\dot{B}} \ln \Psi, \tag{7.7}$$

and as a consequence of (6.1) $\Psi = 0$. This contradiction means that it must be $\Psi = 0$ and, therefore, C_{ABCD} is algebraically degenerate. ■

Lemma III: If a Riemannian structure with algebraically degenerate C_{ABCD} admits a left congruence of null strings, then there exist $\alpha_{\dot{B}}$ and a function $\Phi \neq 0$, such that the Sommers vector has the form

$$Z_{A\dot{B}} = k_A\alpha_{\dot{B}} - \tfrac{1}{2}\nabla_{A\dot{B}} \ln \Phi. \tag{7.8}$$

Consequently, from the point of view of the conformally equivalent space, $g' = \Phi^{-2}g$, with $Z'_{A\dot{B}} = k'_A\alpha_{\dot{B}}$ [compare(4.14)], the congruence determined by Σ is plane $(\Theta' = 0)$.

Proof: Indeed, given the structure of the null strings described by $(k_A, Z_{A\dot{B}})$, we consider a pair of simultaneous linear equations for the single function $\ln \Phi$,

$$k^S(Z_{S\dot{B}} + \tfrac{1}{2}\nabla_{S\dot{B}}\ln \Phi) = 0. \tag{7.9}$$

Denoting $\nabla_{\dot{A}} := k^S\nabla_{S\dot{A}}$, we infer that the integrability condition of these equations amounts to

$$-\tfrac{1}{2}(\nabla_{\dot{A}}\nabla_{\dot{B}} - \nabla_{\dot{B}}\nabla_{\dot{A}})\ln \Phi = \nabla_{\dot{A}}k^SZ_{S\dot{B}} - \nabla_{\dot{B}}k^SZ_{S\dot{A}}. \tag{7.10}$$

It can be shown, employing (4.5) and (7.9) that (7.10) are identically satisfied. Therefore, Eqs. (7.9) are integrable, and being assured of the existence of the Φ, the existence of $\alpha_{\dot{B}}$ in (7.8) follows.

With (7.8) being true, in the conformally equivalent space $g' = \Phi^{-2}g$, $Z'_{A\dot{B}} = k'_A\alpha_{\dot{B}}$ and we have $\Theta'_{A\dot{B}} = 0$. ■

Now we make two simple observations concerning Einstein spaces.

In ES, employing the notation of (7.2) for $\Psi_{(a)}$, from (7.3) (contracted respectively in the Cases II and D with k^Ck^D, in the Case III with k^D, and in the case N directly), one can show that (7.8) is valid with algebraically degenerate $C_{ABCD} \neq 0$, with Φ having the specific values

$$\begin{aligned} &\text{Type II or } D: && \Phi = [\Psi_{(1)}]^{-1/3}, \\ &\text{Type III}: && \Phi = [\Psi_{(2)}]^{-1}, \\ &\text{Type } N: && \Phi = \Psi_{(3)}. \end{aligned} \tag{7.11}$$

Next we remark that, because of Lemma II, (7.8) is *a fortiori* valid in ES which admits a left congruence of null strings.

Lemma IV: If a Riemannian structure with algebraically degenerate C_{ABCD} admits a left congruence of null strings, then there exists a coordinate chart $\{p^{\dot{A}}, q^{\dot{A}}\}$ and structural functions ϕ and $Q^{\dot{A}\dot{B}} = Q^{(\dot{A}\dot{B})}$ in terms of which the metric assumes the form

$$g = 2\phi^{-2}\, dq_{\dot{A}} \underset{s}{\otimes} (dp^{\dot{A}} + Q^{\dot{A}\dot{B}}dq_{\dot{B}}). \tag{7.12}$$

Proof: Indeed, under the current assumptions, Lemma III is valid and, from the point of view of $g' = \Phi^{-2}g$, the congruence determined by Σ is plane and characterized by $k'_A = \Phi k_A$, $Z'_{A\dot{B}} = k'_A\alpha_{\dot{B}}$, so that

$$D'k'^A = 3Z'k'^A. \tag{7.13}$$

Then selecting the SL(2,$\mathbb{C}$) gauge in such a way that $k'^A = \delta^A_1$ it follows from (7.13) and $D'g'^{2\dot{B}} = 0$ that there exist a SL(2,$\mathbb{C}$) gauge (denoted by the same symbols $g'^{A\dot{B}}$), and the functions $\Lambda \neq 0$, $q^{\dot{1}}, q^{\dot{2}}(dq^{\dot{1}} \wedge dq^{\dot{2}} \neq 0)$, such that

$$g'^{2\dot{B}} = \Lambda\, dq^{\dot{B}}. \tag{7.14}$$

Then from $D'g'^{A\dot{B}} = 0$ one obtains

$$dq^{\dot{1}} \wedge dq^{\dot{2}} \wedge d(\Lambda g'^{1\dot{B}}) = 0, \tag{7.15}$$

which is equivalent to the existence of $p^{\dot{B}}$, $Q'^{\dot{B}\dot{C}}$ such that

$$\Lambda g'^{1\dot{B}} = -2(dp^{\dot{B}} + Q'^{\dot{B}\dot{C}}dq_{\dot{C}}). \tag{7.16}$$

(In the last statement, we apply the theorem that if $x^i \in \Lambda^0(M_n)$, $\omega := dx^1 \wedge dx^2 \wedge \ldots \wedge dx^r \neq 0$, and $e \in \Lambda^1(M_n)$, then the condition $\omega \wedge de = 0$ is equivalent to the existence of $(x, y_i) \in \Lambda^0(M_n)$ such that $e = dx + \Sigma^r_{i=1}y_i dx^i$.) From (7.14) and (7.16) it follows that

$$0 \neq g'^{1\dot{1}} \wedge g'^{1\dot{2}} \wedge g'^{2\dot{1}} \wedge g'^{2\dot{2}} = 4dp^{\dot{1}} \wedge dp^{\dot{2}} \wedge dq^{\dot{1}} \wedge dq^{\dot{2}}, \tag{7.17}$$

and so $\{p^{\dot{A}}, q^{\dot{A}}\}$ constitute a chart. Substituting now from (7.14) and (7.16) into $g' = -\tfrac{1}{2}g'_{A\dot{B}} \underset{s}{\otimes} g'^{A\dot{B}} = -g'^2{}_{\dot{B}} \underset{s}{\otimes} g'^{1\dot{B}}$

we have that

$$g' = 2dq_{\dot A} \underset{s}{\otimes} (dp^{\dot A} + Q^{\dot A\dot B} dq_{\dot B}), \tag{7.18}$$

where $Q_{\dot A\dot B} := Q'_{(\dot A\dot B)}$. On the other hand, for the original metric we have $g = \phi^{-2} g', \phi := \Phi^{-1}$; therefore, Lemma IV is true. ∎

As the next point, we should like to examine some specific properties of the deviation form Θ in the subcase of an Einstein space which admits a left congruence of null strings. Quite generally, postulating only the existence of a left congruence of null strings (given in canonical normalization), using (4.5) and (6.9) and the definition of $\Theta_{A\dot B}$ from (4.16), we can evaluate $\nabla^{A}{}_{\dot C}\Theta^{B}{}_{\dot D}$, which, decomposed into irreducible parts, yields

$$\begin{aligned} -\tfrac14 \nabla^{A\dot B}\Theta_{A\dot B} &= 2\rho_{AB}k^A k^B, \\ \tfrac14 \nabla^{S}{}_{(\dot A}\Theta_{|S|\dot B)} &= -k_R k_S \sigma^{RS}{}_{\dot A\dot B}, \\ \tfrac14 \nabla_{(A}{}^{\dot S}\Theta_{B)\dot S} &= -2(\eta + 4\Delta)k_A k_B + 2k^R k_{(A}\rho_{B)R}, \\ \tfrac14 \nabla^{(A}{}_{(\dot C}\Theta^{B)}{}_{\dot D)} &= -k_R\,[k^{(A}\sigma^{R)B}{}_{\dot C\dot D} \\ &\quad + 6k^{(A}Z^{B}{}_{\dot C}Z^{R)}{}_{\dot D}] + k^A k^B \rho_{\dot C\dot D}. \end{aligned} \tag{7.19}$$

Knowing this, one easily determines the general expression for the twist of Θ

$$*(\Theta \wedge d\Theta) = 2\rho_{AB}k^A k^B \Theta - 4k_A k^R Z_R{}^{\dot S} k_P k_Q \sigma^{PQ}{}_{\dot S\dot B} g^{A\dot B}. \tag{7.20}$$

Specializing these formulas for the case of ES, we have, according to Lemma II, C_{ABCD} algebraically degenerate, with $\rho_{AB}k^A k^B = 0$. Moreover, according to (6.15), we have

$$\text{(i)}\quad k^{(A}\chi^{B)}{}_{\dot C\dot D} - \tfrac12 \sigma^{AB}{}_{\dot C\dot D} + Z^{A}{}_{(\dot C}Z^{B}{}_{\dot D)} = 0, \tag{7.21}$$

$$\text{(ii)}\, 2\rho_{\dot A\dot B} = k^S \chi_{S\dot A\dot B},$$

from which, one obtains

$$-k^A\rho_{\dot C\dot D} - \tfrac12 k_R \sigma^{AR}{}_{\dot C\dot D} + k_R Z^{A}{}_{(\dot C}Z^{R}{}_{\dot D)} = 0. \tag{7.22}$$

Using this information, we infer that in ES the deviation form has vanishing twist:

$$ES{:}\ast(\Theta \wedge d\Theta) = 0, \tag{7.23}$$

while from the first of the formulas in (7.19) we see that the codifferential of Θ vanishes:

$$\text{ES}{:}\ast d \ast \Theta = 0. \tag{7.24}$$

At the same time, the shear of Θ reduces to

$$\text{ES}{:}\tfrac14 \nabla^{(A}{}_{(\dot C}\Theta^{B)}{}_{\dot D)} = 3[k^A k^B \rho_{\dot C\dot D} - 2k_R k^{(A}Z^{B)}{}_{(\dot C}Z^{R}{}_{\dot D)}] \tag{7.25}$$

and the rotation coefficients reduce to

$$\text{ES:}\quad \begin{aligned} \tfrac14 \nabla^{S}{}_{(\dot A}\Theta_{|S|\dot B)} &= -2k^R k^S Z_{R\dot A} Z_{S\dot B}, \\ \tfrac14 \nabla_{(A}{}^{\dot S}\Theta_{B)\dot S} &= [-2(\eta + 4\Delta) + \tfrac13 \Psi_{RS}k^R k^S]k_A k_B, \end{aligned} \tag{7.26}$$

Ψ_{AB} being the object from (7.2).

Lemma V: In ES with $\lambda = 0$, the plane congruences of left null strings ($\Theta = 0$) do not exist if C_{ABCD} is of the Type II or D, while with $\lambda \neq 0$, they do not exist if C_{ABCD} is of Type [-], N or III. Moreover, if an ES admits a plane congruence of null strings, then $\rho_{\dot A\dot B} = 0$.

Proof: Indeed, $\Theta = 0$ implies $Z_{A\dot B} = k_A \alpha_{\dot B}$ and consequently $\Delta = 0$. Then from (7.26) and (6.13) it follows that

$$\Psi_{AB}k^A k^B = \lambda. \tag{7.27}$$

Therefore, with $\lambda = 0$, the Type II and D are contradictory, while for $\lambda \neq 0$, the Type [-], N and III are impossible.

The condition $\rho_{\dot A\dot B} = 0$ follows from (7.25). ∎

We notice also that in ES with a plane congruence of left null strings, the invariants of the left conformal curvature must be constant. Indeed, with $-\tfrac12 C_{ABCD} = k_{(A}k_B\Psi_{CD)}$ and (7.27) valid we have

$$6C_{ABCD}C^{ABCD} = \lambda^2. \tag{7.28}$$

At last we notice that the structure of the Sommers vector as given by (7.8) can be still simplified in most degenerated cases.

Lemma VI: Let the Riemennian structure with C_{ABCD} of the Type III, N or [-] admits a left congruence of null strings. Then there exist $\Phi \neq 0$ and χ such that

$$\text{(i)} Z_{A\dot B} = k_A \alpha_{\dot B} - \tfrac12 \nabla_{A\dot B} \ln \Phi, \text{(ii)} \alpha_{\dot B} := \Phi^2 k_S \nabla^{S}{}_{\dot B}\chi, \tag{7.29}$$

where

$$\nabla^{A\dot B}(\Phi^{-2}\nabla_{A\dot B}\chi) = \begin{cases} 0 & \text{for Types } N \text{ and [-]} \\ \neq 0 & \text{for Type III,} \end{cases} \tag{7.30}$$

while the scalar curvature has the form

$$\frac{R}{24} = -\tfrac14 \Phi \Box \Phi^{-1}. \tag{7.31}$$

Proof: Indeed, via Lemma III, we are sure of the existence of $\Phi \neq 0$ and $\alpha_{\dot B}$ with which (7.29) (i) is valid. Consider, then, with $k_A, \Phi \neq 0$ and $\alpha_{\dot B}$ known, the pair of linear equations for the single function χ

$$\Phi^{-2}\alpha_{\dot B} = k_S \nabla^{S}{}_{\dot B}\chi. \tag{7.32}$$

One can show that the integrability conditions of (7.32) are satisfied iff C_{ABCD} is of the Type III, N or [-]. This proves (7.29). Then one can derive (7.30) from (6.14). Working out now, from (7.29), Δ and η, one finds according to (6.13), that

$$\frac{R}{24} = -(\Delta + \eta) = -\frac14 \Phi \Box \Phi^{-1} + \frac14 k_A(\nabla^{A\dot B}\alpha_{\dot B} - \frac32 \alpha_{\dot B}\nabla^{A\dot B}\ln \Phi), \tag{7.33}$$

which, with integrability conditions of (7.32) being valid, reduces to (7.31). ∎

Suppose now that, under the assumptions of Lemma VI, we consider the conformally equivalent metric,

$$g' = \Phi^{-2} g, \tag{7.34}$$

Φ being the same function which appears in (7.29). The left conformal curvature $C'_{ABCD} = \Phi^2 C_{ABCD}$ remains, of course of the Type III, N or [-] correspondingly, and the congruence Σ from the point of view of g' is now plane, characterized by

$$k'_A = \Phi k_A,\ Z'_{A\dot B} = k'_A \alpha_{\dot B},\ \alpha_{\dot B} = k'_S \nabla'^{S}{}_{\dot B}\chi. \tag{7.35}$$

This follows from (4.14) and the fact that, on Λ^0, $\nabla'_{A\dot B} = \Phi \nabla_{A\dot B}$.] Then, from (7.31) and the last line of (4.10) it follows that

$$\text{R}' = 0. \tag{7.36}$$

The condition (7.30) assumes now the form

$$\Box'\chi = \begin{matrix} \nearrow 0 \text{ for Types } N \text{ and } [-] \\ \searrow \neq 0 \text{ for Type III.} \end{matrix} \tag{7.37}$$

[This is so, because left-hand member of (7.30) has the invariant form of $*d*(\Phi^{-2}d\chi)$ and, on Λ^1:$* = \Phi^2*'$, while on Λ^4:$* = \Phi^{-4}*'$.]

Let (M,g) now be an ES which admits a left congruence of null strings characterized by

C_{ABCD} of Type III,N or [-],

$$C_{AB\dot{C}\dot{D}} = 0,\ R = -4\lambda = \text{const.} \tag{7.38}$$

The Lemma VI applies, and consequently according to (7.31) the function Φ^{-1} must fulfill a Klein–Gordon (i.e., Helmholtz) equation:

$$(\Box - \tfrac{2}{3}\lambda)\Phi^{-1} = 0. \tag{7.39}$$

The conformally equivalent structure $(M,g' = \Phi^{-2}g)$ is now characterized, according to (4.10), by the curvature coefficients

$$\begin{gathered} C'_{ABCD} \text{ of Type III, } N \text{ or } [-],\ R' = 0, \\ C'^{AB}{}_{\dot{C}\dot{D}} = -\tfrac{1}{2}\Phi\nabla^{(A}{}_{(\dot{C}}\nabla^{B)}{}_{\dot{D})}\Phi. \end{gathered} \tag{7.40}$$

But from (4.10), replacing the unprimed quantities (including covariant gradients) by primed and vice versa, and Φ by Φ^{-1}, we have the identities

$$\begin{gathered} \text{(i)}\ \Phi^{-2}\Box'\Phi + \Phi^2\Box\Phi^{-1} = 0, \\ \text{(ii)}\nabla'^{(A}{}_{(\dot{C}}\nabla'^{B)}{}_{\dot{D})}\Phi^{-1} + \nabla^{(A}{}_{(\dot{C}}\nabla^{B)}{}_{\dot{D})}\Phi = 0. \end{gathered} \tag{7.41}$$

It follows that (7.39) can be equivalently stated as

$$\Box'\Phi + \tfrac{2}{3}\lambda\Phi^3 = 0, \tag{7.42}$$

while $C'^{AB}{}_{\dot{C}\dot{D}}$ has the form

$$C'^{AB}{}_{\dot{C}\dot{D}} = \tfrac{1}{2}\Phi\nabla'^{(A}{}_{(\dot{C}}\nabla'^{B)}{}_{\dot{D})}\Phi^{-1}. \tag{7.43}$$

One can discuss further the integrability conditions of the string structure as referred to the metric g'. We notice then that $\chi'^{A}{}_{\dot{C}\dot{D}}$ can be represented as

$$\chi'^{A}{}_{\dot{C}\dot{D}} = -k'_S\nabla'^{A}{}_{(\dot{C}}\nabla'^{S}{}_{\dot{D})}\chi + 2k'^{A}(\alpha_{\dot{C}}\alpha_{\dot{D}} + \nu_{\dot{C}\dot{D}}), \tag{7.44}$$

where $\nu_{\dot{C}\dot{D}} = \nu_{(\dot{C}\dot{D})}$ is some spinor field, such that

$$\begin{aligned} &\Phi\nabla'^{(A}{}_{(\dot{C}}\nabla'^{B)}{}_{\dot{D})}\Phi^{-1} + 6k'_S k'^{(A}\nabla'^{B)}{}_{(\dot{C}}\nabla'^{S}{}_{\dot{D})}\chi \\ &\quad = 8k'^{A}k'^{B}\nu_{\dot{C}\dot{D}}. \end{aligned} \tag{7.45}$$

Below we summarize the main results of this section.

Theorem: An Einstein space (abbreviated to ES) admits a left congruence of null strings if and only if the Weyl spinor C_{ABCD} is algebraically degenerate.

The number of left congruences of null strings in an ES is equal to the number of degenerate Debever–Penrose spinor directions. For an ES with $C_{ABCD} \equiv 0$ there are infinitely many local left congruences.

We also have:

(1) The metric of an Einstein space admitting a left congruence of null strings can be represented in the form

$$g = 2\phi^{-2}\, dq_{\dot{A}} \underset{s}{\otimes} (dp^{\dot{A}} + Q^{\dot{A}\dot{B}}\, dq_{\dot{B}}),\ \text{where } Q_{\dot{A}\dot{B}} = Q_{(\dot{A}\dot{B})}.$$

(2) By the conformal rescaling $g' = \phi^2 g$, the congruence $q^{\dot{A}} = \text{const}$ becomes plane.

(3) Any left congruence of null strings in an Einstein space with C_{ABCD} of the Type II or D and with $\lambda = 0$, or with C_{ABCD} of the Type [-], or N or III and $\lambda \neq 0$, is necessarily deviating. ■

As a corollary, notice that our sequence of lemmas provides, among other things, an extension of the theorem of Plebański–Hacyan,[1] which states that the left algebraic degeneration is equivalent to the existence of a congruence of left null strings—modulo the homogeneous Einstein equations—to the case of Einstein spaces with nontrivial λ.

ACKNOWLEDGMENTS

We want to express our gratitude to J. D. Finley III and I. Robinson for helpful discussions and critical remarks.

[1] J. F. Plebański and S. Hacyan, J. Math. Phys. **16**, 2403 (1975).
[2] J. F. Plebański and I. Robinson, Phys. Rev. Let. **37**, 493 (1976).
[3] J. F. Plebański and I. Robinson, in *Proceeding of Symposium on Asymptotic Structures of Space-Time*, Cincinnati, June 1977, edited by F. Eposito and L. Witten (Plenum, New York, 1977).
[4] C. P. Boyer, J. D. Finley, III, and J. F. Plebański, *Einstein Memorial Volume*, Society for General Relativity and Gravitation (Plenum, New York, 1979).
[5] J. D. Finley, III, and J. F. Plebański, J. Math. Phys. **17**, 2207 (1976).
[6] J. D. Finley, III, and J. F. Plebański, J. Math. Phys. **18**, 1662 (1977).
[7] K. Rózga, Rep. Math. Phys. **11**, 197 (1977).
[8] J. D. Finley, III, and J. F. Plebański, J. Math. Phys. **22**, 667 (1981).
[9] C. W. Fette, A. I. Janis, and E. T. Newman, J. Math. Phys. **17**, 660 (1976); Gen. Rev. Grav. **8**, 29 (1977).
[10] J. F. Plebański, J. Math. Phys. **16**, 2395 (1975).
[11] J. D. Finley, III, and J. F. Plebański, J. Math. Phys. **17**, 585 (1976).
[12] C. P. Boyer and J. F. Plebański, J. Math. Phys. **18**, 1022 (1977).
[13] E. Flaherty, *Lecture Notes in Physics*, Vol. 46 (Springer, Heidelberg, 1976).
[14] P. Sommers, Proc. R. Soc. London Ser. A **349**, 309 (1976).
[15] J. F. Plebański, *Spinors, Tetrads and Forms* (Monograph of Centro de Investigacion y Estudios Avanzados del I.P.N., Mexico City, 1974).
[16] I. Robinson, *Space-Time: the Alfred Schild Lectures*, (University of Texas Press, 1981).
[17] E. T. Newman and R. Penrose, J. Math. Phys. **3**, 566 (1962).
[18] F. J. Ernst, J. Math. Phys. **19**, 489 (1978).
[19] A. Schild, Phys. Rev. D **16**, 1722 (1987).
[20] R. K. Sachs, in *Recent Development in General Relativity*, dedicated to Leopold Infeld (Pergamon, New York, and PWN, Warsaw, 1962).
[21] G. C. Debney, R. P. Kerr, and A. Schild, J. Math. Phys. **10**, 1842 (1969).
[22] I. Robinson and A. Schild, J. Math. Phys. **4**, 484 (1963).
[23] R. Penrose, Gen. Rel. Grav. **7**, 31 (1976).
[24] C. R. Le Brun, Proc. R. Soc. London Ser. A **380**, 171 (1982).

Congruences of null strings in complex space-times and some Cauchy–Kovalevski-like problems

Ivor Robinson and Krzysztof Rózga
Program in Mathematical Sciences, The University of Texas at Dallas, Richardson, Texas 75080

(Received 30 December 1982; accepted for publication 9 November 1983)

It is shown that a problem of construction of a local congruence of null strings is equivalent to a natural Cauchy–Kovalevski-like problem, related to an equation for a spinor field k_A defining the congruence. Initial data are specified on two-dimensional submanifolds. In left-conformally-flat spaces, the solution of that problem exists for arbitrary initial data.

PACS numbers: 04.20.Cv

1. INTRODUCTION

It is the existence of a complete set of two-dimensional, totally null (in the sense of vanishing metric) and self-dual submanifolds, which distinguishes complex left-conformally-flat space-times among others.[1] These submanifolds are referred to as null strings or twistor surfaces. For any point p of such a space (M,ds^2) and any null two-dimensional, self-dual subspace W_p of the tangent space $T_p(M)$, there exists a null string tangent to W_p at p.[2] (See also Ref. 3.)

There are, however, congruences of null strings rather than single null strings, which are of great importance when one is trying to find an analytic form of the metric field.[4] In heavens (left-flat spaces) there is a special one-parameter family of nonexpanding congruences of null strings. This fact makes it possible to provide a description of heavens in terms of deformed projective twistor spaces[1] (see also Refs. 5 and 6) as well as to find an almost explicit analytic form of the metric field.[7] (See also Ref. 8.)

As a natural generalization of heavens, one obtains "projective extensions of heavens"[9] characterized by the existence of generalized canonical families of null strings.[10] These spaces still do not exhaust all left-conformally-flat ones.

The objective of this paper is to study the mechanism of generation of local congruences of null strings in left-conformally-flat spaces.

The way in which the constructions can be carried out is essentionally obvious from the geometrical point of view. Indeed, suppose we want to determine a congruence of null strings in some neighborhood of the point $p \in M$. For this purpose, take any two-dimensional submanifold $s:\mathscr{T} \to M$ passing through that point. Next, consider a holomorphic distribution $\widetilde{D}$ of null, two-dimensional and self-dual subspaces of the tangent spaces to M, defined on some open neighborhood $\mathscr{U} \subset \mathscr{T}$ of the point p, such that it is transversal to $\mathscr{U}$. Thus we assume that there is a holomorphic mapping $\widetilde{D}:\mathscr{U} \ni q \to \widetilde{D}(q) \subset T_q(M)$, such that $\dim \widetilde{D}(q) = 2$, $\widetilde{D}(q)$ is null and self-dual and $\widetilde{D}(q) \oplus s_* T_q(\mathscr{T}) = T_q(M)$, for $q \in \mathscr{U}$.[11] Consider next all local two-dimensional geodesic submanifolds generated by geodesic lines passing through the points of $\mathscr{U}$ in directions of $\widetilde{D}(q)$, for $q \in \mathscr{U}$. In this way one obtains a set of null strings,[2] which on some open neighborhood of p, $U \subset M$ defines a congruence. It is also clear that any congruence of left null strings can be obtained by this method, at least locally. In left-conformally-flat spaces there are natural candidates for $\mathscr{T}$, the null strings passing through p. A construction of a congruence of null strings can be carried out also in a slightly different way. Given $\widetilde{D}$ on $\mathscr{T}$, one extends it to on integrable, two-dimensional distribution of totally null, self-dual subspaces and next one constructs its integral submanifolds (null strings).

All these structures can be restated conveniently in terms of spinors.[3] Indeed, any congruence Σ of self-dual null strings defines and is defined locally by an integrable, simple and self-dual 2-form, denoted by the same symbol Σ; more precisely by a class of proportional 2-forms (integrability means, that $d\Sigma = \omega \wedge \Sigma$ for some 1-form ω). Consequently $\Sigma = k_A k_B S^{AB}$, where the S^{AB} constitute a basis for self-dual 2-forms.[11] Now it is convenient to rescale it in such a way that the new 2-form is closed. Then the corresponding nowhere vanishing spinor field k_A is said to be in canonical normalization, and it satisfies an equation of the form[3]

$$\nabla_{A\dot{B}} k_C = 3Z_{A\dot{B}} k_C + 2\epsilon_{AC} k^S Z_{S\dot{B}}, \tag{1.1}$$

where $Z_{A\dot{B}}$ is another spinor field, known as Sommers' vector. It turns out that the geometrical construction of congruences described before, is equivalent to some Cauchy–Kovalevski-like problem related to Eq. (1.1). We discuss it in Sec. 2. As a result, a proof of the existence of local congruences of self-dual null strings in left-conformally-flat spaces is provided.

Any congruence of null strings Σ can be described locally in terms of two independent functions Z^A, $A = 1,2$, as the set of submanifolds defined by the equations $Z^A = \text{const}$. Given a spinor field k_A, such that (1.1) holds, the existence of those functions follows from the fact that the corresponding 2-form Σ is closed, and hence locally

$$\Sigma = 2dZ^1 \wedge dZ^2. \tag{1.2}$$

The equations on Z^A,s take an especially simple form if a coordinate system based on two transversal congruences of null strings Σ_1 and Σ_2 is introduced. Then one obtains the corresponding Cauchy–Kovalevski-like problem. It is discussed in Sec. 3.

2. THE EXISTENCE OF CONGRUENCES OF NULL STRINGS AND A CAUCHY–KOVALEVSKI-LIKE PROBLEM

Let (M,ds^2) be a complex space-time and let p be any point of M. Take an arbitrary two-dimensional, holomorphic distribution D defined on some open neighborhood U of the

 0022-2488/84/061941-06$02.50

point p. Suppose D is integrable. Thus for U being sufficiently small there exist two independent vector fields X and Y, which span that distribution. Moreover there is a coordinate system $\{q^A, p^B\}$ on U such that the equations $q^A = \text{const}$, $A = 1,2$, define integral submanifolds of D and, therefore, X and Y are spanned by the vector fields $\{\partial/\partial p^A\}_{A=1,2}$ only.

Suppose there is on U a congruence of self-dual null strings transversal to the distribution D. Therefore, there exist on U (provided it is sufficiently small) spinor fields k_A and $Z_{A\dot{B}}$ (k_A nowhere vanishing), such that Eq. (1.1) is satisfied. The condition of transversality means that for any point $q \in U$ a direct sum of the tangent spaces at q to a null string and to an integral submanifold of D, both of them passing through q, is equal to the tangent space $T_q(M)$. This property can be restated conveniently in terms of components of X and Y in some null-tetrad $\{\partial_{A\dot{B}}\}$.[12] Indeed, the tangent spaces to null strings of the congruence are spanned by the vectors of the form $k^A l^{\dot{B}} \partial_{A\dot{B}}$ with $l^{\dot{B}}$ being arbitrary. Consequently, the congruence is transversal to D iff the spinor fields

$$\sigma^{\dot{B}} := k_A X^{A\dot{B}} \quad \text{and} \quad \sigma^{\dot{B}} := k_A Y^{A\dot{B}}, \tag{2.1}$$

where $X^{A\dot{B}}$ and $Y^{A\dot{B}}$ are defined by $X = -\frac{1}{2}X^{A\dot{B}}\partial_{A\dot{B}}$ and $Y = -\frac{1}{2}Y^{A\dot{B}}\partial_{A\dot{B}}$, respectively, are linearly independent (this condition does not depend on the particular choice of X and Y).

Next we show that the Sommers vector $Z_{A\dot{B}}$ is determined completely by covariant derivatives of k_C in directions of D. For this purpose it is convenient to introduce the following notation:

$$\nabla_X := -\tfrac{1}{2}X^{A\dot{B}}\nabla_{A\dot{B}},\ \nabla_Y := -\tfrac{1}{2}Y^{A\dot{B}}\nabla_{A\dot{B}},\ Z_X := X^{A\dot{B}}Z_{A\dot{B}},$$
$$Z_Y := Y^{A\dot{B}}Z_{A\dot{B}}, \quad \text{and} \quad \Theta_{\dot{B}} := k^S Z_{S\dot{B}}. \tag{2.2}$$

Then from (1.1) one easily infers that

$$-2\nabla_X k_C = 3Z_X k_C - 2X_C{}^{\dot{B}}\Theta_{\dot{B}}$$
and
$$-2\nabla_Y k_C = 3Z_Y k_C - 2Y_C{}^{\dot{B}}\Theta_{\dot{B}}. \tag{2.3}$$

Next, by contraction of both sides of (2.3) with k^C and employing the fact that $\sigma_{\dot{B}}$ and $\rho_{\dot{B}}$ (2.1) are linearly independent, one obtains

$$\Theta_{\dot{B}} = (\rho^{\dot{N}}\sigma_{\dot{N}})^{-1}\{(k^C\nabla_X k_C)\rho_{\dot{B}} - (k^C\nabla_Y k_C)\sigma_{\dot{B}}\}. \tag{2.4}$$

(One can easily check that the form of $\Theta_{\dot{B}}$ does not depend on a specific choice of vector fields X and Y which span D.)

The Sommers vector $Z_{A\dot{B}}$ can be represented as a linear combination of $X_{A\dot{B}}$, $Y_{A\dot{B}}$, $k_A\sigma_{\dot{B}}$ and $k_A\rho_{\dot{B}}$. One easily finds that

$$\begin{aligned} Z_{A\dot{B}} = {} & (\rho^{\dot{N}}\sigma_{\dot{N}})^{-1}\{(\sigma^{\dot{M}}\Theta_{\dot{M}})Y_{A\dot{B}} - (\rho^{\dot{M}}\Theta_{\dot{M}})X_{A\dot{B}}\} \\ & + (\rho^{\dot{N}}\sigma_{\dot{N}})^{-2}k_A\sigma_{\dot{B}}\,[(\rho^{\dot{M}}\Theta_{\dot{M}})(Y^{C\dot{D}}X_{C\dot{D}}) \\ & - (\sigma^{\dot{M}}\Theta_{\dot{M}})(Y^{C\dot{D}}Y_{C\dot{D}})] \\ & + (\rho^{\dot{N}}\sigma_{\dot{N}})^{-2}k_A\,\rho_{\dot{B}}\,[(\sigma^{\dot{M}}\theta_{\dot{M}})(X^{C\dot{D}}Y_{C\dot{D}}) \\ & - (\rho^{\dot{M}}\Theta_{\dot{M}})(X^{C\dot{D}}X_{C\dot{D}})] \\ & + (\rho^{\dot{N}}\sigma_{\dot{N}})^{-1}(k_A Z_Y\sigma_{\dot{B}} - k_A Z_X\,\rho_{\dot{B}}), \end{aligned} \tag{2.5}$$

where $\Theta_{\dot{A}}$ is defined by (2.4), and $k_A Z_X$ and $k_A Z_Y$ can be obtained from (2.3).

Thus we see that it is the tangential counterpart of (1.1), which can be satisfied by an arbitrary spinor field k_A, such that the corresponding spinor fields $\sigma_{\dot{B}}$ and $\rho_{\dot{B}}$ are lineary independent, provided the Sommers vector $Z_{A\dot{B}}$ is defined according to (2.5). The transversal counterpart of (1.1) provides a nontrivial constraint on k_A

$$k^A\nabla_{A\dot{B}}k_C = k_C\Theta_{\dot{B}}, \quad \text{which depends on } D. \tag{2.6}$$

Now we infer a Lemma.

Lemma 1: Let D be a two-dimensional distribution defined on some open subset $U \subset M$. Then:

(i) Any solution k^A of Eq. (2.6), nowhere vanishing on U, such that the spinor fields $\sigma_{\dot{B}}$ and $\rho_{\dot{B}}$ are linearly independent, defines a congruence of self-dual null strings transversal to D.

(ii) If D' is another two-dimensional distribution, such that $\sigma_{\dot{B}}'$ and $\rho_{\dot{B}}'$ related to it by (2.1) with the same k^A are linearly independent, then the spinor fields $\Theta_{\dot{B}}'$ and $Z'_{A\dot{B}}$ are identical with $\Theta_{\dot{B}}$ and $Z_{A\dot{B}}$, respectively.

Proof: Indeed, for k^A being a solution of (2.6), with $\Theta_{\dot{B}}$ defined in terms of D, it follows from the discussion above that k^A is also a solution of (1.1).

To prove (ii), assume that the vector fields X' and Y' span the distribution D'. Then they can be represented in the form of

$$\begin{aligned} {}'X^{A\dot{B}} &= aX^{A\dot{B}} + bY^{A\dot{B}} + \alpha k^A\sigma^{\dot{B}} + \beta k^A\rho^{\dot{B}}, \\ {}'Y^{A\dot{B}} &= cX^{A\dot{B}} + dY^{A\dot{B}} + \gamma k^A\sigma^{\dot{B}} + \delta k^A\rho^{\dot{B}}, \end{aligned} \tag{2.7}$$

with $a, b, c, d, \alpha, \beta, \gamma$, and δ being some functions. Hence ${}'\sigma^{\dot{B}} = a\sigma^{\dot{B}} + b\rho^{\dot{B}}$ and ${}'\rho^{\dot{B}} = c\sigma^{\dot{B}} + d\rho^{\dot{B}}$, and necessarily $ad - bc \neq 0$. Further ${}'\Theta_{\dot{B}}$ can be calculated according to (2.4). Finally employing (2.6) one obtains ${}'\Theta_{\dot{B}} = \Theta_{\dot{B}}$. In the same way one checks that ${}'Z_{A\dot{B}} = Z_{A\dot{B}}$. ■

We now study Eq. (2.6) further. In the coordinate system $\{q^A, p^B\}$, related to an integrable distribution D, the vector fields $\partial_{A\dot{B}}$ can be represented in the form of

$$\partial_{A\dot{B}} = \mu^C{}_{\dot{B}A}\frac{\partial}{\partial q^C} + \nu^C{}_{\dot{B}A}\frac{\partial}{\partial p^C}, \tag{2.8}$$

with $\mu^C{}_{\dot{B}A}$ and $\nu^C{}_{\dot{B}A}$ being some holomorphic functions. We have also $X = -\frac{1}{2}X^{A\dot{B}}\mu^C{}_{\dot{B}A}\partial/\partial p^C$, and $Y = -\frac{1}{2}Y^{A\dot{B}}\nu^C{}_{\dot{B}A}\partial/\partial p^C$ since X and Y are spanned by $\partial/\partial p^C$. Now, since X, Y, $k^A\partial_{A\dot{1}}$ and $k^A\partial_{A\dot{2}}$ are assumed to be linearly independent [this is equivalent to the linear independence of $\sigma_{\dot{B}}$ and $\rho_{\dot{B}}$ determined by (2.1)], it follows that the (2×2) matrix

$$(b^A{}_{\dot{B}}) := (k^S\mu^A{}_{S\dot{B}}) \tag{2.9}$$

is nonsingular. This fact, together with an identity

$$\nabla_{A\dot{B}}(\equiv\nabla_{\partial_{A\dot{B}}}) = \mu^C{}_{\dot{B}A}\nabla_{\partial/\partial q^C} + \nu^C{}_{\dot{B}A}\nabla_{\partial/\partial p^C}, \tag{2.10}$$

which is the result of (2.8) and properties of a covariant derivative imply that, as far as transversal to D congruences are considered, Eq. (2.6) is equivalent to another one of the following general form:

$$\frac{\partial\theta^\alpha}{\partial q^A} = H^\alpha{}_A\left(q^B, p^B, \theta^\beta\frac{\partial\theta^\beta}{\partial p^B}\right), \tag{2.11}$$

where $\alpha, \beta, = 1,2,\dots,N$ and $H^\alpha{}_A$ is linear in $\partial\theta^\beta/\partial p^B$. In our case $N = 2$ and $(\theta^\alpha) = (k^A)$.

There is a natural *Cauchy–Kovalevski-like problem* related to (2.11) and so to (2.6) as well. Indeed, consider a point $p \in U$ and a two-dimensional submanifold $\mathcal{T} \subset U$ determined by the equations $q^A = q^A(p)$. ($\mathcal{T}$ is an integral submanifold of D passing through the point p.) Suppose the *initial data* are specified in some open neighborhood $\mathcal{U}$ of the point p on $\mathcal{T}$, $\mathcal{U} \subset \mathcal{T}$, i.e., holomorphic functions $\tilde{\theta}^\alpha$ on $\mathcal{U}$ are given, such that for any point $q \in \mathcal{U}$, $(q^B, p^B, \tilde{\theta}^\beta\, \partial\tilde{\theta}^B/\partial p^C)$ belongs to the domain of Eq. (2.11). Then the problem is to find a local solution of (2.11), which agrees with those initial data. (The standard Cauchy–Kovalevski theorems of existence and uniqueness are related to equations for which initial data are specified on a hypersurface[13] or at a point.[14]

Differentiation of both sides of (2.11) with respect to q^B, and (2.11) itself, provide an expression for $\partial^2\theta^\alpha/\partial q^B \partial q^A$ in terms of q's and the derivatives of θ^β with respect to p^C's up to the second order. Then as a consequence of $\partial^2\theta^\alpha/\partial q^A \partial q^B = \partial^2\theta^\alpha/\partial q^B \partial q^A$ the functional relations between q's and those derivatives follow. We call them the *integrability conditions* of Eq. (2.11). An equation of the form (2.11) is said to be *completely integrable*, if its integrability conditions are satisfied identically. By the integrability conditions of Eq. (2.6) we mean the integrability conditions, in the sense just explained for an equation of the form (2.11), equivalent to (2.6).

Lemma 2: Equation (2.6) is completely integrable if and only if the space-time is left-conformally flat.

Proof: Indeed, to find integrability conditions of (2.6) it is convenient to use (2.6) itself. Let $\nabla_{\dot{A}}$ be defined by $\nabla_{\dot{A}} := k^A \nabla_{A\dot{A}}$. Then acting with that operator on both sides of (2.6), forming the comutator $(\nabla_{\dot{A}}\nabla_{\dot{B}} - \nabla_{\dot{B}}\nabla_{\dot{A}})k_C$, employing the Ricci identity,[12]

$$\tfrac{1}{2}\nabla_{(A}{}^{\dot{M}}\nabla_{B)\dot{M}}k^C = k^S\left[-C^C{}_{SAB} + \frac{R}{12}\epsilon_{S(A}\delta^C{}_{B)}\right], \tag{2.12}$$

and the Eq. (2.6), one obtains

$$-2C_{ABCD}k^Bk^Ck^D = (\nabla^{\dot{M}}\Theta_{\dot{M}})k_A. \tag{2.13}$$

The expression for $\nabla^{\dot{M}}\Theta_{\dot{M}}$ can be found from (2.4). The formula, which results, contains the second-order derivative of the form $\nabla_{\dot{D}}\nabla_Y k_C$ and $\nabla_{\dot{D}}\nabla_X k_C$. These can be expressed in terms of $\nabla_Y\nabla_{\dot{D}}k_C$ and $\nabla_X\nabla_{\dot{D}}k_C$, respectively, due to the Ricci identities (2.12) and

$$\tfrac{1}{2}\nabla^M{}_{(\dot{A}}\nabla_{|M|\dot{B})}k_C = k^S C_{SC\dot{A}\dot{B}}. \tag{2.14}$$

Finally the first-order derivatives of k_C can be eliminated by (2.6), $\nabla_X k_C$ and $\nabla_Y k_C$.

A substitution of $\nabla^{\dot{M}}\Theta_{\dot{M}}$ into (2.13) results in

$$\begin{aligned}&-2C_{ABCD}k^Bk^Ck^D\\ &\quad = (\rho^{\dot{M}}\sigma_{\dot{M}})^{-1}(\rho_{\dot{B}}X^{M\dot{B}} - \sigma_{\dot{B}}Y^{M\dot{B}})C_{MBCD}k^Bk^Ck^Dk_A.\end{aligned} \tag{2.15}$$

Now it is clear that Eq. (2.15) is an identity with respect to k_E, if and only if $C_{ABCD} = 0$. This, however, means that the space-time is left-conformally flat. ■

The properties of completely integrable equations of the form (2.11) are summarized in the next two Lemmas.

Lemma 3: For an arbitrary completely integrable equation of the form (2.11) and arbitrary initial data defined on an open neighborhood $\mathcal{U} \subset \mathcal{T}$ of the point p, there exists a formal solution of the corresponding Cauchy–Kovalevski problem in the form of unique formal power series at p. These are the power series which, if they converge, represent a unique holomorphic solution of the Cauchy–Kovalevski problem in some open neighborhood $U_p \subset M$ of the point p.

Proof: Indeed, we notice that the coefficients of the expansions of functions θ^α at the point p are determined by the initial data and Eq. (2.11) itself. This can be accomplished by the process of differentiations and substitutions[13]; for arbitrary initial data, the coefficients are defined uniquely by this procedure. To make sense in a power-series expansion, they must be symmetrical in the appropriate indices. For this, the complete integrability is sufficient. If an expansion coverges in some neighborhood of p, then its sum represents the unique solution of the corresponding Cauchy–Kovalevski problem because of the way in which it has been defined.[13] ■

Lemma 4: If the equation of the form (2.11) is completely integrable, then for arbitrary initial data defined on some open neighborhood $\mathcal{U} \subset \mathcal{T}$ of the point p, there exists a unique holomorphic solution of the corresponding Cauchy–Kovalevski problem defined on some open neighborhood U_p of that point; $U_p \subset M$.

Proof: It can be accomplished by the use of a technique of majorizing functions.[13]

Finally taking into account a relation between the Eqs. (1.1) and (2.6), and Lemma 1, 2, and 4 one infers

Theorem 1: Suppose that a complex space-time (M, ds^2) is left-conformally flat. Let D be a two-dimensional and integrable distribution on an open subset $U \subset M$. Let p by any point of U and $\mathcal{T}$ an integral submanifold of D passing through that point.

Consider an arbitrary spinor field $\tilde{k}^A$, defined along $\mathcal{T}$, $\tilde{k}^A \neq 0$ everywhere, such that the null two-dimensional subspaces of tangent spaces to M, determined by $\tilde{k}^A$ at points of $\mathcal{T}$, are transversal to $D|_{\mathcal{T}}$.[11] Then:

(i) there exists an open connected neighborhood $U_p \ni p$ in M, and a unique nowhere vanishing spinor field k^A on it, such that $k^A{}_{|\mathcal{T}\cap U_p} = \tilde{k}^A{}_{|\mathcal{T}\cap U_p}$ and k^A is a solution of (2.6);

(ii) k^A is a unique solution of (1.1) on U_p, which agrees with the initial data $\tilde{k}^A{}_{|\mathcal{T}\cap U_p}$ and consequently $\tilde{k}^A$ defines on U_p a unique congruence of self-dual null strings such that its null strings are tangent to null subspaces defined by the spinor field $\tilde{k}^A$ at points of $\mathcal{T}\cap U_p$.

Proof: Indeed, the first part of Theorem 1 is a result of complete integrability of (2.6), Lemma 2, and Lemma 4. Next because of Lemma 1, k^A is also a solution of (1.1). If D' is any integrable distribution on U_p with $\mathcal{T}$ being its integral submanifold, then the same k^A satisfies (2.6) with $'\Theta_{\dot{B}}$ related to D'; indeed, $'\Theta_{\dot{B}} = \Theta_{\dot{B}}$ (Lemma 1). Such k^A is a determined, however, uniquely, by the initial conditions $\tilde{k}^A$. Hence the assertion (ii) is true. ■

We remark that proportional initial data on $\mathcal{T}$ determines the same congruence. Indeed, it is a class of proportional spinor fields which defines the same congruence. Consider now a pair of proportional initial data $\tilde{k}^A$ and $\tilde{\varphi}\cdot\tilde{k}^A$ ($\tilde{\varphi}$ a function nowhere vanishing on $\mathcal{T}$). Let k^A and $'k^A$ denote the corresponding, unique solutions and let (Z^A) be holomorphic functions defining locally a congruence related to k^A. The mapping $(p^1, p^2) \rightarrow (Z^1(q^A(p), p^B), Z^2(q^A(p), p^B))$ is invertible in some neighborhood of the point $(p^B(p))$. Let its inverse be denoted by H. Then $'k^A = \varphi\cdot k^A$, where $\varphi(q^A, p^B) = (\tilde{\varphi}\circ H)(Z^C(q^A, p^B))$. Indeed, φk^A satisfies the condition $d(S^{AB}\varphi^2 k_A k_B) = 0$ [equivalent to (1.1)[3]], because k^A does ($S^{AB}k_A k_B = 2dZ^1 \wedge dZ^2$) and it agrees with the initial data $\tilde{\varphi}\cdot\tilde{k}^A$ for $'k^A$.

In the next section we make a more specific choice for $\mathscr{T}$ and a distribution D. And so $\mathscr{T}$ is a member of a congruence of self-dual null strings (or its open subset) and D is determined by that congruence.

3. A CONSTRUCTION OF CONGRUENCES BASED ON TWO TRANSVERSAL CONGRUENCES OF NULL STRINGS

We begin this section by recalling the following facts.

Proposition 1: Let M be a complex four-dimensional manifold and let Σ_1, Σ_2, and Σ be three congruences of two-dimensional submanifolds, such that any two of them are transversal, i.e., in some open neighborhood of any point $p\in M$ there are holomorphic functions q^A, $q^{A'}$, and Z^A such that $dq^1\wedge dq^2\wedge dq^{1'}\wedge dq^{2'}\neq 0$, $dq^1\wedge dq^2\wedge dZ^1\wedge dZ^2\neq 0$, and $dq^{1'}\wedge dq^{2'}\wedge dZ^1\wedge dZ^2\neq 0$ and locally $\Sigma_1: q^A = \text{const}$, $\Sigma_2: q^{A'} = \text{const}$, and $\Sigma: Z^A = \text{const}$; $A = 1,2$, $A' = 1',2'$.

Then M admits a unique conformal structure in which Σ_1, Σ_2, and Σ are congruences of null strings. It is represented by local metrics of the form

$$ds^2 = \Phi^{-1} Z_{A,B} Z^A_{\ ,C'}\, dq^B\, dq^{C'}, Z_A := \epsilon_{AB} Z^B \quad [\text{see } (3.4)] \tag{3.1}$$

with Φ being an arbitrary nowhere vanishing function.

Proof: Indeed, any local metric structure admitting Σ_1, Σ_2, and Σ as congruences of null strings is of the form

$$ds^2 = 2\Omega_{AB'}\, dq^A\, dq^{B'} = 2\omega_{AB}\, dq^A\, dZ^B.$$

Then necessarily $\omega_{(A|B|}Z^B_{\ ,C)} = 0$. This condition provides three algebraic equations on four components of ω_{AB}. In consequence of that, one obtains $\Omega_{AB'} = \Phi^{-1} Z_{C,A} Z^C_{\ ,B'}$, for some function $\Phi \neq 0$. It is clear that taking Φ arbitrary ($\Phi\neq 0$) we obtain a class of conformally related metrics, defined uniquely by the requirement that Σ_1, Σ_2, and Σ are congruences of null strings. It is also easy to check that the conformal structures defined on two overlapping coordinate neighborhoods are identical on their intersection. ∎

Proposition 2:

(i) A congruence of null strings transversal to congruence of self-dual null strings is self-dual itself.

(ii) Congruences of self-dual null strings which are nowhere transversal to each other are identical.

Proof: Indeed, (i) follows from the fact that the tangent spaces to a self-dual null string consist of the vectors of the form $k^A l^{\dot{B}}\partial_{A\dot{B}}$ for k^A fixed, while those for anti-self-dual ones are spanned by the vectors $l^A k^{\dot{B}}\partial_{A\dot{B}}$ with $k^{\dot{B}}$ fixed[1] ($\partial_{A\dot{B}}$ denotes a contravariant null tetrad). So the tangent spaces of self-dual and anti-self-dual null strings intersect along null directions. To obtain (ii) one observes that the tangent spaces to null strings of such congruences, passing through the same point are identical. ∎

Now suppose a conformal structure on M is given, which admits two self-dual congruences of null strings. One can represent it locally by metric fields of the form

$$ds^2 = 2\Omega_{AB'}\, dq^A\, dq^{B'}, \tag{3.2}$$

with $\Omega_{AB'}$ restricted conveniently by

$$\det(\Omega_{AB'}) = 1. \tag{3.3}$$

We introduce also the convention, which permits us to change the position of indices $A,B,\ldots$ and $A',B',\ldots$, according to the standard rules,[12] employing skew-symmetric Levi-Cività symbols:

$$(\epsilon_{AB}) = (\epsilon^{AB}) = (\epsilon_{A'B'}) = (\epsilon^{A'B'}) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{3.4}$$

($\Omega_A{}^{B'} = \Omega_{AC'}\epsilon^{C'B'}$, $\Omega^A{}_{B'} = \Omega_{CB'}\epsilon^{CA}$ etc.). With this convention (3.3) takes the form of $\Omega_{AB'}\Omega^{AB'} = 2$.

There is a natural choice for the null tetrad one forms $g^{A\dot{B}}$.[12] The self-dual two forms S^{AB} related to $g^{A\dot{B}}$ are $S^{11} = 2dq^{1'}\wedge dq^{2'}$, $S^{22} = 2dq^1\wedge dq^2$, and $S^{12} = \Omega_{AB'}dq^A\wedge dq^{B'}$.

By the standard process[12] one obtains the components of left-conformal curvature spinor C_{ABCD}:

$$\begin{aligned} C_{1111} &= 0 = C_{2222}, \\ C_{1112} &= -\tfrac{1}{4}\Omega_{BC',}{}^{CD'}\Omega^B{}_{D'}, \\ C_{1222} &= \tfrac{1}{4}\Omega_{AB',}{}^{AB}\Omega_B{}^{B'}, \\ C_{1212} &= \tfrac{1}{3}(-\Omega_{AB',}{}^{AB'} + \Omega_{AC',}{}^{A}\Omega_{BD',}{}^{D'}\Omega^{BC'}). \end{aligned} \tag{3.5}$$

It follows from Propositions 1 and 2 that to construct all local congruences of self-dual null strings transversal to the congruences $\Sigma_1: q^A = \text{const}$ and $\Sigma_2: q^{A'} = \text{const}$, it suffices to find local holomorphic functions Z^A and Φ (Φ nowhere vanishing), such that

$$Z_{A,B}Z^A{}_{,C'} = \Phi\Omega_{BC'}. \tag{3.6}$$

Let $F^{A'}{}_B$ denote the inverse of $Z^A{}_{,B'}$, i.e., $F^{A'}{}_B$ is such that

$$F^{A'}{}_B Z^B{}_{,C'} = \delta^{A'}{}_{C'}. \tag{3.7}$$

Then (3.6) can be rewritten as

$$Z_{A,B} = \Phi\Omega_{BB'}F^{B'}{}_A. \tag{3.8}$$

By differentiation of both sides of (3.8) with respect to q^C one finds

$$Z_{A,BC} = \Phi_{,C}\Omega_{BB'}F^{B'}{}_A + \Phi F^{B'}{}_A\Omega_{BB',C} + \Phi\Omega_{BB'}F^{B'}{}_{A,C}, \tag{3.9}$$

while a similar proccess applied to (3.7) plus (3.8) results in

$$\begin{aligned} F^{B'}{}_{A,C} = &-\tfrac{1}{2}\Phi_{,C'}\Omega_C{}^{B'}F^{C'}{}_A f \\ &-\tfrac{1}{2}\Phi\Omega_C{}^{B'}{}_{,C'}F^{C'}{}_A f - \Phi\Omega_{CN'}F^{N'L}{}_{,C'}F^{C'}{}_A F^{B'}{}_L, \end{aligned} \tag{3.10}$$

where $f := F_{A'B}F^{A'B}$.

Now the condition $Z^A{}_{,BC} = Z^A{}_{,CB}$ together with (3.9) and (3.10) provides an equation, which can be solved with respect to $\Phi_{,A}$, and

$$\Phi_{,A} = f\Phi\Phi_{,B'}\Omega_A{}^{B'} + \tfrac{1}{2}\Phi^2\Omega_A{}^{B'}f_{,B'} + \Phi\Omega^C{}_{B',C}\Omega_A{}^{B'}. \tag{3.11}$$

There are two important systems of differential equations to be discussed later. The first one referred to as I, consists of Eqs. (3.8) and (3.11), while the second, auxiliary, referred to as II consists of (3.8), (3.10), and (3.11). In I there are Z^A and Φ, which are unrelated, whereas $F^{A'}{}_B$ is determined by (3.7). In II, all of them Z^A, $F^{A'}{}_B$, and Φ are unrelated.

Next, one observes that the structure of both of them is of the form

$$\frac{\partial\theta^\alpha}{\partial q^A} = H^\alpha_A\left(q^B, q^{B'}, \theta^\beta, \frac{\partial\theta^\beta}{\partial q^{B'}}, \frac{\partial^2\theta^\beta}{\partial q^{B'}\partial q^{C'}}\right), \tag{3.12}$$

where $\alpha = 1,2,3$ and $(\theta^\alpha) = (Z^A, \Phi)$ for I, and $\alpha = 1,2,...,7$, $(\theta^\alpha) = (Z^A, F^{B'}{}_C, \Phi)$ for II.

There is a natural Cauchy–Kovalevski-like problem related to (3.12). The initial data are assumed on a submanifold given by $q^A = \text{const}$, $A = 1,2$. Also integrability conditions of (3.12) can be defined in a similar way as those for (2.11). However, a dependence of its right-hand sides on the second-order derivatives requires special attention. (We recall a well-known fact[13] that the equation $\partial u/\partial t = \partial^2 u/\partial x^2$ does not have local holomorphic solutions for arbitrary local holomorphic initial data on the hypersurface $t = 0$.)

Now we notice that

Lemma 5: The system I is completely integrable iff the metric structure $ds^2 = \Omega_{AB'}\, dq^A\, dq^{B'}$ is left-conformally flat.

Proof: Indeed, integrability conditions of (3.8) are satisfied identically being equivalent to (3.11). Then rather lengthy but straightforward calculation provides integrability conditions of (3.11), in the form of

$$4\Phi C_{1222} + 3f\Phi^2 C_{1122} + \Phi^3 f^2 C_{1112} = 0, \tag{3.13}$$

where C_{ABCD} are defined by (3.5). It is now obvious that (3.13) becomes an identity if and only if $C_{ABCD} = 0$. ■

For completely integrable equations of the form (3.12) there is also an assertion of Lemma 3 which is true. Indeed, it is just complete integrability, which makes it possible to construct unique formal power series expansions from arbitrary initial data. However, in general they fail to be convergent. A system I is exceptional in this respect. Indeed, it is easily seen that any formal solution of I determines a formal solution of II. Moreover, a convergence of one of them implies that of the second. Since a system II is of the form (2.11) it follows that the corresponding formal power series solutions converge. Thus we have

Theorem 1': Suppose the complex space-time (M, ds^2) is left-conformally flat. Let Σ_1 and Σ_2 be two transversal congruences of self-dual null strings defined on some open subset $U \subset M$: Σ_1:$q^A = \text{const}$, Σ_2:$q^{A'} = \text{const}$, and $dq^1 \wedge dq^2 \wedge dq^{1'} \wedge dq^{2'} \neq 0$.

Let p be any point of U and $\mathscr{T}$ a null string of the congruence Σ_1 passing through it; $\mathscr{T}$: $q^A = q^A(p)$.

Consider arbitrary holomorphic functions $\widetilde{Z}^A$ and $\widetilde{\Phi}$ defined on $\mathscr{T}$, such that $\det(\widetilde{Z}^A{}_{,B'})$ and $\widetilde{\Phi}$ nowhere vanish.

Then there exists an open neighborhood $U_p \ni p$ in M, $U_p \subset U$ and unique holomorphic functions Z^A and Φ on U_p (Φ nowhere vanishing) such that $Z^A{}_{|U_p \cap \mathscr{T}} = \widetilde{Z}^A{}_{|U_p \cap \mathscr{T}}$, $\Phi_{|U_p\cap\mathscr{T}} = \widetilde{\Phi}_{|U_p\cap\mathscr{T}}$ and (3.6) are satisfied.

The equations $Z^A = \text{const}$, $A = 1,2$, define on U_p a congruence of self-dual null strings transversal to Σ_1 and Σ_2 and any congruence with that property can be obtained in this way.

Remark 1: One easily finds, employing (3.3), (3.6), and (3.7), that $2dZ^1 \wedge dZ^2 = k_A k_B S^{AB}$ with

$$\pm k^A = (\tfrac{1}{2} Z^{M,N'} Z_{M,N'})^{1/2} \delta_2^A + \Phi(\tfrac{1}{2} Z^{M,N'} Z_{M,N'})^{-1/2} \delta_1^A. \tag{3.14}$$

δ_1^A and δ_2^A are components of the spinor fields defining congruences Σ_1 and Σ_2, respectively. The condition (3.13) is then equivalent to $C_{ABCD} k^A k^B k^C k^D = 0$.

We can restrict the initial data $\widetilde{\Phi}$ and $\widetilde{Z}^A$ conveniently, requiring, for example, $\widetilde{\Phi} = 1$. Then a special class of congruences can be obtained if an additional condition $(\widetilde{Z}^{A,B'} \widetilde{Z}_{A,B'}) = \text{const} \neq 0$ is imposed. If (M, ds^2) is a left-flat space-time and Σ_1 and Σ_2 are nonexpanding congruences of null strings related to covariantly constant spinor fields, then those new congruences defined by (3.14) are again of that type. Also $Z^{M,N'} Z_{M,N'} = \text{const}$, $\Phi = 1$, and k^A is a covariantly constant spinor field.[1,8]

Remark 2: "Projective extensions of heavens" and in particular "heavens" (left-flat spaces) can be characterized by the existence of special one-parameter families of congruences of self-dual null strings.[10]

For a complex space-time to be of that type it has to admit, at least locally, two linearly independent spinor fields k^A and k'^A such that

$$\begin{aligned} \nabla_{A\dot{B}} k_C &= \mu_{A\dot{B}} k_C + \nu_{C\dot{B}} k_A, \\ \nabla_{A\dot{B}} k'_C &= \mu'_{A\dot{B}} k'_C + \nu'_{C\dot{B}} k'_A, \end{aligned} \tag{3.15}$$

and

$$\mu_{A\dot{B}} + \nu_{A\dot{B}} = \mu'_{A\dot{B}} + \nu'_{A\dot{B}}. \tag{3.16}$$

[There is some obvious ambiguity in μ's and ν's, which does not effect the condition (3.16).]

Then any spinor field $l^A = ak^A + bk'^A$, where a and b are arbitrary constants, such that $(a,b) \neq (0,0)$, defines a congruence of null strings. In this way one obtains a generalized canonical family of congruences of null strings parametrized by ratios of a and b. We notice that in general l^A is not in a canonical normalization, even if the space-time is a left-flat one[10] [see also (1.1)].

Now let Φ and Φ' be such that the spinor fields $m^A = \Phi k^A$ and $m'^A = \Phi' k'^A$ are normalized canonically, i.e., they satisfy the equations of the form (1.1). Then employing (3.15) one obtains

$$\begin{aligned} 3Z_{A\dot{B}} &= \nabla_{A\dot{B}} \ln \Phi + \mu_{A\dot{B}} + \nu_{A\dot{B}}, \\ 3Z'_{A\dot{B}} &= \nabla_{A\dot{B}} \ln \Phi' + \mu'_{A\dot{B}} + \nu'_{A\dot{B}}, \end{aligned} \tag{3.17}$$

$$\begin{aligned} k^A(Z_{A\dot{B}} - \tfrac{1}{2}\nu_{A\dot{B}}) &= 0, \\ k'^A(Z'_{A\dot{B}} - \tfrac{1}{2}\nu'_{A\dot{B}}) &= 0. \end{aligned} \tag{3.18}$$

Next (3.16) together with (3.17) implies

$$dZ = dZ', \tag{3.19}$$

where $Z = -\tfrac{1}{2} Z_{A\dot{B}} g^{A\dot{B}}$, $Z' = -\tfrac{1}{2} Z'_{A\dot{B}} g^{A\dot{B}}$, and $\{g^{A\dot{B}}\}$ are null-tetrad 1-forms.[12]

Conversely, the condition (3.19) implies that $Z_{A\dot{B}}$ and $Z'_{A\dot{B}}$ are of the form

$$Z_{A\dot{B}} = \tfrac{1}{3}\nabla_{A\dot{B}} \ln \Phi + \tfrac{1}{3}\rho_{A\dot{B}}$$

and

$$Z'_{A\dot{B}} = \tfrac{1}{3}\nabla_{A\dot{B}} \ln \Phi' + \tfrac{1}{3}\rho_{A\dot{B}},$$

for some functions Φ, Φ' and a spinor field $\rho_{A\dot{B}}$.

If $\nu_{A\dot{B}}$ and $\nu'_{A\dot{B}}$ are chosen according to (3.18) with m^A, m'^A instead of k^A and k'^A (this can be always done), then the spinor fields $\Phi^{-1} m^A$ and $\Phi'^{-1} m'^A$ satisfy (3.15) with $\mu_{A\dot{B}}$

$=\rho_{A\dot{B}} - \nu_{A\dot{B}}$ and $\mu'_{A\dot{B}}, = \rho_{A\dot{B}} - \nu'_{A\dot{B}}$. The condition (3.16) is also fulfilled. Thus we have

Proposition 3: A complex space-time is a "projective extension of heavens" if and only if there exist two transversal congruences of self-dual null strings, such that Eq. (3.19) holds.

We notice also that the condition (3.19) is conformally invariant.[3]

Proposition 4: A complex space-time is conformally related to a "heaven" if and only if there exist two transversal congruences of self-dual null strings, such that $Z = Z'$ and $dZ = dZ' = 0$.

Proof: Indeed, if $dZ = dZ' = 0$ and $Z = Z'$, it follows that there exists a conformal gauge in which $Z = 0 = Z'$.[3] This means that the corresponding spinor fields are covariantly constant and consequently $C_{ABCD} = 0$. The converse assertion follows by a similar argument.■

[1]R. Penrose, Gen. Relativ. Gravit. **7**, 31 (1976).

[2](M,ds^2) is a pair consisting of a complex four-dimensional manifold and holomorphic metric structure, restricted by the condition that its left-conformal-curvature spinor $C_{ABCD} = 0$. (Left- is here equivalent to self-dual.) By the tangent space $T_p(M)$ is meant a complex four-dimensional vector space of complex tangent vectors of the type (1,0). All submanifolds are understood as complex submanifolds. A congruence of two-dimensional submanifolds on $U \subset M$ (U-open) is a set $\mathscr{C}$ of two-dimensional submanifolds of U, such that for any point $p \in U$:

(i) there is exactly one member of $\mathscr{C}$ passing through that point;

(ii) there exists a coordinate system $(U_p, \{z^1,z^2,z^3,z^4\})$, $U_p \subset U$, such that all members of $\mathscr{C}$ passing through the points of U_p are given on U_p by the equations of the form $z^1 = \text{const}$, $z^2 = \text{const}$.

[3]J. F. Plebański and K. Rózga, to appear.

[4]J. F. Plebański and I. Robinson, Phys. Rev. Lett. **37**, 493 (1976).

[5]R. S. Ward, Commun. Math. Phys. **78**, 1 (1980).

[6]C. R. Le Brun, Proc. R. Soc. London Ser. A **380**, 171 (1982).

[7]J. F. Plebański, J. Math. Phys. **16**, 2395 (1975).

[8]E. T. Newman, J. R. Porter, and K. P. Tod, Gen. Relativ. Gravit. **9**, 1129 (1978).

[9]J. D. Finley, III and J. F. Plebański, J. Math. Phys. **17**, 585 (1976).

[10]I. Robinson and K. Rózga, to appear.

[11]For most choices of $\mathscr{T}$ such a distribution can be found. In what follows we assume that this is the case.

[12]J. F. Plebański, *Spinors, Tetrads and Forms*, Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas, UNAM, Centro de Investigacion y Estudios Avanzados, IPN, Mexico City (1981).

[13]I. G. Petrovski, *Partial Differential Equations* (Saunders, Philadelphia, 1967).

[14]L. P. Eisenhart, *Non-Riemannian Geometry* (American Math. Soc., New York, 1968).

Approximate causal solutions for a class of wave equations with backscatter[a)]

James L. Anderson
Department of Physics, Stevens Institute of Technology, Hoboken, New Jersey 07030

(Received 3 December 1982; accepted for publication 7 October 1983)

The construction of solutions of wave equations with backscatter usually requires the use of some kind of approximation scheme. A problem then arises—how can one ensure that the approximate solution is causal? In this paper we present a method of constructing approximte causal wave-zone solutions that is applicable to a class of wave equation which reduce in form, in the limit as an expansion parameter goes to zero, to the scalar wave equation in flat space-time expressed in terms of the null coordinate $u_0 = t - r$. An extension of the method which is applicable to higher spin wave equations is also given.

PACS numbers: 04.20.Jb

I. INTRODUCTION

The construction of solutions of a system of wave equations requires either initial value data or some kind of radiation condition. If the equations are linear and the underlying space-time is flat and empty except for a region of compact spatial support containing the sources of the waves, there are several forms this condition might take. One can require that (i) the radiation is everywhere outgoing, or (ii) the solutions are causal, or (iii) there is no incoming radiation at past null infinity. Furthermore, it is a relatively easy matter to construct solutions that satisfy all of these conditions. One way of doing so is to use a retarded Green function which can be given in closed form for these equations. Such solutions automatically satisfy conditions (i) and (ii) and, if the sources do not grow too rapidly with time in the distant past, one can show[1] that condition (iii) is also satisfied.

The presence of backscattered radiation due to nonlinear terms in the wave equation which act as secondary emitters, the presence of a medium with variable index of refraction, a curved background space-time, or some combination of these possibilities complicates the construction of radiative solutions for a number of reasons. For one thing, one almost always has to resort to some kind of approximation scheme for this purpose. Furthermore, because of the presence of backscattered radiation, condition (i) above is not applicable. One should, of course, require that there be only outgoing radiation at future null infinity. However, this condition is not sufficient to exclude incoming radiation.[2]

Ehlers[3] and Leipold and Walker[1] have advocated the use of a Sommerfeld-type condition to be imposed at past null infinity for such problems. The imposition of such a condition is supposed to exclude the presence of incoming radiation. However, it is not clear what form such a condition should take in all cases. Even when one has such a condition, it is often difficult to employ in practice since solutions are usually given as functions of a retarded null coordinate while the approach to past null infinity is along curves of a constant advanced null coordinate. In general the relation between these two coordinates is not simple and usually depends on the expansion parameter.

[a)] Work supported in part by NSF Grant No. 79-11664-01.

In this paper we present a method of constructing approximate wave-zone solutions for a class of wave equations that satisfy condition (ii) above. A condition of causality is, of course, required on physical grounds. It is also, as we shall see, a condition that can be easily imposed on approximate solutions since its application involves the retarded rather than the advanced null coordinate. It should be noted that the imposition of a causality condition does not automatically ensure the absence of incoming radiation; one can easily give examples of causal solutions that violate a no-incoming radiation condition.[1] However, in these cases, the sources grow with time in the infinite past. Thus, a no-incoming radiation condition can be considered to be a condition on the behavior of sources associated with causal solutions and hence, if it exists, can be imposed to bound the motion of these sources after a causal solution has been found. That, however, is a problem which does not concern us here.

The type of problem to which our method of construction can be applied is one in which the source of the emitted radiation has compact spatial support. In what follows we shall refer to the region exterior to the source as the wave or outer zone. Our procedure leads to approximate causal solutions in this zone that do not require a knowledge of the solution in the region occupied by the source. As a consequence they will be seen to depend on one or more arbitrary functions of a retarded null coordinate. These functions must then be determined in some manner from a knowledge of the solution in the region occupied by the source. If, for example, the source is in slow motion, that is, if the ratio of the propagation time across the source to its period is small compared to unity, one can most easily use the method of matched asymptotic expansions (MAE)[4] for this purpose. We will not concern ourselves here, however, with the aspect of the problem.

We now make three additional assumptions concerning the nature of the problem with which we are dealing. We first assume that there exists in the outer zone a well-defined family of nonintersecting future-directed characteristic surfaces u = constant. Here u is a retarded null coordinate which is determined by the characteristic equation for the problem. The determination of u may be independent of the solution of the wave equation associated with the problem, or it may

 0022-2488/84/061947-04$02.50

depend on it as it does in the case of general relativity, where the metric used to determine u is not known initially. In this case, u must also be determined by some approximation scheme which is consistent with that used to determine the solution of the wave equation.

Our second assumption concerns the nature of the wave equation for the problem. We will assume that, in the limit in which a suitable expansion parameter ϵ goes to zero, this equation reduces in form, when expressed in terms of u and the spatial coordinates, to that of the scalar wave equation in flat, empty space-time when the latter is expressed in terms of the flat space-time characteristic coordinate $u_0 = r - t$. While $u = u_0$ in the limit in which our expansion parameter goes to zero, we emphasize that, in taking the above-mentioned limit, u is to be considered to be an independent variable. And finally we assume that, in the limit $u =$ const, $r \to \infty$, the effective source terms in the wave equation, i.e., the terms that go to zero in the limit $\epsilon \to 0$, are $O(1/r^3)$. This last assumption is necessary in order that the integrals which appear in the wave-zone expansion exist.

While the above assumptions may seem overly restrictive, they are in fact satisfied by a number of problems of physical interest, including the radiation of gravitational waves by gravitationally bound systems in general relativity and similar problems in Yang-Mills theory. We also expect that the method will prove to be applicable to other physically interesting problems involving electromagnetic radiation in a material medium and the radiation of acoustic waves.

II. A MODEL PROBLEM

In order to fix the ideas discussed above, and at the same time to illustrate the general approach used in our method, we will consider the following model problem for a scalar field satisfying a wave equation of the form

$$\{(1 - f(r))^2 \partial_t{}^2 - \nabla^2\}\psi = 4\pi\rho(\epsilon t, r), \tag{2.1}$$

where

$$\begin{aligned} f(r) &= \kappa, \quad r \leqslant 1 \\ &\kappa/r, \quad r \geqslant 1, \end{aligned} \tag{2.2}$$

and where

$$\rho(\epsilon t, r) = 0, \quad r \geqslant 1. \tag{2.3}$$

In what follows we will assume that the "coupling constant" κ is at least $O(\epsilon)$.

In the wave-zone zone one introduces stretched coordinates $\bar{t} = \epsilon t$, and $\bar{r} = \epsilon r$, and also a stretched coupling constant $\bar{\kappa} = \epsilon\kappa$. The outer equation, when expressed in terms of these quantities, takes the form

$$\{(1 - \kappa/r)^2 \partial_t{}^2 - \nabla^2\}\psi = 0, \tag{2.4}$$

where we have dropped the bars over all quantities.

The retarded null coordinate for this problem is determined by the characteristic equation

$$(1 - \kappa/r)^2(\partial_t u)^2 - (\partial_r)^2 = 0. \tag{2.5}$$

It follows that the retarded null coordinate is given by

$$u = t - t - \kappa \ln r. \tag{2.6}$$

Similarly one finds that the advanced null coordinate is given by

$$v = t + r + \kappa \ln r. \tag{2.7}$$

Let us now rewrite Eq. (2.4) in terms of the variables u and r and at the same time set

$$\psi = \phi / r. \tag{2.8}$$

The resulting equation is

$$\begin{aligned} &(1/r)(2\partial_u \partial_r - \partial^2{}_r)\phi \\ &\quad + \kappa\{(2/r^2)\partial_u \partial_r - (1/r^3)\partial_u \phi\} = 0. \end{aligned} \tag{2.9}$$

We see that our equation is of the type discussed in the introduction. The flat space-time D'Alembertian, when expressed in terms of u_0 operating on ϕ / r, has just the form of the first term in the above equation. Furthermore, if in the limit $u =$ const, $r \to \infty$, $\phi = O(1)$, then the remaining terms are all $O(1/r^3)$.

In order to construct an approximate solution to Eq. (2.9), we will expand in a usually asymptotic power series in of the form

$$\phi \sim \sum_{n=0}^{\infty} \kappa^n \phi_n. \tag{2.10}$$

When this expansion is substituted into Eq. (2.9) and the coefficients of the various powers of κ are equated to zero, one obtains a sequence of equations for the ϕ_n of the form

$$\partial_r(2\partial_u - \partial_r)\phi_n = S_n(u,r), \tag{2.11}$$

where $S_0 = 0$ and where S_n depends on $\phi_{n-1},...,\phi_0$ and is thus a known function of u and r. It is this system of equations that we must solve subject to a causality condition. In the next section we show how this task is accomplished.

III. CAUSAL APPROXIMATE SOLUTIONS

In the above example we considered a special problem with spherical symmetry; in this section we drop that assumption. However, we will assume that the field can be expanded in spherical harmonics as well as in an asymptotic series in the coupling constant κ. In this case the hierarchy of equations to be solved for the $\phi_{n,l,m}$ which appear in this double expansion have the general form

$$\{\partial_r(2\partial_u - \partial_r) + (1/r^2)l(l+1)\}\phi_{n,l,m} = S_{n,l,m}, \tag{3.1}$$

where again $S_{0,l,m} = 0$ and $S_{n,l,m}$ is a function of lower order ϕ's.

Consider first the equation with $n = 0$. The most general solution to this equation is a linear combination of an outgoing and an incoming solution of D'Alembert's equation of multipolarity l with u_0 and v_0 replaced by u and v, respectively. For $l = 0$

$$\phi_{0,00} = g(u) + h(v), \tag{3.2}$$

where g and h are arbitrary functions of their arguments. For $l = 1$

$$\begin{aligned} \phi_{0,1m} &= \{g_m(u) + (1/r)g_m(u)\} \\ &\quad + \{h_m(v) - (1/r)h_m(v)\}, \end{aligned} \tag{3.3}$$

where again g_m and h_m are arbitrary functions of their arguments and similarly for other values of l. The arbitrary functions appearing in these solutions are determined in the method of MAE by matching these solutions to corresponding solutions of the inner problem. In order that our overall solution be causal we shall require that, if the sources are stationary for times $t < t_0$, then there exists some u_0 which is related to t_0 such that the above solutions are also stationary for all $u < u_0$. The only way in which this can happen, of course, is if these solutions are independent of v. In this case they represent purely outgoing waves plus, perphaps, some everywhere static field.

Let us now consider the remaining equations in the hierarchy (3.1). In order that our overall solution will be causal we shall require that each $\phi_{n,lm}$ be stationary for $u < u_0$. It follows that the effective sources on the right-hand sides of these equations will also possess this property. They can therefore be written as the sum of two terms, one of which is everywhere stationary and another that vanishes for $u < u_0$. Thus, each $\phi_{n,lm}$ will consist of an everywhere stationary part which has as its source the stationary part of $S_{n,lm}$, and another part which has the part of $S_{n,lm}$ that vanishes for $u < u_0$ as its source. Since no causality condition is involved in solving the stationary parts of the hierarchy equations, we will not concern ourselves with them here, but consider only the problem of solving the nonstationary parts of these equations.

We thus require solutions of equations of the form (3.1) where the effective sources vanish for $u < u_0$ which also vanish for $u < u_0$. The problem of finding such solutions is complicated by the fact that the effective sources are only known as functions of u and r in the outer zone so that the usual retarded Green function, whose use requires a knowledge of the sources everywhere, cannot be employed. In spite of this difficulty, we have found a particular causal solution to equations of the form (3.1). It is

$$\phi_{n,lm} = P_l \int_{-\infty}^{u} k_{n,lm}(u', y)\, du', \tag{3.4}$$

where $y = (u - u')/2 + r$ and where the operator P_l is given by

$$P_l = (\partial_r - 2/r)(\partial_r - 4/r)...(\partial_r - 2l/r). \tag{3.5}$$

The integrand appearing in Eq. (3.4) in turn satisfies

$$2\partial_r P_l k_{n,lm} = S_{n,lm}. \tag{3.6}$$

This last equation can be integrated directly to give

$$k_{n,lm} = r^{2l} \int_r^\infty dr_1\, r_1^{-2} \int_{r_1}^\infty dr_2\, r_2^{-2} \times \cdots \times \int_{r_{l-1}}^\infty dr_l r_l^{-2} H_{n,lm}(u,r_l), \tag{3.7}$$

where

$$H_{n,lm}(u,r) = -(1/2)\int_r^\infty dr'\, S_{n,lm}(u,r'). \tag{3.8}$$

The proof that Eqs. (3.4)–(3.8) constitute a particular solution to Eq. (3.1) is given in the Appendix. Here we note that, as follows from Eq. (3.7), we only need a knowledge of the source function $S_{n,lm}$ in the outer region to construct $k_{n,lm}$, since the integrals appearing in Eq. (3.7) do not extend into the inner zone. Furthermore, we see that if $S_{n,lm}$ is zero for $u < u_0$ then $k_{n,lm}$ will also have this property. It follows therefore that $\phi_{n,lm}$ will also have this property. Since it is a particular solution of Eq. (3.1), we see that we can add to it any solution of the homogeneous equation that also vanishes for $u < u_0$ to obtain the general causal solution of this equation. This latter function will then be determined by matching to the inner solution. Finally, we should mention that Bardeen and Press[5] have constructed approximate solutions for wave equations in a Schwarzschild space-time that are causal in the sense employed here. We have applied our method to their equation for a scalar field for the cases $l = 0$ and 1 and have shown that the two sets of solutions are equal to each other.

IV. CAUSAL SOLUTIONS FOR HIGHER-ORDER SPIN FIELDS

In their paper referred to above, Bardeen and Press also considered wave equations for higher-order spin fields in a Schwarzschild space-time. These equations have the form

$$\{2\partial_u\partial_r + (2/r)(s + p + 1)\partial_u - \partial^2{}_r - (2/r)(s + 1)\partial_r + (1/r^2)(l + s + 1)(l - s)\}\psi_l = S_l(u,r), \tag{4.1}$$

where S_l is a known function of u and r which vanishes for $u < u_0$ and is $O(r^{-3})$. In these equations s is the spin of the field and can take on values $0,1,\ldots$. The parameter p is the spin weight of the particular component of the field being considered and has the values $\pm s$. If we make the substitution $\psi_l = \phi_l/r^m$, where $m = s - p + 1$, Eq. (4.1) takes the form

$$(1/r^m)\{\partial_r(2\partial_u - \partial_r) + (2p/r)(2\partial_u - \partial_r) + (1/r^2)(l + p)(l - p + 1)\}\phi_l = S_l(u,r). \tag{4.2}$$

A causal solution to this equation is again of the form

$$\phi_l = P_l \int_{-\infty}^{u} k_l(u', y) du', \tag{4.3}$$

where now

$$P_l = (\partial_r - 2/r)(\partial_r - 4/r)...(\partial_r - 2(p + l)/r) \tag{4.4}$$

with $l \geqslant -p$ if p is negative or $l \geqslant 0$ if p is positive. The function k_l now satisfies

$$2\{\partial_r + (2p/r)\}P_l k_l(u,r) = S_l(u,r) \tag{4.5}$$

which can again be integrated to yield an expression for k_l similar to that given by Eq. (3.7).

V. DISCUSSION

In this paper we have shown how to construct an approximate causal solution to a class of wave equations with backscatter. We want to emphasize that although we have employed the notion of a source that is stationary for times less than some t_0, there is nothing that precludes the possibility that t_0 lies in the infinite past. The solution given above is applicable in this case as well, provided only that the integral in Eq. (3.4) exists; this will be the case if the source does not grow too rapidly in the infinite past. We also want to emphasize that, even if one only requires the first term in the approximation, it is important to know that it is the first term in a consistent approximation procedure.

ACKNOWLEDGMENTS

I would like to thank Martin Walker and Jurgen Ehlers for several helpful discussions.

APPENDIX: PROOF OF SOLUTION

To prove that $\phi_{n,lm}$ given by Eq. (3.4) is a solution of Eq. (3.1) we will first prove by induction that

$$\partial_r(2\partial_u - \partial_r)P_l K = \partial_r P_l(2\partial_u - \partial_r)K - l(l+1)(P_l K)/r^2 \tag{A1}$$

for all l where P_l is given by Eq. (3.5) and where K is any reasonably well-behaved function of u and r. Because ∂_u commutes with P_l, this equation reduces to

$$\partial^2{}_r P_l K = \partial_r P_l \partial_r K + l(l+1)(P_l K)/r^2. \tag{A2}$$

Equation (A2) is obviously satisfied when $l = 0$. Let us assume now that it is satisfied for some l and make the substitution

$$K \rightarrow (\partial_r - 2(l+1)/r)K \tag{A3}$$

in Eq. (A2). The result of this substitution is

$$\partial_r{}^2 P_{l+1} K = \partial_r P_{l+1} \partial_r K + (l+1)(l+2)(P_{l+1}K)/r^2 + \partial_r P_l(K/r^2) - (1/r^2)P_l(\partial_r - 2(l+1)/r)K. \tag{A4}$$

Thus we see that Eq. (A1) will hold for $l = l + 1$ if it holds for l provided that

$$\partial_r P_l(K/r^2) = (1/r^2)P_l(\partial_r - 2(l+1)/r)K \tag{A5}$$

for all l and any K. It is any easy matter to show that this equation holds for $l = l + 1$ if it holds for l by now making the substitution

$$K \rightarrow (\partial_r - 2(l+2)/r)K \tag{A6}$$

into it. Since it obviously holds when $l = 0$, we can again conclude by induction that it holds for all l. We therefore conclude that Eq. (A1) holds for all l.

With the use of Eq. (A1) we can now rewrite equation (3.1) after dropping unnecessary subscripts in the form

$$\partial_r P_l(2\partial_u - \partial_r)K_l = S_l. \tag{A7}$$

If we then take

$$K_l = \int_{-\infty}^{u} k_l(u', y)du', \tag{A8}$$

we see that k_l satisfies

$$2\partial_r P_l k_l = S_l, \tag{A9}$$

which is just Eq. (3.6), thus proving that $\phi_{n,lm}$ given by Eq. (3.4) is indeed a solution of Eq. (3.1). A similar proof can be given for the case of the higher-spin equations.

[1]G. Leipold and M. Walker, Ann. Inst. H. Poincaré, 27 (1977).
[2]M. Walker (private communication).
[3]J. Ehlers, in *Isolated Gravitating Systems in General Relativity, Proceedings of the International School of Physics Enrico Fermi*, Course XLVII, edited by J. Ehlers (Academic, New York, 1979).
[4]See, for example, A. Nayfeh, *Pertrubation Methods* (Wiley, New York, 1973).
[5]J. M. Bardeen and W. M. Press, J. Math. Phys. **14**, 7 (1973).

Electrovac type D solutions with cosmological constant

Alberto García Díaz[a)]
Centro de Investigación y de Estudios Avanzados del IPN, Departamento de Física, Apartado Postal 14-740, 07000 México D.F., Mexico

(Received 1 February 1983; accepted for publication 18 November 1983)

All type D electrovac solutions with cosmological constant for an algebraically general electromagnetic field aligned along the Debever–Penrose directions are obtained. From a single canonical metric element both the null and the non-null orbit solutions are derived simultaneously. The explicit expressions of the obtained solutions are listed in the table.

PACS numbers: 04.20.Jb

1. INTRODUCTION

In this work we determine, in a systematic and unified manner, the complete set of solutions for the class $\mathscr{D}$ of electrovac metrics with λ, i.e., solutions of Petrov type D to the Einstein–Maxwell equations with cosmological constant for an algebraically general electromagnetic field aligned along the double Debever–Penrose (DP) vectors. The $\mathscr{D}$ solutions possess always a two-parameter Abelian isometry group with orbits of two kinds—null and non-null orbits.

The general structures for the non-null and null orbit $\mathscr{D}$ metrics have been established, without explicit integration, by Debever and McLenaghan.[1]

In our earlier paper,[2] in what follows referred to as I, we have established all solutions of the null orbit $\mathscr{D}$ class, and also we have shown, in Sec. 5 of I, the canonical form of the metric for the non-null orbit family of $\mathscr{D}$ solutions together with the complete set of field equations to be satisfied by the structural functions.

Here, starting from a single canonical metric element, we determine, by integrating the field equations pointed out in Sec. 5 of I, all the non-null orbit $\mathscr{D}$ solutions and the null orbit $\mathscr{D}$ class as well. The obtained solutions contain, as it should be, all known type D metrics; for instance, the seven-parameter Plebański–Demiański solution,[3] by canceling out the electric and magnetic charges, one arrives at the vacuum type D metrics with λ, by setting λ to zero, one obtains the Kinnersley electrovac solutions,[4] and so forth.

2. GENERAL METRICS. EINSTEIN–MAXWELL EQUATIONS AND THEIR SOLUTIONS

The study of the $\mathscr{D}$ class of solutions, possessing an isometry group with null or non-null orbits, we begin from the canonical metric

$$g = H^{-2}\left\{ \frac{\Delta}{P} dx^2 + \frac{P}{\Delta}(d\tau + p\, d\sigma)^2 + 2\, dy(d\tau + m\, d\sigma) - \frac{Q}{\Delta}(d\tau + m\, d\sigma)^2 \right\}, \tag{2.1}$$

and the electromagnetic two-form

$$\omega = H^{-2}(\mathscr{E} + i\check{\mathscr{B}})\{(d\tau + m\, d\sigma)\wedge dy + i(d\tau + p\, d\sigma)\wedge dx\}, \tag{2.2}$$

where the structural functions $m(x), p(y), \Delta := m - p, P(x), Q(y)$, and $H(x,y)$ satisfy the canonical set of equations

$$\Delta(m_{xx} - p_{yy}) - (m_x)^2 - (p_y)^2 = 0, \tag{2.3a}$$

$$\begin{aligned} &4\Delta H_{xx} = \mathscr{H},\ 4\Delta H_{yy} = -\mathscr{H}, \\ &2m_x H_x + 2p_y H_y = \mathscr{H}, \\ &2\Delta H_{xy} - m_x H_y + p_y H_x = 0,\ \mathscr{H} := (m_{xx} + p_{yy})H, \end{aligned} \tag{2.3b}$$

$$\begin{aligned} &-4\lambda\Delta = D_x P + D_y Q, \\ &D_z F := H^2 F_{zz} - 6HH_z F_z - 4F(HH_{zz} - 3(H_z)^2), \end{aligned} \tag{2.3c}$$

and the electromagnetic functions $\mathscr{E}(x,y)$ and $\check{\mathscr{B}}(x,y)$ fulfill the Maxwell equations

$$d\ln(\mathscr{E} + i\check{\mathscr{B}})H^{-2}\Delta - i\left(\frac{m_x}{\Delta}dy + \frac{p_y}{\Delta}dx\right) = 0, \tag{2.4}$$

which states that ω from (2.2) is a closed two-form. The integrability condition of this last equation is just Eq. (2.3a).

If the function Q is equated to zero, one obtains the null-orbit $\mathscr{D}$ metric studied in I.

For the non-null orbit $\mathscr{D}$ class is more convenient, in order to facilitate the identification of the solutions studied with previously known solutions, the symmetric form of the metric (2.1), namely

$$\begin{aligned} g &= H^{-2}\left\{ \frac{\Delta}{P} dx^2 + \frac{P}{\Delta}(du + p\, dv)^2 + \frac{\Delta}{Q} dy^2 - \frac{Q}{\Delta}(du + m\, dv)^2 \right\}, \\ \omega &= H^{-2}(\mathscr{E} + i\check{\mathscr{B}})\{(du + m\, dv)\wedge dy + i(du + p\, dv)\wedge dx\}, \end{aligned} \tag{2.5}$$

which is obtainable from (2.1) by executing the transformation

$$d\tau = du - \frac{p}{Q}dy,\ d\sigma = dv + \frac{dy}{Q}. \tag{2.6}$$

Without any loss of generality, one can consider optionally the metric structure (2.5), with the structural functions satisfying the canonical set of Eqs. (2.3) and (2.4), as the basic canonical metric structure for the $\mathscr{D}$ class of solutions possessing a group with null or non-null orbits. From this point of view, the null-orbit solutions become limiting contractions of the solutions for the metric (2.5) when Q is setting to zero, having executed previously the transformation (2.6).

[a)] Also at Sección de Graduados, Escuela Superior de Ingeniería Mecáncia y Eléctrica del IPN, México, D.F.

 0022-2488/84/061951-04$02.50

TABLE I. Type $\mathscr{D}$ electrovac solutions with λ.

$$g = H^{-2}\left\{\frac{\Delta}{P}\,dx^2 + \frac{P}{\Delta}(d\tau + p\,d\sigma)^2 + 2dy(d\tau + m\,d\sigma) - \frac{Q}{\Delta}(d\tau + m\,d\sigma)^2\right\},\ \Delta = m - p,$$
$$\omega = H^{-2}(\mathscr{E} + i\check{\mathscr{B}})\{(d\tau + m\,d\sigma)\wedge dy + i(d\tau + p\,d\sigma)\wedge dx\}.$$
$$T:\left\{x = x,\ y = y,\ d\tau = du - \frac{p}{Q}\,dy,\ d\sigma = dv + \frac{dy}{Q}\right\},$$
$$g = H^{-2}\left\{\frac{\Delta}{p}\,dx^2 + \frac{P}{\Delta}(du + p\,dv)^2 + \frac{\Delta}{Q}\,dy^2 - \frac{Q}{\Delta}(du + m\,dv)^2\right\}.$$

N	Static Metrics	Z
$B - R$	$S(\mu z) := 1 - \mu z^2,\ m = 1 = H,\ p = 0,\ P = S[(\lambda + e^2 + g^2)x],\ Q = S[(\lambda - e^2 - g^2)y],$ $\omega = d\{(e + ig)(y\,d\tau + ix\,d\sigma)\} =: \omega(1).$ $Q = 0,\ P = S(2\lambda x),\ \lambda = e^2 + g^2.$	0
$gR - N$	$\epsilon := \{-1, 0, 1\},\ m = 1,\ p = 0,\ H = y,\ \omega = \omega(1),$ $P = S(\epsilon x),\ Q = -\frac{\lambda}{3} + \epsilon y^2 + 2my^3 + (\epsilon^2 + y^2)\,y^4.$	Re Z
$g^*R - N$	$m = 1,\ p = 0,\ H = x,\ \omega = \omega(1),\ Q = S(\epsilon y),$ $p = -\frac{\lambda}{3} + \epsilon x^2 + 2nx^3 - (e^2 + g^2)x^4.$ $Q = 0,\ P = -\frac{\lambda}{3} + 2nx^3 - (e^2 + g^2)x^4.$	0
gC	$m = 1,\ p = 1,\ H = x + y,\ \omega = \omega(1),$ $p = \left(-\frac{\lambda}{6} + \gamma\right) - \epsilon x^2 + 2mx^3 - (e^2 + g^2)x^4,$ $Q = \left(-\frac{\lambda}{6} - \gamma\right) + \epsilon y^2 + 2my^3 + (e^2 + g^2)y^4.$	Re Z
	Stationary Metrics	
$g\tilde{B}(+)$	$m = e^{2x},\ p = 0,\ H = e^x\cos y,$ $T:\{x \to -\frac{1}{2}\ln x,\ y \to \arctan y/l,\ P \to (2lx)^{-2}P,\ Q \to (l^2 + y^2)^{-2}Q\},$ $g = \delta\left(P\,d\sigma^2 + \frac{dx^2}{P}\right) + \frac{\delta}{Q}\,dy^2 - \frac{Q}{\delta}(d\tau - 2lx\,d\sigma)^2,\ \delta = l^2 + y^2,$ $P = S(\epsilon x),\ Q = e^2 + g^2 - \epsilon l^2 + l\lambda^4 - 2my + (\epsilon - 2\lambda l^2)\,y^2 - \frac{\lambda}{3}\,y^4,$ $\omega = -d\left\{(e + ig)\left(\frac{d\tau}{y + il} - i\frac{y - il}{y + il}x\,d\sigma\right)\right\}.$	Z
$g\tilde{B}(-)$	$m = 0,\ p = e^{2y},\ H = e^y\cos x,$ $T:\{x \to \arctan x/l,\ y \to -\frac{1}{2}\ln y,\ P \to (l^2 + x^2)^{-2}P,\ Q \to (2ly)^{-2}Q\},$ $g = \frac{P}{\delta}(d\tau + 2ly\,d\sigma)^2 + \frac{\delta}{p}\,dx^2 + \delta(2\,dy\,d\sigma - Q\,d\sigma^2),\ \delta = l^2 + x^2,$ $P = -e^2 - g^2 - \epsilon l^2 + \lambda l^4 + 2nx + (\epsilon - 2\lambda l^2)x^2 - \frac{\lambda}{3}x^4,\ Q = S(\epsilon x),$ $\omega = d\left\{(e + ig)\left(\frac{d\tau}{l + ix} + \frac{l - ix}{l + ix}y\,d\sigma\right)\right\}.$ $Q = 0,\ P = -e^2 - g^2 + 2nx + \lambda\left(l^4 - 2l^2x^2 - \frac{x^4}{3}\right).$	0
$P - C[A]$	$m = x^2,\ p = -y^2,\ H = 1,\ \omega = -\left\{\frac{e + ig}{y + ix}(d\tau + ixy\,d\sigma)\right\} =: \omega(2),$ $P = b - g^2 + 2nx - \epsilon x^2 - \frac{\lambda}{3}x^4,\ Q = b + e^2 - 2my + \epsilon y^2 - \frac{\lambda}{3}\,y^4.$ $Q = 0,\ P = -(e^2 + g^2) + 2nx.$	Z
$P - D$	$m = x^2,\ p = -y^2,\ H = 1 - xy,\ \omega = \omega(2),$ $P(x) = \left(-\frac{\lambda}{6} - g^2 + \gamma\right) + 2nx - \epsilon x^2 + 2mx^3 + \left(-\frac{\lambda}{6} - e^2 - \gamma\right)x^4,$ $Q(x) = \left(-\frac{\lambda}{6} + e^2 + \gamma\right) + 2my - \epsilon y^2 + 2ny^3 + \left(-\frac{\lambda}{6} + g^2 - \gamma\right)y^4.$	Z

To integrate the field equations [(2.3) and (2.4)] one has to proceed as follows: First, solve Eq. (2.3a) for m and p. Second, integrate Eqs. (2.3b) for H. Finally, integrate Eq. (2.3c) for P and Q. Equations (2.3a) and (2.3c) yield always to variable separable equations. The Maxwell equations (2.4) are trivially integrable by virtue of the (2.3a) equation. Basically, all solutions of Eq. (2.3a) and (2.3b) for m, p, and H have been obtained in I, therefore, in the present generalization with Q different from zero, it remains only to integrate the scalar curvature equation (2.3c) for P and Q.

In the present generalization, the following theorem for the congruences of electromagnetic eigenvectors, say l and n, applies: Within the class $\mathscr{D}$ of metrics, when m_x and p_y do not vanish simultaneously, if the congruence $l\,(n)$ is nondiverging, then it is also twist-free.

Proof: For the $\mathscr{D}$ metric in the (2.5) form, the complex rotation Z of $l\,(n)$ is given by

$$Z(l) = Z(n) = \left(\frac{Q}{2\Delta}\right)^{1/2}\left\{H_y + \frac{1}{2}\frac{p_y}{\Delta}H + \frac{i}{2}\frac{m_x}{\Delta}H\right\} =: \left(\frac{Q}{2\Delta}\right)^{1/2}(\rho + i\omega). \tag{2.7}$$

[For the $\mathscr{D}$ metric given by (2.1) the complex rotations are $Z(l) = \rho + i\omega$, $Z(n) = (Q/2\Delta)(\rho + i\omega)$]. Let the divergence ρ to be zero, $\rho = 0$, differentiating with respect to y this equation for ρ, and substituting H_{yy} from (2.3) in the so obtained equation, one arrives at $H(m_x)^2 = 0$, thus the twist ω is zero. Inversely, suppose that the twist ω is zero, $\omega = 0 \Rightarrow m_x = 0$, hence, from (2.3a), $p_{yy} = \Delta^{-1}(p_y)^2$. Substituting this p_{yy} into the equation with first derivatives of H, one arrives at $p_y\rho = 0$. Therefore, provided that $m_x = 0 \neq p_y$, ρ is equal to zero.

If $m_x = 0 = p_y$, there exist twist-free $\mathscr{D}$ solutions endowed with divergence $\rho = H_y$.

This theorem allows us to subclassify the solutions according with the (vanishing or different from zero) values of the derivatives m_x and p_y. The determination of P and Q from Eq. (2.3c), as was pointed out above, is straightforward once the functions m, p, and H are known. Omitting this evaluation, we list all type D (electrovac and vacuum) solutions with cosmological constant (if present) in the table below.

In the table we give the solutions in canonical coordinates, omitting spurious parameters. When Q is equal to zero, one arrives at the null orbit solutions; in the case $Q \neq 0$, one is dealing with the non-null orbit solutions. The symbols used are N (name symbol); BR (Bertotti–Robinson), RN (Reissner–Nordstrem), C (metric of Levi-Cività), $[A]$, $\widetilde{B}(+)$, $\widetilde{B}(-)$ (Carter solutions),[5] P (Plebański),[6] and PD (Plebański–Demiański).[7] The symbols g and * stand for "generalized" and "anti." T:{ } denote transformations; we omit linear transformations of the Killingian variables $\tau(u)$ and $\sigma(v)$. Z denotes complex expansion.

Specializing the values of the parameters one arrives at more special subclasses of solutions. For complete references, see Ref. 8.

For the sake of completeness, the explicit expressions of the secondary curvature quantities are given in the Appendix.

3. CONCLUSIONS

The $\mathscr{D}$ metrics (null and non-null orbit) are exhausted by the solutions presented in the table. Specializing the values of the parameters one arrives at more particular solutions; for instance, the vacuum and charged Kinnersley solutions, the $\mathscr{D}$'s with λ, the Kerr–Newman solution, the generalized Schwarzschild metric,... . The seven-parameter Plebański–Demiański solution is the most general of the $\mathscr{D}$ metrics.

Note added in proof: After this paper was concluded, we learned of the work of R. Debever, N. Kamran, and R. G. McLenaghan on closely related subjects.[9]

APPENDIX

The curvature quantities can be easily determined by using the null tetrad formalism. The metric g and the two-form ω for the $\mathscr{D}$ class of solutions can be given as

$$\begin{aligned} &g = 2e^1 \otimes e^2 + 2e^3 \otimes e^4, \\ &\omega = (\mathscr{E} + i\check{\mathscr{B}})(e^1 \wedge e^2 + e^3 \wedge e^4), \\ &e^1 = \bar{e}^2,\ e^3 = \bar{e}^3,\ e^4 = \bar{e}^4. \end{aligned} \tag{A1}$$

Thus, the e^3 and e^4 null vectors are aligned along the electromagnetic eigenvectors l and n, and consequently, along the DP directions. Hence, the only nonvanishing component of the traceless Ricci tensor C_{ab} is $C_{12} = -\mathscr{E}^2 - \check{\mathscr{B}}^2$, and the only different from zero complex curvature quantity $C^{(a)}$ $(a = 1,...,5)$, which characterize the Weyl tensor, is $C^{(3)}$.

For the expression (2.1) the choice of ϵ^a is

$$\begin{aligned} &\left.\begin{matrix}\epsilon^1\\ \epsilon^2\end{matrix}\right\} = H^{-1}\left(\frac{P}{2\Delta}\right)^{1/2}\left\{\frac{\Delta}{P}dx \pm i(d\tau + p\,d\sigma)\right\}, \\ &\epsilon^3 = H^{-1}(d\tau + m\,d\sigma), \\ &\epsilon^4 = H^{-1}[dy - (Q/2\Delta)(d\tau + m\,d\sigma)], \end{aligned} \tag{A2}$$

while for the symmetric metric (2.5), the e^a are

$$\begin{aligned} &\left.\begin{matrix}e^1\\ e^2\end{matrix}\right\} = H^{-1}\left(\frac{P}{2\Delta}\right)^{1/2}\left[\frac{\Delta}{P}dx \pm i(du + m\,dv)\right], \\ &\left.\begin{matrix}e^3\\ e^4\end{matrix}\right\} = H^{-1}\left(\frac{Q}{2\Delta}\right)^{1/2}\left[\frac{\Delta}{Q}dy \pm (du + m\,dv)\right]. \end{aligned} \tag{A3}$$

The tetrad e^a from (A3) is related with ϵ^a from (A2) by a σ-gauge, with $\sigma = \ln(Q/2\Delta)^{1/2}$, accompanied by the coordinate transformation (2.6). Under this gauge-coordinate transformation one has

$$\begin{aligned} &e^1 \to \epsilon^1,\ e^2 \to \epsilon^2,\ (Q/2\Delta)^{-1/2}\ e^3 \to \epsilon^3,\ (Q/2\Delta)^{1/2}\ e^4 \to \epsilon^4, \\ &C^{(3)}(e) \to C^{(3)}(\epsilon),\ C_{12}(e) \to C_{12}(\epsilon). \end{aligned} \tag{A4}$$

With respect to these e^a or ϵ^a tetrads the curvature quantities $C^{(3)}$ and C_{12} are given by

$$\begin{aligned} 6\Delta H^{-2}C^{(3)} &= P_{xx} - 3\frac{m_x}{\Delta}P_x - 2P\left[\frac{m_{xx}}{\Delta} - 2\left(\frac{m_x}{\Delta}\right)^2 + \left(\frac{p_y}{\Delta}\right)^2\right] \\ &\quad + Q_{yy} + 3\frac{p_y}{\Delta}Q_y \\ &\quad + 2Q\left[\frac{p_{yy}}{\Delta} + 2\left(\frac{p_y}{\Delta}\right)^2 - \left(\frac{m_x}{\Delta}\right)^2\right], \\ 4\Delta H^{-2}C_{12} &= -4\Delta H^{-2}(\mathscr{E}^2 + \check{\mathscr{B}}^2) \\ &= P_{xx} - 2\left(\frac{H_x}{H} + \frac{m_x}{\Delta}\right)P_x \\ &\quad - \frac{2}{\Delta}\left(p_{yy} - 2m_x\frac{H_x}{H}\right)P \\ &\quad - Q_{yy} + 2\left(\frac{H_y}{H} - \frac{p_y}{\Delta}\right)Q_y \\ &\quad - \frac{2}{\Delta}\left(m_{xx} - 2p_y\frac{H_y}{H}\right)Q. \end{aligned} \tag{A5}$$

[1]R. Debever and R. G. McLenaghan, J. Math. Phys. **8**, 1711 (1981).
[2]A. García, D. and H. Salazar I, "All null orbit type *D* electrovac solutions with cosmological constant," J. Math. Phys. **24**, 2498 (1983).
[3]J. F. Plebański and M. Demiański, Ann. Phys. (N.Y.) **98**, 98 (1976).
[4]W. Kinnersley, Ph.D. thesis, California Institute of Technology (1969).
[5]B. Carter, Commun. Math. Phys. **10**, 280 (1968).
[6]J. F. Plebański, J. Math Phys. **20**, 1946 (1979).
[7]J. F. Plebański, Ann. Phys. (N.Y.) **90**, 196 (1975).
[8]D. Kramer, H. Stephani, M. MacCallum, and E. Herlt, *Exact Solutions of Einstein's Field Equations* (Deutscher Verlag, Berlin, 1980).
[9]R. Debever, N. Kamran, and R. G. McLenaghan, J. Math. Phys. **25**, 1955 (1984).

Exhaustive integration and a single expression for the general solution of the type D vacuum and electrovac field equations with cosmological constant for a nonsingular aligned Maxwell field[a)]

R. Debever
Département de Mathématique, Université Libre de Bruxelles, Brussels, Belgium

N. Kamran and R. G. McLenaghan
Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada

(Received 1 June 1983; accepted for publication 18 November 1983)

We present an exhaustive integration of the type D vacuum and electrovac field equations with cosmological constant admitting a nonsingular aligned Maxwell field satisfying the generalized Goldberg–Sachs theorem. We derive a single expression for the general solution from which one may obtain all particular cases known until now in partial versions. We also investigate in detail the separability properties of the Hamilton–Jacobi equation for the charged particle orbits and of the Klein–Gordon equation for a massive charged spin-zero test particle and their corresponding Killing tensors.

PACS numbers: 04.20.Jb, 02.30.Jr

1. INTRODUCTION

A major contribution to the theory of type D vacuum and electrovac solutions with cosmological constant was made by Carter[1–3] who, while searching for the solutions admitting a two-parameter abelian invertible isometry group and separable Hamilton–Jacobi and Klein–Gordon equations, determined essentially all the nonaccelerating type D solutions[4] in a coordinate system in which the components of the metric and Maxwell field are rational functions. Another important contribution was made about the same time by Kinnersley[5] who, by choosing a null tetrad and coordinates specialized along one of the pair of repeated null directions of the Weyl tensor whose existence is guaranteed by the Goldberg–Sachs[6] theorem, obtained all the vacuum solutions the most complicated of which, the C-NUT (Case III.B in his paper), a twisting generalization of the accelerating C-metric of Levi-Cività, is expressed in terms of Jacobi elliptic functions. All the vacuum D solutions admit, as was observed by Kinnersley and later proved more directly by Hughston and Sommers,[7] Weir and Kerr[8] (in the expanding case), and Czapor and McLenaghan[9] (referred to as CM in the sequel), at least a two-parameter abelian isometry group and, as has been shown by Matravers[10] and CM, a separable Hamilton–Jacobi equation for the null geodesics. The next advance was the integration by Plebański and Demiański[11] of the electrovac field equations with cosmological constant for a particular metric ansatz similar in form to one considered by Carter[12] but incorporating a more general conformal factor derived earlier by Debever[13] in his analysis of Carter's hypotheses, yielding a seven-parameter family of type D solutions expressed in rational functions and interpretable as a charged version with cosmological constant of the C-NUT solution. This result was obtained independently by Weir and Kerr in the vacuum case, from their analysis of expanding algebraically special solutions admitting two commuting Killing vectors. The explicit coordinate transformation relating the seven-parameter solution to the C-NUT solution generalized to include charge and cosmological constant has been given independently by Debever and Kamran[14] based on previous work of Debever[15] and by Ishikawa and Miyashita.[16] The picture of the type D electrovac solutions with cosmological constant is completed by a class of solutions apparently first brought to light by Debever and McLenaghan[17,18] and Debever and Kamran,[19] for which the orbits of the two parameter abelian isometry group are null. The existence of these solutions seems to have been overlooked in the literature because of the generally held assumption that the above mentioned isometry group is invertible. All the previously known solutions, including the null orbit solutions, are special cases of a single expression for the general solution given recently by the authors[20] (in a letter denoted by DKM 2 in what follows).

The purpose of the present paper is to prove the above result, stated without proof in DKM 2, by giving an exhaustive integration[21] of the type D vacuum and electrovac field equations with cosmological constant for an aligned nonsingular Maxwell field (the results of which have been presented without proof in an earlier paper[22] denoted by DKM 1 in the sequel), followed by the derivation, by analytic continuation from the most complicated solution obtained, of the above mentioned single expression for the general solution. The results of this integration procedure are given in Theorems 1 and 2 of Sec. 2. In passing, we give a derivation of the metric which was the starting point of Plebański and Demiański's calculation.

This work is guided by the following principles: (i) unification by a single treatment of all the various particular cases previously considered separately, (ii) a choice of coordinates yielding maximal algebraic simplicity in the expressions for the components of the metric and Maxwell field. The integration is modelled on and completes the work of Carter as generalized by Debever[23] and has as its starting point the canonical forms for the metric and Maxwell field given in the electrovac case by Debever and McLenaghan[24] (DM in the

a) This work was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

sequel) and in the vacuum case by CM, where in the invertible case the two principal null directions are treated in a symmetrical[25] fashion, and where the existence of at least a two-parameter abelian orthogonally transitive isometry group and the separability of the Hamilton–Jacobi equation are shown without explicit integration. Another feature of the work is the proof of generalized versions of Carter's Hamilton–Jacobi and Klein–Gordon separability hypotheses, Theorems 3 and 4 of Sec. 2, as consequences of the type D field equations for the class of solutions considered. These results enable us to exhibit the corresponding (conformal) Killing tensors admitted by the solutions and to distinguish between the accelerating and nonaccelerating solutions.

In Sec. 2 we state the hypotheses and results of the paper and give their relation to the work of others. Sections 3, 4, 5, and 6 contain the proofs of the theorems. We perform the calculations using the Newman–Penrose[26] (NP) formalism and the equivalent complex vectorial formalism of Cahen, Debever, and Defrise.[27] The relationship between these formalisms is given in Debever, McLenaghan, and Tariq[28] (DMT).

2. HYPOTHESIS AND STATEMENT OF RESULTS

We consider solutions of Einstein's vacuum and electrovac field equations with cosmological constant, which may be written as

$$R_{ij} - \tfrac{1}{2}Rg_{ij} + \lambda g_{ij} = F_{ik}F_j{}^k - \tfrac{1}{4}g_{ij}F_{kl}F^{kl}, \tag{2.1a}$$

$$F_{ik;}{}^k = 0, \quad F_{[ij;k]} = 0, \tag{2.1b}$$

where we permit the cosmological constant λ and the electromagnetic field tensor F_{ij} to vanish, which satisfy the following conditions:

H1. The Weyl tensor C_{ijkl} is everywhere of Petrov type D which is equivalent to the existence of real null vector fields l and n satisfying at every point

$$l^j l^k C_{ijk[l} l_{m]} = n^j n^k C_{ijk[l} n_{m]} = 0. \tag{2.2}$$

H2. If the Maxwell field tensor F_{ij} is nonzero it is nonsingular with its principal null directions aligned with the principal null directions of the Weyl tensor, that is we have

$$l^i F_{i[j} l_{k]} = n^i F_{i[j} n_{k]} = 0. \tag{2.3}$$

H3. The invariants of the Weyl tensor and the trace-free Ricci tensor $S_{ij} := R_{ij} - \tfrac{1}{4}Rg_{ij}$ satisfy at least one of the following inequalities:

$$C_{ijkl}{}^*C^{ijkl} \neq 0, \tag{2.4a}$$

$$C_{ijkl}C^{ijkl} \neq \tfrac{4}{3}S_{ij}S^{ij}. \tag{2.4b}$$

The last hypothesis is required to insure that the principal null congruences defined by l and n, respectively, are both geodesic and shear-free, that is that the generalized Goldberg–Sachs[29,30] theorem holds. The exceptional case when H3 is not satisfied has been studied by Plebański and Hacyan[31] and Garcia and Plebański.[32] It is worth noting that H3 is automatically satisfied in the case of the vacuum field equations with or without cosmological constant. Following DM we shall denote the class of solutions of Eqs. (2.1) satisfying H1, H2, and H3 by $\mathfrak{D}$ and the subclass which are solutions of the vacuum equations with or without cosmological constant by $\mathfrak{D}_0$.

The main results of this paper are given in the following theorems.

Theorem 1: For every solution in $\mathfrak{D}$ there exists a system of local coordinates (u,v,w,x) in which the metric and self-dual Maxwell field have the form[33]

$$\begin{aligned} ds^2 = \left[\frac{|\epsilon_1 p(w) - \epsilon_2 m(x)|}{T^2(w,x)}\right]\Bigg\{ & fU(w)\left[\frac{\epsilon_1\,du + m(x)dv}{\epsilon_1 p(w) - \epsilon_2 m(x)}\right]^2 + (1-f^2)\left(\frac{2}{1+f^2}\right)\left[\frac{dw(\epsilon_1\,du + m(x)dv)}{\epsilon_1 p(w) - \epsilon_2 m(x)}\right] \\ & - \left(\frac{2f}{1+f^2}\right)^2\left[\frac{dw^2}{fU(w)}\right] - V(x)\left[\frac{\epsilon_2\,du + p(w)dv}{\epsilon_1 p(w) - \epsilon_2 m(x)}\right]^2 - \frac{dx^2}{V(x)}\Bigg\}, \end{aligned} \tag{2.5a}$$

$$\overset{+}{F} = B(w,x)\left\{dw \wedge \left[\frac{\epsilon_1\,du + m(x)dv}{\epsilon_1 p(w) - \epsilon_2 m(x)}\right] - i\,dx \wedge \left[\frac{\epsilon_2\,du + p(w)dv}{\epsilon_1 p(w) - \epsilon_2 m(x)}\right]\right\}, \tag{2.5b}$$

where ϵ_1, ϵ_2, and f are constants satisfying $\epsilon_1{}^2 + \epsilon_2{}^2 \neq 0$, and all functions are real valued except B which is complex.

The above forms follow by combining the non-null group orbit (when $f \neq 0$) and the null group orbit (when $f = 0$) metrics and Maxwell fields given separately in Theorem 2 of DM (electrovac case) and CM (vacuum case). The group orbits are timelike, spacelike, or null at a given point according to whether $fU(w)$ is positive, negative, or zero, respectively. If $f \neq 0$ it may, under the assumption $U(w) > 0$, be set equal to 1 or -1 corresponding to timelike or spacelike orbits, respectively, in which cases the group is manifestly invertible and hence orthogonally transitive. If $f = 0$ the group is orthogonally transitive but not invertible.

A real vector potential for the Maxwell field F has the following form[34]:

$$A = \left(\frac{\epsilon_2 G(x) + \epsilon_1 H(w)}{\epsilon_1 p(w) - \epsilon_2 m(x)}\right)du + \left(\frac{p(w)G(x) + m(x)H(w)}{\epsilon_1 p(w) - \epsilon_2 m(x)}\right)dv, \tag{2.5c}$$

where G and H are real valued functions of the indicated variables. It should be noted that $F := 2\,dA$ is more special than the F implied by Eq. (2.5b). However, it will be shown that vector potentials of the form (2.5c) exist for every solution in the class $\mathfrak{D}$.

Integration of the remaining Einstein field equations yields the following:

Theorem 2: The general solution $\widetilde{A}$ * in the class $\mathfrak{D}$ may be obtained by specifying the constants ϵ_1 and ϵ_2 and the functions m, p, T, fU, V, and B appearing in (2.5) as follows:

$$\epsilon_1 = b^2 \cos\gamma, \ \ \epsilon_2 = \sin\gamma, \tag{2.6a}$$

$$m(x) = -[c^2x^2 + b^2k^2 + l^2(1-b^2\cos^2\gamma)/\sin^2\gamma] \times \sin\gamma - 2clx, \tag{2.6b}$$

$$p(w) = [b^2c^2w^2 + k^2(b^2 - \sin^2\gamma)/\cos^2\gamma + l^2] \times \cos\gamma + 2b^2ckw, \tag{2.6c}$$

$$T(w,x) = a(cw\cos\gamma + k)(cx\sin\gamma + l) + 1, \tag{2.6d}$$

$$fU(w) = c^2(b^4g_4 + 2a^2B_0\overline{B}_0)w^4\cos^2\gamma + cf_3w^3 \times \cos\gamma + f_2w^2 + f_1w + f_0, \tag{2.6e}$$

$$\begin{aligned} V(x) = {} & c^2g_4x^4\sin^2\gamma + c[acf_1\cos\gamma - 2akf_2 \\ & + 3ak^2f_3 + 4(l - ab^4k^3)g_4 - 8a^3k^3B_0\overline{B}_0]x^3\sin\gamma \\ & + [3aclf_1\cos\gamma - (1+6akl)f_2 + 3k(1+3akl)f_3 \\ & + 6(l^2 - b^4k^2(1+2akl))g_4 - 12a^2k^2 \\ & \times(1+2akl)B_0\overline{B}_0]x^2 + g_1x + g_0, \end{aligned} \tag{2.6f}$$

$$B(w,x) = B_0\left(\frac{b^2(cw\cos\gamma + k) + i(cx\sin\gamma + l)}{b^2(cw\cos\gamma + k) - i(cx\sin\gamma + l)}\right), \tag{2.6g}$$

where $f_0, f_1, f_2, f_3, g_0, g_1, g_4, B_0, a, b, c, k, l$, and γ are real parameters and B_0 is a complex parameter which satisfies the following relations:

$$\begin{aligned} & acg_1\sin\gamma - 3a^2cl^2f_1\cos\gamma \\ & + 2al(1+3akl)f_2 - [1 + 3akl(2+3akl)]f_3 \\ & + 4[b^4k - al^3 + 3ab^4k^2l(1+akl)]g_4 \\ & + 8a^2k[1 + 3akl(1+akl)]B_0\overline{B}_0 = 0, \end{aligned} \tag{2.6h}$$

$$\begin{aligned} & c^2(g_0\sin^2\gamma - b^4f_0\cos^2\gamma) \\ & + c[(2al^3 + b^4k)f_1\cos\gamma - lg_1\sin\gamma] \\ & - (b^4k^2 + l^2 + 4akl^3)f_2 \\ & + (b^4k^3 + 3kl^2 + 6ak^2l^3)f_3 \\ & + (3l^4 - b^8k^4 - 6b^4k^2l^2 - 8ab^4k^3l^3)g_4 \\ & + 2(1 - 6a^2k^2l^2 - 8a^2k^3l^3 - a^2b^4k^4)B_0\overline{B}_0 = 0, \end{aligned} \tag{2.6i}$$

and are otherwise restricted to a range such that the metric is nonsingular and of signature minus two. This requires that $V(x) > 0$ which may also limit the range of the coordinate x. The cosmological constant λ is given in terms of the above parameter by

$$\lambda = -3[a^2c^2f_0\cos^2\gamma - a^2ckf_1\cos\gamma + a^2k^2f_2 - a^3k^3f_3 + (1 + a^2b^4k^4)g_4 + 2a^4k^4B_0\overline{B}_0]. \tag{2.6j}$$

We now list the main special cases of the general solution described in Theorem 2 which are obtained by giving the indicated values to the parameters a, b, c, k, l, and γ in the Eqs. (2.5) and (2.6). The nonzero NP component Ψ_2 of the Weyl tensor is also given in each case. In Sec. 4, it will be shown by direct integration that these solutions exhaust the class $\mathfrak{D}$.

*Case A**: $b = 1, c = \sqrt{2},\ k = l = 0$, and $\gamma = \pi/4$, implying (2.7a)

$$\epsilon_1 = \epsilon_2 = 1/\sqrt{2},\ m(x) = -\sqrt{2}x^2,\ p(w) = \sqrt{2}w^2,\ T(w,x) = awx + 1, \tag{2.7b}$$

$$fU(w) = [a^2(2B_0\overline{B}_0 - f_0) - \tfrac{1}{3}\lambda]w^4 + ag_1w^3 + f_2w^2 + f_1w + f_0, \tag{2.7c}$$

$$V(x) = -(a^2f_0 + \tfrac{1}{3}\lambda)x^4 + af_1x^3 - f_2x^2 + g_1x + f_0 - 2B_0\overline{B}_0, \tag{2.7d}$$

$$B(w,x) = B_0(w+ix)/(w-ix), B_0 = :b_0(\cos\alpha + i\sin\alpha), \tag{2.7e}$$

$$G(x) = b_0x\sin\alpha,\ H(w) = -b_0w\cos\alpha. \tag{2.7f}$$

$$\Psi_2 := \tfrac{1}{2}C_{ijkl}l^in^j(m^k\overline{m}^l - l^kn^l) = \tfrac{1}{2}(1+awx)^3(w-ix)^{-3}[f_1 + ig_1 + 4B_0\overline{B}_0(1-awx)(w+ix)^{-1}]. \tag{2.7g}$$

Case B^0_-: $a = 0,\ b = c = 1,\ l = 0,\ \gamma = \pi/2$, implying (2.8a)

$$\epsilon_1 = 0, \epsilon_2 = 1,\ m(x) = -x^2 - k^2,\ p(w) = 2kw,\ T(w,x) = 1, \tag{2.8b}$$

$$fU(w) = f_2w^2 + f_1w + f_0, \tag{2.8c}$$

$$V(x) = -\tfrac{1}{3}\lambda x^4 - (f_2 + 2k^2\lambda)x^2 + g_1x + k^2f_2 + k^4\lambda - 2B_0\overline{B}_0, \tag{2.8d}$$

$$B(w,x) = B_0(k+ix)/(k-ix),\ B_0 = :b_0(\cos\alpha + i\sin\alpha) \tag{2.8e}$$

$$G(x) = b_0(x\sin\alpha - k\cos\alpha),\ H(w) = -b_0w\cos\alpha. \tag{2.8f}$$

$$\Psi_2 = \tfrac{1}{6}(k-ix)^{-3}[-2k(3f_2 + 4k^2\lambda) + 3ig_1 + 12B_0\overline{B}_0(k+ix)^{-1}]. \tag{2.8g}$$

Case B^0_+: $a = 0,\ b = c = 1,\ k = 0,\ \gamma = 0$, implying (2.9a)

$$\epsilon_1 = 1,\ \epsilon_2 = 0,\ m(x) = -2lx,\ p(w) = w^2 + l^2,\ T(w,x) = 1, \tag{2.9b}$$

$$fU(w) = -\tfrac{1}{3}\lambda w^4 - (g_2 + 2l^2\lambda)w^2 + f_1w + l^2g_2 + l^4\lambda + 2B_0\overline{B}_0, \tag{2.9c}$$

$$V(x) = g_2x^2 + g_1x + g_0, \tag{2.9d}$$

$$B(w,x) = B_0(w+il)/(w-il), B_0 = :b_0(\cos\alpha + i\sin\alpha), \tag{2.9e}$$

$$G(x) = b_0x\sin\alpha,\ H(w) = b_0(l\sin\alpha - w\cos\alpha). \tag{2.9f}$$

$$\Psi_2 = \tfrac{1}{6}(w-il)^{-3}[3f_1 - 2il(3g_2 + 4l^2\lambda) + 12B_0\overline{B}_0(w+il)^{-1}]. \tag{2.9g}$$

*Case C**: $a = 1,\ b = 0,\ c = \sqrt{2},\ k = l = 0,\ \gamma = \pi/4$, implying (2.10a)

$$\epsilon_1 = 0,\ \epsilon_2 = 1/\sqrt{2},\ m(x) = -\sqrt{2}x^2,\ p(w) = 0,\ T(w,x) = wx + 1, \tag{2.10b}$$

$$fU(w) = 2B_0\overline{B}_0w^4 + f_3w^3 + f_2w^2 + f_1w + f_0, \tag{2.10c}$$

$$V(x) = -(f_0 + \tfrac{1}{3}\lambda)x^4 + f_1 x^3 - f_2 x^2 + f_3 x - 2B_0\overline{B}_0, \tag{2.10d}$$

$$B(w,x) = -B_0 =: -b_0(\cos\alpha + i\sin\alpha), \tag{2.10e}$$

$$G(x) = b_0 x \sin\alpha, \quad H(w) = -b_0 w \cos\alpha, \tag{2.10f}$$

$$\Psi_2 = \tfrac{1}{2}x^{-4}(1+wx)^3[f_3 x + 4B_0\overline{B}_0(1-wx)]. \tag{2.10g}$$

$$\text{Case } C^{00}\text{:} \quad a=0,\ b=1,\ c=0,\ k=0,\ l=1,\ \gamma=0, \text{ implying} \tag{2.11a}$$

$$\epsilon_1 = 1,\ \epsilon_2 = 0,\ p(w) = 1,\ m(x) = 0,\ T(w,x) = 1, \tag{2.11b}$$

$$fU(w) = (2B_0\overline{B}_0 - \lambda)w^2 + f_1 w + f_0, \tag{2.11c}$$

$$V(x) = -(2B_0\overline{B}_0 + \lambda)x^2 + g_1 x + g_0, \tag{2.11d}$$

$$B(w,x) = -B_0 =: -b_0(\cos\alpha + i\sin\alpha), \tag{2.11e}$$

$$G(x) = -b_0 x\sin\alpha, \quad H(w) = -b_0 w\cos\alpha. \tag{2.11f}$$

$$\Psi_2 = -\tfrac{1}{3}\lambda. \tag{2.11g}$$

We now describe the relation between our solution and the main classes of type D solutions of Eq.(2.1) given in the literature. By setting $a = 0$, $b = f = 1$ in Eqs. (2.5) and (2.6) we obtain the following family of solutions:

$$\epsilon_1 = \cos\gamma,\ \epsilon_2 = \sin\gamma,\ T(w,x) = 1, \tag{2.12a}$$

$$m(x) = -(c^2x^2 + k^2 + l^2)\sin\gamma - 2clx, \quad p(w) = (c^2w^2 + k^2 + l^2)\cos\gamma + 2ckw, \tag{2.12b}$$

$$U(w) = -\tfrac{1}{3}\lambda c^2 w^4\cos^2\gamma - \tfrac{4}{3}\lambda ckw^3\cos\gamma + f_2 w^2 + f_1 w + f_0, \tag{2.12c}$$

$$V(x) = -\tfrac{1}{3}\lambda c^2x^4\sin^2\gamma - \tfrac{4}{3}\lambda clx^3\sin\gamma - [f_2 + 2\lambda(k^2 + l^2)]x^2 + g_1 x + g_0, \tag{2.12d}$$

$$B_0\overline{B}_0 = \tfrac{1}{2}[c^2(f_0\cos^2\gamma - g_0\sin^2\gamma) + c(lg_1\sin\gamma - kf_1\cos\gamma) + (k^2+l^2)f_2 + \lambda(k^2+l^2)^2]. \tag{2.12e}$$

Modulo an obvious substitution and relabeling this is Carter's[35] $[\widetilde{A}]$ class of solutions which he characterized as the most general solution of the Eqs. (2.1) for which the metric and Maxwell fields admit a two-parameter, abelian, invertible isometry group and for which the Klein–Gordon equation for a massive, charged spin-zero field is soluble by separation of variables (Conditions I, II, IIIS, and IV of his paper). Such solutions are necessarily of Petrov type D. In Theorem III to follow we shall give a converse of this result. If we set $c = \sqrt{2}$, $k = l = 0$, $\gamma = \pi/4$ in the Eqs. (2.12) or equivalently $a = 0$, $b = f = l$ in our A^* solutions we obtain Carter's $[A]$ class which has been rediscovered by Plebański.[36] This family of solutions is the most complicated branch of the $[\widetilde{A}]$ solutions depending as it does on six essential arbitrary parameters and admitting only a two-parameter isometry group. It reduces to the Demiański–Newman[37] solution when $\lambda = 0, f_2 > 0$, $g_1{}^2 + 4f_2(f_0 - 2B_0\overline{B}_0) > 0$, to the charged Kerr solution of Newman *et al.*[38] when in addition $g_1 = 0$ and to the Kerr[39] solution when furthermore $B_0 = 0$. When $f = 1$ in our B^0_- and B^0_+ solutions we obtain Carter's $[\widetilde{B}(-)]$ and $[\widetilde{B}(+)]$ class of solutions, respectively, which include the charged version with cosmological constant of the Taub[40]-NUT[41] solutions and their charged generalization due to Brill.[42] These solutions depend in general on five essential arbitrary constants and possess a four-parameter isometry group. Setting $f = 1$ in our C^{00} solution yields Carters $[D]$ class of solutions, the complete class of which was first discovered by Bertotti[43] and Robinson.[44] These solutions depend in general on three essential arbitrary constants and admit a six-parameter isometry group. Kinnersley's[45] class of vacuum type D solutions may be obtained in a single coordinate system by setting $B_0 = \lambda = 0$ in the Eqs. (2.6) defining our general solution $\widetilde{A}^*$. Setting $a = f = 1$ in our A^* solution yields after the substitution $u \to \sqrt{2}u$, $v \to (1/\sqrt{2})v$ and upon setting

$$u = \tau,\ v = \sigma,\ w = q,\ x = -p \tag{2.13a}$$

$$B_0\overline{B}_0 = \tfrac{1}{2}(e_0{}^2 + g_0{}^2), \quad f_0 = \gamma + e_0{}^2 - \tfrac{1}{6}\lambda,$$

$$f_1 = -2m, \quad f_2 = \epsilon, \quad g_1 = -2n, \tag{2.13b}$$

and a change of signature, the seven-parameter family of solutions given by Plebański and Demiański[46] in Eqs. (3.30) and (3.31) of their paper. These solutions generalize Carter's $[A]$ class of solutions to include the charged version with cosmological constant of Kinnersley's Case III.B solutions in a different coordinate system where the metric and Maxwell field functions are rational functions of the coordinates rather than elliptic functions (see also Weir and Kerr,[47] Debever and Kamran,[48] and Ishikawa and Miyashita[49]). The generalization is effected by allowing a more general conformal factor (see Debever[50]) which implies nonzero third degree terms in the metric functions U and V given by Eqs. (2.8b) and (2.8c). A charged version with cosmological constant of Levi-Civitá's[51] vacuum solution due to Plebański and Demiański, Newman and Tamburino,[52] Robinson and Trautman,[53] and Ehlers and Kundt[54] may be obtained from our C^* solution by setting $f = 1$ in the Eqs. (2.5) and (2.10). This solution depends on six essential arbitrary parameters and admits a two-parameter isometry group. Both of the above solutions have the property which will be stated precisely in Theorems III and IV that the Hamilton–Jacobi equation for the massless particle orbits and the conformally invariant wave equation satisfied by massless spin-zero fields are solvable by separation of variables[55,56] while the corresponding massive equations are not.

If $f = -1$ in the family of solutions A^*, B^0_+, B^0_-, C^*, and C^{00} we obtain solutions with spacelike group orbits. However, these solutions are not essentially different from the ones enumerated above where $f = 1$ if we allow the sign of the function U appearing therein to be negative. This sign is determined by the relevant metric parameters and the range of the coordinate w.

It remains to consider the solutions in $\mathfrak{D}$ corresponding to $f=0$ when the orbits of the two-parameter abelian isometry group are null. These solutions seem to have been overlooked in literature and have only recently been determined. The most complicated of these solutions which we denote by A_0 is obtained from the class A^* by setting $a=f=0$ in the Eqs. (2.5) and (2.7) which yields on rescaling the coordinates u and v the following metric and Maxwell field:

$$ds^2 = (w^2+x^2)\left[\frac{2\,dw(du - x^2\,dv)}{w^2+x^2} - (g_1x - 2B_0\overline{B}_0)\times\left(\frac{du + w^2\,dv}{w^2+x^2}\right)^2 - \frac{dx^2}{g_1x - 2B_0\overline{B}_0}\right], \tag{2.14a}$$

$$\overset{+}{F} = B_0\left(\frac{w+ix}{w-ix}\right)\left[dw\wedge\left(\frac{du - x^2\,dv}{w^2+x^2}\right) - i\,dx\wedge\left(\frac{du+w^2\,dv}{w^2+x^2}\right)\right], \tag{2.14b}$$

$$\lambda = 0. \tag{2.14c}$$

Up to an obvious relabeling this is the solution first brought to light by two of us.[57] It depends on three essential arbitrary constants, admits only to a two-parameter abelian isometry group, and possesses separable Hamilton–Jacobi and Klein–Gordon equations. It reduces to a vacuum solution *iff* $B_0=0$, in which case one recovers Kinnersley's Case II.E solution with $b=m=0$. When $B_0\neq 0$, the above solution is a special case of a class of type D solutions given by Debever[58] [obtained by setting $b=0$ in his Eq. (2.12)]. It is also a special case of a family of solutions obtained by Leroy[59] under different hypotheses[60] [given by his Eq. (3.36) with $b=0$]. For further reference we denote by A^0 the solution obtained by setting $a=0$ in the A^* solution. The A^0 solution contains Carter's $[A]$ solution and the null orbit solution A_0.

A second class of null orbit solutions is obtained from the family B^0_- by setting $f=0$ in Eqs. (2.5) and (2.8) which yields after the substitution $u\to -u$, $v\to -v$ the following metric and Maxwell field:

$$ds^2 = (x^2+k^2)\left[2\,dw\,dv - V(x)\times\left(\frac{du+2kw\,dv}{x^2+k^2}\right)^2 - \frac{dx^2}{V(x)}\right], \tag{2.15a}$$

$$\overset{+}{F} = B_0\left(\frac{k+ix}{k-ix}\right)\left[dw\wedge dv + i\,dx\wedge\left(\frac{du+2kw\,dv}{x^2+k^2}\right)\right], \tag{2.15b}$$

where

$$V(x) = -\tfrac{1}{3}\lambda x^4 - 2k^2\lambda x^2 + g_1x + k^4\lambda - 2B_0\overline{B}_0. \tag{2.15c}$$

This solution which we denote by $B^0_{-,0}$ has been given previously by us in DKM 1 and DKM 2. The solution depends, in general, on four essential arbitrary constants since the parameter k if nonzero may be set equal to 1 by rescaling the coordinates. If $k=0$ one obtains a more special null orbit solution[61] which we denote by $C^0_{-,0}$ in DKM 1 and which depends on three essential arbitrary constants. The $B^0_{-,0}$ solution is in fact a special case of Carter's $[\widetilde{B}(-)]$ family obtained by setting $h=0$ in his Eqs. (13) and (14) or equivalently of our B^0_- family obtained by setting $f=1, f_2=0$ in the Eqs. (2.5) and (2.8) which condition is equivalent to the vanishing of the Gaussian curvature of the two-dimensional metric $U(w)dv^2 - U^{-1}(w)dw^2$. The explicit coordinate transformation which yields the $B^0_{-,0}$ solution from the above B^0_- solution is given in the case $f_1\neq 0$ by

$$u\to u + kvw + 4kf_0f_1^{-2}\tanh^{-1}[(w+2v)/(w-2v)], \tag{2.16a}$$

$$v\to 2f_1^{-1}\tanh^{-1}[(w+2v)/(w-2v)],\quad w\to -\tfrac{1}{2}f_1vw - f_0f_1^{-1},\quad x\to x, \tag{2.16b}$$

with a similar but simpler transformation holding when $f_1=0, f_0\neq 0$.

A third class of null orbit solutions may be obtained from the family C^{00} by setting $f=0$ in the Eqs. (2.5) and (2.11) which gives the following metric and Maxwell field:

$$ds^2 = 2\,dv\,dw - V(x)dv^2 - (V(x))^{-1}\,dx^2, \tag{2.17a}$$

$$\overset{+}{F} = B_0(du\wedge dw + i\,dx\wedge dv), \tag{2.17b}$$

where

$$V(x) = -2\lambda x^2 + g_1x + g_0, \tag{2.17c}$$

$$B_0\overline{B}_0 = \tfrac{1}{2}\lambda. \tag{2.17d}$$

This solution which we denote by $C^{00}{}_{,0}$ has been given previously by the authors in DKM 1 and in a slightly different coordinate system in DKM 2. In general it depends on two essential arbitrary constants. This solution, in analogy with the property of the $B^0_{-,0}$ just discussed, is a special case of Carter's $[D]$ family (the Bertotti–Robinson solutions) which may be obtained by setting $\Lambda = -e^2$ in his Eqs. (17) and (18) or equivalently a special case of our C^{00} family which may be obtained by setting $f=1$, $\lambda = 2B_0\overline{B}_0$ in Eqs (2.5) and (2.11). The explicit coordinate transformation which yields the $C^{00}{}_{,0}$ solution from the above described C^{00} solution is given in the case of $f_1\neq 0$ by $u\to u$ and the Eq. (2.16b) and in the case $f_1=0, f_0\neq 0$, by a similar but simpler transformation.

In Sec. 5 a proof (a sketch of which is presented in DKM 1) is given that A_0, $B^0_{-,0}$, $C^0_{-,0}$, and $C^{00}{}_{,0}$ are the only solutions in $\mathfrak{D}$ for which the orbits of the two-parameter abelian isometry group are everywhere null. We denote by $\widetilde{A}^0$ the solution in $\mathfrak{D}$ obtained by setting $a=0$ in the solution $\widetilde{A}^*$. The $\widetilde{A}^0$ solution contains Carter's $[\widetilde{A}]$ solutions and all the null orbit solutions.

The separability of the Hamilton–Jacobi (HJ) equation for the zero rest-mass particle orbits (null geodesics) in a canonical coordinate system for every solution in $\mathfrak{D}$ has been established in DM (electrovac case) and CM (vacuum case). The corresponding result for the HJ equation for the charged particle orbits is given in the following theorem:

Theorem 3: The Hamilton–Jacobi equation for a particle of mass m and charge e

$$\frac{1}{2}g^{ij}\left(\frac{\partial S}{\partial x^i} - eA_i\right)\left(\frac{\partial S}{\partial x^j} - eA_j\right) = \frac{\partial S}{\partial \tau} = \frac{1}{2}m^2, \tag{2.18}$$

where τ is an affine parameter, the g^{ij} are the contravariant components the metric tensor defined by Eq. (2.5a) and A_i

are the covariant components of the vector potential defined by Eq. (2.5c), is solvable by separation of variables in the canonical coordinates (u,v,w,x) of Theorem 1 if $m = 0$. If $m \neq 0$, the equation is solvable by separation of variables iff the condition

$$\frac{\partial^2}{\partial w\, \partial x}\, [T^{-2}(\epsilon_1 p - \epsilon_2 m)] = 0 \qquad (2.19)$$

is satisfied.

Proof: If S of the form

$$S = \tfrac{1}{2}m^2\tau + \alpha u + \beta v + S_1(w,\alpha,\beta,\gamma,\delta) + S_2(x,\alpha,\beta,\gamma,\delta), \qquad (2.20)$$

where α, β, γ, and δ are constant, is substituted into the Eq. (2.18) one obtains the following equation:

$$\begin{aligned}
&-fU(w)(S_1'(w))^2 + 2(1+f^2)^{-1}(1-f^2)\\
&\quad\times S_1'(w)(\alpha p(w) - \epsilon_2\beta - eH(w))\\
&\quad + 4f(1+f^2)^{-2}(U(w))^{-1}(\alpha p(w) - \epsilon_2\beta\\
&\quad - eH(w))^2 - V(x)(S_2'(x))^2\\
&\quad - (V(x))^{-1}(\epsilon_1\beta - \alpha m(x) - eG(x))^2\\
&\qquad = m^2(T(w,x))^{-2}(\epsilon_1 p(w) - \epsilon_2 m(x)). \qquad (2.21)
\end{aligned}$$

If $m = 0$, the above equation clearly separates into a pair of ordinary differential equations for the functions S_1 and S_2, respectively. If $m \neq 0$, this separation is possible iff the additional condition

$$(T(w,x))^{-2}(\epsilon_1 p(w) - \epsilon_2 m(x)) = g(w) + h(x) \qquad (2.22)$$

is satisfied for some single variable functions g and h. This condition is clearly equivalent to the condition (2.19).

An examination of the Eqs. (2.7) to (2.11) shows that the condition (2.19) is satisfied in the cases B^0_-, B^0_+, and C^{00}, and in the case A^* if $a = 0$ and is not satisfied in the case C^*. Since these cases exhaust the class $\mathfrak{D}$ we have proved the following corollary:

Corollary: The Hamilton–Jacobi equation (2.18) for a particle of mass m and charge e is solvable by separation of variables in the case $m = 0$ in the canonical coordinates (u,v,w,x) of Theorem 1 for every solution in the class $\mathfrak{D}$. In the case $m \neq 0$ the equation is solvable by separation of variables only in the cases A^0, B^0_-, B^0_+, and C^{00} that is in the $\widetilde{A}^0$ family of solutions.

This corollary shows that Carter's separability condition III is satisfied in only the subclass A^0 of the class $\mathfrak{D}$ of solutions but that it holds in the entire class $\mathfrak{D}$ if it is weakened to apply to the HJ equation for the zero rest-mass (charged) particles only.

The existence of a conformal Killing tensor for every solution in $\mathfrak{D}$ arising from the separability of the HJ equation has been shown in DM (electrovac case) and CM (vacuum case). The existence of a full Killing tensor B^{ij} in the subclass $\widetilde{A}^0$ may be established in a similar manner from the separability therein of the HJ equation for the massive charged particle orbit. Explicitly one has

$$\begin{aligned}
&B^{ij}(p_i - eA_i)(p_j - eA_j)\\
&\quad := (g+h)^{-1}\{h[-fUp_w^2 + 2(1+f^2)^{-1}\\
&\qquad \times(1-f^2)p_w(p(p_u - eA_u) - \epsilon_2(p_v - eA_v))\\
&\qquad + 4f(1+f^2)^{-2}U^{-1}(p(p_u - eA_u) - \epsilon_2(p_v - eA_v))^2]\\
&\qquad + g[Vp_x^2 + V^{-1}(\epsilon_1(p_v - eA_v) - m(p_u - eA_u))^2]\}\\
&\qquad =: K,
\end{aligned} \qquad (2.23a)$$

where

$$p_i = \frac{\partial S}{\partial x^i} \qquad (2.23b)$$

denote the components of the canonical momentum. The quantity K appearing in Eq. (2.23) is precisely the separation constant that arises from Eq. (2.21) when Eq. (2.22) holds. As such it must be constant along each charged particle orbit. This property may also be established by showing the vanishing of the Poisson bracket (see Carter[62])

$$[K,H] = 0, \qquad (2.24)$$

where

$$\begin{aligned}
H :&= \tfrac{1}{2}g^{ij}(p_i - eA_i)(p_j - eA_j) \qquad (2.25)\\
&= \tfrac{1}{2}T^2Z^{-1}[-fUp_w^2 + 2(1-f^2)\\
&\quad\times(1+f^2)^{-1}p_w(pp_u - \epsilon_2 p_v - eH)\\
&\quad + 4f(1+f^2)^{-2}U^{-1}(pp_u - \epsilon_2 p_v - eH)^2\\
&\quad - Vp_x^2 - V^{-1}(\epsilon_1 p_v - mp_u - eG)^2], \qquad (2.26)
\end{aligned}$$

is the Hamiltonian of the charged particle orbits and Eq. (2.22) again holds. It thus follows[63] that the tensor B_{ij} defined by Eq. (2.23a) must satisfy the Killing tensor equation

$$B_{(ij;k)} = 0, \qquad (2.27)$$

and the additional condition

$$B_{k(i}F_{j)}{}^k = 0. \qquad (2.28)$$

The tensor B_{ij} gives an explicit form for the Killing tensor whose existence in the $\widetilde{A}^0$ solutions (in the case $\lambda = 0$) has been shown by Hughston and Sommers[64] from the existence of an appropriate two-index Killing spinor.

The class of solutions $\mathfrak{D}$ also possesses stronger separability properties an example of which is provided by the separability of the Klein–Gordon (KG) equation which governs the motion of a massive, charged, spin-zero test particle in an external field in the first quantized limit (for a discussion of such a theory see Carter[65]). The precise result for this case is given in the following theorem:

Theorem 4: The Klein–Gordon equation for a spin-zero field of mass m and charge e given by

$$\begin{aligned}
\mathbf{H}\psi :&= \left(g^{-1/2}\frac{\partial}{\partial x^j}g^{1/2} - ieA_j\right)g^{jk}\left(\frac{\partial}{\partial x^k} - ieA_k\right)\psi\\
&\quad + \left(\frac{1}{6}R + m^2\right)\psi = 0, \qquad (2.29)
\end{aligned}$$

where g^{ij} are the contravariant components of the metric tensor defined by Eq. (2.5a), where A_i are the covariant components of the vector potential defined by Eq. (2.5c), where

$$g = |\det(g_{ij})|, \qquad (2.30)$$

and where R denotes the curvature scalar, admits in the case $m = 0$ an R-separable solution[66] of the form

$$\psi(u,v,w,x) = e^{i(\alpha u + \beta v)}\psi_1(w)\psi_2(x)T(w,x), \qquad (2.31)$$

where (u,v,w,x) are the canonical coordinates of Theorem 1, provided that the Weyl tensor of the metric (2.5a) is of Petrov type D or equivalently that the metric functions p and m satisfy

$$(\epsilon_1 p(w) - \epsilon_2 m(x))p''(w) - \epsilon_1\epsilon_2(p'(w))^2 = (\epsilon_1 p(w) - \epsilon_2 m(x))m''(x) - \epsilon_1\epsilon_2(m'(x))^2. \tag{2.32}$$

If $m \neq 0$, the Eq. (2.29) admits an R-separable solution of the form of Eq. (2.31) iff the condition (2.19) is satisfied.

Proof: A proof of this theorem will be given in Sec. 6.

A comparison of Theorem 3 and Theorem 4 reveals that KG separability is more restrictive than HJ separability since the metric functions p and m are restricted by the type D condition (2.32) in the former case but not in the latter. By hypothesis all solutions in $\mathfrak{D}$ are of Petrov type D thus satisfying the condition (2.32). Furthermore an examination of the Eqs. (2.7) to (2.11) shows that the condition (2.19) is satisfied in the cases A^0, B^0_-, B^0_+, and C^{00} but is not satisfied in the case A^* with $a \neq 0$ nor in the case C^*. Since these cases exhaust the class $\mathfrak{D}$ we have proved the following corollary to Theorem 3:

Corollary: The Klein–Gordon equation (2.29) for a particle of mass m and charge e admits an R-separable solution of the form (2.31) in the canonical coordinates of Theorem 2 in the case $m = 0$ for every solution in the class $\mathfrak{D}$. In the case $m \neq 0$ the equation admits an R-separable solution only for the cases A^0, B^0_-, B^0_+, and C^{00} that is in the $\widetilde{A}^0$ family of solutions.

The above corollary shows that, in analogy with Carter's separability condition III, his Schrödinger separability condition IIIS is satisfied only in the subclass $\widetilde{A}^0$ of $\mathfrak{D}$ but that it holds in the entire class $\mathfrak{D}$ if it is weakened to apply to the KG equation (2.29) in the case $m = 0$ when the equation is conformally invariant. Some care is required to reach the above conclusion since Carter's Klein–Gordon equation [Eq. (39) in his paper] differs from our Eq. (2.29) by the omission of the term $\frac{1}{6}R\psi$. However, this is immaterial since R is constant for the $\mathfrak{D}$ solutions implying that the two equations are identical modulo a redefinition of the mass of Eq. (2.29). However, if one wishes to work in the more general context of the canonical forms of Theorem 1 without imposing the type D condition or the remaining electrovac field equations one finds a different separability result for Carter's form of the KG equation than that given in Theorem 4: namely, the KG Eq. (2.29) where the term $\frac{1}{6}R\psi$ has been omitted admits a separable solution of the form

$$\psi(u,v,w,x) = e^{i(\alpha u + \beta v)}\psi_1(w)\psi_2(w), \tag{2.33}$$

in the case $m \neq 0$ provided that the condition

$$T(w,x) = \text{const} \tag{2.34}$$

is satisfied. We note that the imposition of the type D condition (2.32) is not required to obtain this result. Thus we have shown that Carter's separability condition IIIS holds for the canonical metric and vector potential of Theorem 1 when the condition (2.34) is satisfied. However, if the metric function T is not constant the equation does not in general admit an R-separable solution of the form (2.31) even in the case $m = 0$. This is the reason one is lead to consider the conformally invariant equation (2.29).

In analogy with the case of the HJ equation the separability of the KG equation in $\mathfrak{D}$ gives rise to a symmetry operator $\mathbf{K}$ of the form

$$\mathbf{K} := (\nabla_j - ieA_j)\widehat{B}^{jk}(\nabla_k - ieA_k), \tag{2.35}$$

where ∇_j denotes the covariant derivative, which commutes with the KG operator $\mathbf{H}$ in the case $m \neq 0$ and which R-commutes with this operator when $m = 0$, that is, it satisfies

$$[\mathbf{K},\mathbf{H}] = r\mathbf{H}, \tag{2.36}$$

where the bracket denotes the usual operator commutator and r is some function. The operator $\mathbf{K}$ has the property that the R-separable solution (2.31) is an eigenfunction of $\mathbf{K}$ corresponding to the separation constant as eigenvalue. The symmetric tensor $\widehat{B}_{ij}$ implied by the existence of $\mathbf{K}$ satisfies the Killing tensor Eq. (2.27) in the case $m \neq 0$ and is identical with the Killing tensor defined by Eq. (2.23). In the case $m = 0$, the tensor $\widehat{B}_{ij}$ satisfies the conformal Killing tensor equation

$$\widehat{B}_{(ij;k)} = \tfrac{1}{3}g_{(ij}\widehat{B}_{k)}{}^{l}{}_{;l}, \tag{2.37}$$

and is identical with the conformal Killing-tensor defined by Eq. (2.23) which exists in this case. Details of the above results will be given elsewhere.

3. PROOF OF THEOREM 1

It is shown in DM (electrovac case) and CM (vacuum case) that the hypotheses H_1 to H_3 and the field equations (2.1) imply the existence of a canonical null tetrad $(l,n,m,\overline{m})$ such that l and n are principal null vectors of the Weyl tensor C_{ijkl} and the Maxwell field F_{ij} and such that the corresponding NP spin coefficients satisfy the conditions

$$\kappa = \sigma = \nu = \lambda = 0, \tag{3.1a}$$

$$I_1 := \pi\overline{\pi} - \tau\overline{\tau} = 0, \tag{3.1b}$$

$$I_2 := \overline{\mu}\rho - \mu\overline{\rho} = 0, \tag{3.1c}$$

with $\Psi_2 = :\Psi$, $\Phi_{11} = :\Phi$, and $\Lambda = R/24 = \frac{1}{6}\lambda$ being the only nonzero curvature components. On account of the conditions (3.1b) and (3.1c), the transformation

$$l' = e^a l, \quad n' = e^{-a}n, \quad m' = e^{ib}m \tag{3.2}$$

which preserves the directions of l and n may be used to set

$$\pi = \tau, \tag{3.3a}$$

$$\mu = f\rho, \tag{3.3b}$$

where f is a constant which may be zero. We note that if $f \neq 0$ we may use the transformation (3.2) to set $f = -e$, where $e^2 = 1$; this is the choice that was made in DM and CM where in addition one has $\pi = -e\tau$ instead of Eq. (3.3a). However, as will be shown, we are able by leaving f arbitrary to treat simultaneously the non-null group orbit ($f \neq 0$) and the null group orbit ($f = 0$) cases.

When Eqs. (3.3) are taken into account in the NP curvature equations [DM Eqs. (4.21) and CM Eqs. (3.22) to (3.41)] we obtain in a manner similar to that employed in DM and CM for the Cases, I, IIa, IIb, and III defined therein the following additional relations between the spin coefficients

$$\gamma = f\epsilon, \tag{3.4a}$$

$$\beta = \alpha, \tag{3.4b}$$

$$\tau + \overline{\tau} = 2(\alpha + \overline{\alpha}), \tag{3.4c}$$

$$\rho - \bar{\rho} = 2(\epsilon - \bar{\epsilon}), \quad (3.4d)$$

and the following derivative conditions

$$\mathscr{D}\rho = \mathscr{D}\tau = \mathscr{D}\epsilon = \mathscr{D}\alpha = 0, \quad (3.4e)$$

$$\mathscr{L}\rho = \mathscr{L}\tau = \mathscr{L}\epsilon = \mathscr{L}\alpha = 0, \quad (3.4f)$$

where

$$\mathscr{D} := fD + \Delta, \quad \mathscr{L} := \delta + \bar{\delta}, \quad (3.4g)$$

where D, Δ, δ, and $\bar{\delta}$ denote the NP operators.

A quantity which will be useful in the sequel is the complex 1-form θ defined by

$$\theta := -2(\rho\theta^1 - \mu\theta^2 + \tau\theta^3 - \pi\theta^4), \quad (3.5)$$

where the θ^a denote a null tetrad of 1-forms defined by

$$\theta^1 := n_i\, dx^i, \quad \theta^2 := l_i\, dx^i, \quad \theta^3 := -\bar{m}_i\, dx^i = :\bar{\theta}^4. \quad (3.6)$$

The metric and, in view of $H\,2$, the self-dual Maxwell field of the solutions may be expressed in terms of these 1-forms as

$$ds^2 = 2\theta^1\theta^2 - 2\theta^3\theta^4, \quad (3.7)$$

$$\overset{+}{F} = BZ^2, \quad (3.8)$$

where B is a complex valued function and where

$$Z^2 = \theta^1 \wedge \theta^2 - \theta^3 \wedge \theta^4. \quad (3.9)$$

Under the conformal transformation of the basis 1-forms

$$\tilde{\theta}^a = e^{\psi}\theta^a, (a = 1,2,3,4), \quad (3.10)$$

the 1-form θ transforms as

$$\tilde{\theta} = \theta + 2\, d\psi. \quad (3.11)$$

It follows by choosing the conformal factor to satisfy

$$d\psi = -\tfrac{1}{4}(\theta + \bar{\theta}), \quad (3.12)$$

that

$$\tilde{\theta} + \bar{\tilde{\theta}} = 0. \quad (3.13)$$

The choice (3.12) is always possible on account of the integrability condition for the vacuum Bianchi identities[67] or for Maxwell's equations[68] (2.1) which may be expressed as

$$d\theta = 0. \quad (3.14)$$

By virtue of Eqs. (3.3) and (3.5), the Eq. (3.13) implies

$$\bar{\tilde{\rho}} = -\tilde{\rho}, \quad \bar{\tilde{\tau}} = \tilde{\tau}. \quad (3.15)$$

In an analogous manner to that employed in DM it may be shown that the conformal transformation (3.10), where ψ is a solution of Eq. (3.12) preserves the relations (3.1), (3.3), and (3.4) satisfied by the spin coefficients and that in addition the transformed curvature components satisfy

$$\tilde{\Psi}_0 = \tilde{\Psi}_1 = \tilde{\Psi}_3 = \tilde{\Psi}_4 = 0, \quad (3.16a)$$

$$\tilde{\Phi}_{22} = f^2\tilde{\Phi}_{00}, \quad \tilde{\Phi}_{20} = \tilde{\Phi}_{02}, \quad \tilde{\Phi}_{12} = f\tilde{\Phi}_{10}. \quad (3.16b)$$

In view of the relations (3.1a), (3.3), (3.4), and (3.15) Cartan's first structure equations have the form

$$d\tilde{\theta}^1 = f[(2\tilde{\epsilon} - \tilde{\rho})\tilde{\theta}^1 \wedge \tilde{\theta}^2 - 2\tilde{\rho}\tilde{\theta}^3 \wedge \tilde{\theta}^4], \quad (3.17a)$$

$$d\tilde{\theta}^2 = (2\tilde{\epsilon} - \tilde{\rho})\tilde{\theta}^1 \wedge \tilde{\theta}^2 - 2\tilde{\rho}\tilde{\theta}^3 \wedge \tilde{\theta}^4, \quad (3.17b)$$

$$d\tilde{\theta}^3 = -2\tilde{\tau}\tilde{\theta}^1 \wedge \tilde{\theta}^2 + (2\tilde{\alpha} - \tilde{\tau})\tilde{\theta}^3 \wedge \tilde{\theta}^4 = d\tilde{\theta}^4. \quad (3.17c)$$

It follows from the Frobenius integration theorem that there exists a local coordinate system (u,v,w,x) such that

$$\tilde{\theta}^1 = (\sqrt{2}/(1+f^2))[f(L\, du + M\, dv) + R\, dw], \quad (3.18a)$$

$$\tilde{\theta}^2 = (\sqrt{2}/(1+f^2))[L\, du + M\, dv - fR\, dw], \quad (3.18b)$$

$$\tilde{\theta}^3 = (1/\sqrt{2})[N\, du + P\, dv + iT\, dx], \quad (3.18c)$$

where L, M, N, P, R, and T are real-valued functions of the four coordinates. If one replaces the differential forms in Eqs. (3.17) by their values given by Eqs. (3.18) and equates the corresponding coefficients of the differentials one obtains expressions for the spin coefficients in terms of the tetrad functions and a set of restrictions on these functions. If one then performs an appropriate transformation of the coordinates u and v which preserves the form of the tetrad (3.18) and which is permitted by virtue of the Eqs.(3.1), (3.3), (3.4), and (3.16) following a procedure which is described in Sec. 9 of DMT (non-null orbit case) and Sec. 7 of DM (null orbit case), one finds that all the functions appearing in the transformed tetrad of the form (3.18) are independent of u and v.

In addition it follows that

$$R_x = T_w = 0 \quad \left(R_x = \frac{\partial R}{\partial x}, \text{etc.}\right), \quad (3.19a)$$

$$(LZ)_w = (MZ)_w = (NZ)_x = (PZ)_x = 0, \quad (3.19b)$$

where

$$Z := (LP - MN)^{-1}, \quad (3.19c)$$

and that the function ψ appearing in the conformal transformation (3.10) satisfies

$$\psi_u = \psi_v = 0. \quad (3.20)$$

It now follows from Eqs. (3.7), (3.8), (3.10), and (3.18) that general forms of the metric and self-dual Maxwell field for the $\mathfrak{D}$ solutions are

$$\begin{aligned} ds^2 = e^{2\psi}[&4f(1+f^2)^{-2}(L\, du + M\, dv)^2 \\ &+ 4(1-f^2)(1+f^2)^{-2}R\, dw(L\, du + M\, dv) \\ &- 4f(1+f^2)^{-2}R^2\, dw^2 - (N\, du + P\, dv)^2 - T^2\, dx^2], \end{aligned} \quad (3.21a)$$

$$\begin{aligned} \overset{+}{F} = B\,[&2(1+f^2)^{-1}R\, dw \wedge (L\, du + M\, dv) \\ &- iT\, dx \wedge (N\, du + P\, dv)], \end{aligned} \quad (3.21b)$$

where the substitutions $\psi \to -\psi$ and $B \to e^{2\psi}B$ have been made. We note that the expressions (3.21a) for ds^2 and (3.21b) for $\overset{+}{F}$ reduce respectively to DM Eq. (2.12) [or CM Eq. (2.10) in the vacuum case] and DM Eq. (2.6b) when $f = -e$, where $e^2 = 1$, and DM Eq. (2.6c) is taken into account modulo the substitution $w \to -w$, and reduces to DM Eq. (2.13) [or CM Eq. (2.11) in the vacuum case] and DM Eq. (2.7b) when $f = 0$ upon the substitution $w \to w/2$. Thus we have shown that the null and non-null canonical forms for the metric and Maxwell field of the $\mathfrak{D}$ solutions presented separately in DM (electrovac case) and CM (vacuum case) are special cases, respectively, of the single canonical form (3.21a) for the metric and (3.21b) for the Maxwell field.

We now proceed with our derivation of the canonical forms (2.5) by the integration of the Eqs. (3.19b) which yields

$$L = l/Z, \quad M = m/Z, \quad N = n/Z, \quad P = p/Z, \quad (3.22)$$

where l and m are functions independent of w, and n and p

are functions indepedent of x. It follows from the Eqs. (3.19b) and (3.19c) that

$$Z = lp - mn. \tag{3.23}$$

The tetrad (3.18) may thus be reexpressed as

$$\tilde{\theta}^1 = (\sqrt{2}/(1+f^2))[f(RZ)^{-1}(l\,du + m\,dv) + R\,dw], \tag{3.24a}$$

$$\tilde{\theta}^2 = (\sqrt{2}/(1+f^2))[(RZ)^{-1}(l\,du + m\,dv) - fR\,dw], \tag{3.24b}$$

$$\tilde{\theta}^3 = (1/\sqrt{2})[(TZ)^{-1}(n\,du + p\,dv) + iT\,dx], \tag{3.24c}$$

where the substitutions $l \to lT$, $m \to mT$, $n \to nR$, $p \to pR$, have been made. It follows from the Eqs. (3.19) that the spin coefficients corresponding to the tetrad (3.24) have the form

$$\tilde{\rho} = (i/\sqrt{2})(1+f^2)^{-1}(RZ)^{-1}W_2, \tag{3.25a}$$

$$\tilde{\tau} = (1/4\sqrt{2})(1+f^2)(TZ)^{-1}W_1, \tag{3.25b}$$

$$\tilde{\epsilon} = (1/2\sqrt{2})R^{-1} \times \{Z^{-1}[mn' - lp' + i(1+f^2)W_2] - R^{-1}R'\}, \tag{3.25c}$$

$$\tilde{\alpha} = (1/2\sqrt{2})T^{-1} \times \{Z^{-1}[\tfrac{1}{4}(1+f^2)W_1 + i(pl' - mn')] + iT^{-1}T'\}, \tag{3.25d}$$

where W_1 and W_2 are Wronskian determinants defined by

$$W_1 := np' - pn', \tag{3.25e}$$

$$W_2 := lm' - ml'. \tag{3.25f}$$

In a similar manner to that used to obtain the tetrad (3.24) we find that the metric (3.21a) and self-dual Maxwell field (3.21b) may be expressed as

$$ds^2 = e^{2\psi}\{4(1+f^2)^{-2}[f(RZ)^{-2}(l\,du + m\,dv)^2 + (1-f^2)Z^{-1}dw(l\,du + m\,dv) - fR^2\,dw^2] - (TZ)^{-2}(n\,du + p\,dv)^2 - T^2\,dx^2\}, \tag{3.26a}$$

$$\overset{+}{F} = BZ^{-1}[2(1+f^2)^{-1}\,dw \wedge (l\,du + m\,dv) - i\,dx \wedge (n\,du + p\,dv)]. \tag{3.26b}$$

The metric (3.26a) is identical with the corresponding expression given in Eq. (1.8) of DKM 1 after the substitutions $e^{\psi} \to S$, $R \to 1/W$, $T \to 1/X$, $l \to gl$, $m \to gm$, $n \to n/g$, $p \to p/g$, $u \to gu$, $v \to gv$ have been made where

$$g := ((1+f^2)/2)^{1/2}. \tag{3.27}$$

The fact remarked at the beginning of this section that one may set $f = -e$ (where $e^2 = 1$) by the transformation (3.2) when $f \neq 0$ is reflected in the canonical forms (3.26) by the fact that the coordinate transformation

$$u \to |f|^{-1/4}gu - \frac{1}{2}f^{-1}|f|^{1/2}(1-f^2)\int pR^2\,dw, \tag{3.28a}$$

$$v \to |f|^{-1/4}gv + \frac{1}{2}f^{-1}|f|^{1/2}(1-f^2)\int nR^2\,dw, \tag{3.28b}$$

$$w \to |f|^{1/2}w, \quad x \to x, \tag{3.28c}$$

has the effect of transforming (3.26a) and (3.26b) into expressions of the same form, when the functions l, m, n, and p are suitably rescaled, with $f = -e$. The same result applies *mutatis mutandis* to the tetrad (3.24) provided an appropriate tetrad rotation is made. The latter result is that of Theorem 1 of DKM 1.

It remains to be shown that in the expressions (3.26) for the metric and Maxwell field or equivalently in the tetrad (3.24) it is always possible to choose l and n to be constant functions by means of form preserving coordinate transformations. In order to achieve this it is necessary to consider the following possibilities:

Case A: $\tilde{\rho}\tilde{\tau} \neq 0$. Because of the Eqs. (3.15) this inequality is equivalent to

$$(\rho - \bar{\rho})(\tau + \bar{\tau}) \neq 0, \tag{3.29}$$

a form which is invariant under the conformal transformation (3.10) defined by (3.12) which induces the transformations

$$\tilde{\rho} = e^{-\psi}(\rho - D\psi), \quad \tilde{\tau} = e^{-\psi}(\tau - \delta\psi). \tag{3.30}$$

It follows from the Eqs. (3.25) that the defining condition for Case A is equivalent to the inequality $(lm' - ml')(np' - pn') \neq 0$, on some open set which implies the inequality $lmnp \neq 0$. The last inequality permits us modulo a form preserving coordinate transformation and a redefinition of the functions m, p, R, and T to rewrite the tetrad (3.24) as follows:

$$\tilde{\theta}^1 = (\sqrt{2}/(1+f^2))[f(RZ)^{-1}(e_1\,du + m\,dv) + R\,dw], \tag{3.31a}$$

$$\tilde{\theta}^2 = (\sqrt{2}(1+f^2))[(RZ)^{-1}(e_1\,du + m\,dv) - fR\,dw], \tag{3.31b}$$

$$\tilde{\theta}^3 = (1/\sqrt{2})[(TZ)^{-1}(e_2\,du + p\,dv) + iT\,dx], \tag{3.31c}$$

where e_1 and e_2 are constants satisfying $e_1e_2 \neq 0$ and where

$$Z = e_1p - e_2m. \tag{3.31d}$$

We note that our Case A metric canonical form reduces to Carter's[69] [a] canonical form when $f = 1$ and the conformal factor is suitably restricted.

Case B_-: $\tilde{\rho} = 0$, $\tilde{\tau} \neq 0$. By the Eqs. (3.15) these conditions are equivalent to

$$\rho - \bar{\rho} = 0, \quad \tau + \bar{\tau} \neq 0 \tag{3.32}$$

which by the Eqs. (3.30) are invariant under the conformal transformation defined by (3.12). Because of the Eqs. (3.25) the conditions (3.32) can be written as

$$lm' - ml' = 0, \tag{3.33a}$$

$$np' - pn' \neq 0, \tag{3.33b}$$

the last of which implies $np \neq 0$. Since $Z \neq 0$, not both l and m may vanish. Assuming $m \neq 0$ Eq. (3.33a) implies $l = cm$, where c is constant. When we substitute for l in (3.24) we obtain, after a suitable coordinate transformation and redefinitions, the tetrad (3.31) with $e_1 = 0$. In the case $l \neq 0$, Eq. (3.33a) implies $m = \hat{c}l$, where $\hat{c}$ is some constant. An argument identical to that given above again yields a tetrad of the form (3.31) with $e_1 = 0$.

We thus conclude that in the Case B_- there exists a coordinate system in which the tetrad has the form (3.31) with $e_1 = 0$.

Case B_+: $\tilde{\tau} = 0$, $\tilde{\rho} \neq 0$. These conditions are equivalent to

$$\tau+\bar{\tau}=0,\ \rho-\bar{\rho}\neq 0, \tag{3.34}$$

because of the Eqs. (3.15) and as such are invariant under the conformal transformation (3.10) where ψ satisfies (3.12). By an argument similar to that employed in the Case B_- we are able to conclude that in the Case B_+ there exists a coordinate system in which the tetrad has the form (3.37) with $e_2=0$. We note that the metric canonical forms in our Case B_- and B_+ reduce to Carter's [b (ϵ)] canonical forms when $f=1$ and the conformal factor is suitably restricted.

Case C: $\bar{\rho}=\bar{\tau}=0$. Because of the Eqs. (3.15) these conditions may also be written as

$$\rho-\bar{\rho}=0,\quad \tau+\bar{\tau}=0 \tag{3.35}$$

in which form they are invariant under the conformal transformation defined by Eq. (3.12). In view of the expressions (3.25) for the spin coefficients the above conditions have the form

$$lm'-ml'=0, \tag{3.36a}$$

$$np'-pn'=0. \tag{3.36b}$$

The inequality (3.41) implies $lp\neq 0$ or $mn\neq 0$. We need only consider the first possibility since consideration of the second leads to an equivalent result. The Eqs. (3.36) thus imply that $m=c_1 l$, $n=c_2 p$, where c_1 and c_2 are constants satisfying the inequality $1-c_1c_2\neq 0$. When the above values of m and n are substituted in the tetrad (3.24) and a suitable coordinate transformation is made we obtain the tetrad (3.31), where $e_2=0$, $m=0$, and $p(w)=1$. It should be noted in this case that other choices of coordinates are possible, some of which will be used in the sequel.

The above results for the Cases A, B_-, B_+, and C are combined in the following statement: for every solution in $\mathfrak{D}$ there exists a local coordinate system in which the canonical null tetrad for the class has up to a conformal transformation the form (3.31).

An alternate form of the tetrad required for the proof of Theorem 1 and useful in the sequel may be obtained by the transformations $u\to gu$, $v\to gv$, $m\to gm$, $p\to p/g$, $R\to 1/U^{1/2}$, $T\to 1/V^{1/2}$, where g is defined by (3.27), yielding

$$\tilde{\theta}^1=(1/\sqrt{2})[fU^{1/2}Z^{-1}(\epsilon_1\,du+m\,dv)+g^{-2}U^{-1/2}\,dw], \tag{3.37a}$$

$$\tilde{\theta}^2=(1/\sqrt{2})[U^{1/2}Z^{-1}(\epsilon_1\,du+m\,dv)-fg^{-2}U^{-1/2}\,dw], \tag{3.37b}$$

$$\tilde{\theta}^3=(1/\sqrt{2})[V^{1/2}Z^{-1}(\epsilon_2\,du+p\,dv)+iV^{-1/2}\,dx], \tag{3.37c}$$

where

$$\epsilon_1:=g^{-1}e_1,\quad \epsilon_2:=ge_2,\quad Z:=\epsilon_1 p-\epsilon_2 m. \tag{3.37d}$$

A canonical tetrad for the class $\mathfrak{D}$ is obtained from the above tetrad by the conformal transformation

$$\theta^a=e^{-\psi}\tilde{\theta}^a\quad (a=1,2,3,4), \tag{3.38}$$

which is the inverse of the transformation (3.10) defined by (3.12) and whereby Eq. (3.20) ψ may be expressed as

$$\psi(w,x)=\ln|T(w,x)|-\tfrac{1}{2}\ln|\epsilon_1 p(w)-\epsilon_2 m(x)|. \tag{3.39}$$

The expressions for ds^2 and $\overset{+}{F}$ given in Theorem 1 follow from the Eqs. (3.7), (3.8), (3.9) where the values for the θ^a are given by (3.38), (3.37), and (3.39).

We complete this section by giving expressions needed in the sequel for the NP operators and curvature components for some of the tetrads considered in this section. For the tetrad (3.37) the corresponding NP operators are

$$\tilde{D}\phi=(1/\sqrt{2})[fg^{-1}U^{-1/2}(p\phi_u-\epsilon_2\phi_v)+U^{1/2}\phi_w], \tag{3.40a}$$

$$\tilde{\Delta}\phi=(1/\sqrt{2})[g^{-1}U^{-1/2}(p\phi_u-\epsilon_2\phi_v)-fU^{1/2}\phi_w], \tag{3.40b}$$

$$\tilde{\delta}\phi=(1/\sqrt{2})[V^{-1/2}(\epsilon_1\phi_u-m\phi_v)-iV^{1/2}\phi_x], \tag{3.40c}$$

For the tetrad (3.38) the nonzero curvature components are

$$\Psi_1=\tfrac{1}{8}e^{2\psi}(UV)^{1/2}[\epsilon_2(Z^{-1}p_w)_w-\epsilon_1(Z^{-1}m_x)_x], \tag{3.41a}$$

$$\begin{aligned}\Psi_2=\tfrac{1}{12}\epsilon^{2\psi}\{&f[U_{ww}-3\epsilon_1 Z^{-1}U_w(p_w-im_x)-2\epsilon_1 UZ^{-2}(Zp_{ww}-2\epsilon_1 p_w^2+\epsilon_1 m_x^2\\&+3i\epsilon_1 m_x p_w)]+V_{xx}+3\epsilon_2 Z^{-1}V_x(m_x+ip_w)+2\epsilon_2 VZ^{-2}(Zm_{xx}+2\epsilon_2 m_x^2-\epsilon_2 p_w^2+3i\epsilon_2 m_x p_w)\},\end{aligned} \tag{3.41b}$$

$$\Psi_3=f\Psi_1, \tag{3.41c}$$

$$\Phi_{00}=\tfrac{1}{8}Ue^{2\psi}[4T^{-1}T_{ww}-2\epsilon_1(Z^{-1}p_w)_w+\epsilon_1^2Z^{-2}(m_x^2-p_w^2)], \tag{3.41d}$$

$$\Phi_{02}=-\tfrac{1}{8}Ve^{2\psi}[4T^{-1}T_{xx}+2\epsilon_2(Z^{-1}m_x)_x+\epsilon_2^2Z^{-2}(p_w^2-m_x^2)], \tag{3.41e}$$

$$\Phi_{01}=\tfrac{1}{8}(UV)^{1/2}e^{2\psi}\{Z^{-1}[\epsilon_2 p_{ww}+\epsilon_1 m_{xx}-2T^{-1}(\epsilon_2 p_w T_w+\epsilon_1 m_x T_x)]-2iT^{-1}[2T_{wx}+Z^{-1}(m_xT_w-p_wT_x)]\}, \tag{3.41f}$$

$$\begin{aligned}\Phi_{11}=\tfrac{1}{16}e^{2\psi}\{&f[2U_{ww}-4U_w(T^{-1}T_w+\epsilon_1 Z^{-1}p_w)+U(8\epsilon_1(TZ)^{-1}p_wT_w-4T^{-1}T_{ww}+3\epsilon_1^2Z^{-2}(p_w^2+m_x^2)\\&-2\epsilon_1 Z^{-1}p_{ww})]-2V_{xx}+4V_x(T^{-1}T_x-\epsilon_2 Z^{-1}m_x)+V(8\epsilon_2(TZ)^{-1}m_xT_x\\&+4T^{-1}T_{xx}-3\epsilon_2^2Z^{-2}(p_w^2+m_x^2)-2\epsilon_2 Z^{-1}m_{xx})\},\end{aligned} \tag{3.41g}$$

$$\Phi_{21}=f\Phi_{01}, \tag{3.41h}$$

$$\Phi_{22}=f^2\Phi_{00}, \tag{3.41i}$$

$$\begin{aligned}6\Lambda=\tfrac{1}{8}e^{2\psi}\{&f[-2U_{ww}+12T^{-1}U_wT_w+U(12T^{-1}T_{ww}-24T^{-2}T_w^2+\epsilon_1^2Z^{-2}(p_w^2+m_x^2)-2\epsilon_1 Z^{-1}p_{ww})]\\&-2V_{xx}+12T^{-1}V_xT_x+V(12T^{-1}T_{xx}-24T^{-2}T_x^2\\&+\epsilon_2^2Z^{-2}(p_w^2+m_x^2)+2\epsilon_2 Z^{-1}m_{xx})\}.\end{aligned} \tag{3.41j}$$

4. PROOF OF THEOREM 2

The general idea of the proof is first to demonstrate by integration of the Einstein–Maxwell field equations (2.1) that the solutions in the subclasses A^*, B^0_-, B^0_+, C^*, and C^{00} enumerated in Sec. 2 exhaust the class $\mathfrak{D}$. Then adopting an idea of Carter[70] an appropriate parametrized coordinate transformation is performed on the solutions of the form A^* to obtain the solution denoted by $\widetilde{A}^*$ in Theorem 2. By an analyticity argument we conclude that the form $\widetilde{A}^*$ is a solution even for parameter values for which the defining coordinate transformation is singular. The proof of the theorem is completed by remarking (as has been done already in Sec. 2) that the solutions A^*, B^0_-, B^0_+, C^*, and C^{00} are special cases of the solution $\widetilde{A}^*$.

We begin the proof with the integration of the Einstein–Maxwell field equation (2.1) for the canonical forms established in Sec. 3. The first step in the integration is to impose the Petrov type D condition

$$\Psi_1 = \Psi_3 = 0 \tag{4.1}$$

on the tetrad (3.37). This condition is equivalent[71] to the existence of a Maxwell field of the form (3.8) the integrability condition for which is given by Eq. (3.14). Because of the Eqs. (3.41a) and (3.41c) the condition (4.1) may be expressed as

$$\epsilon_2[(\epsilon_1 p - \epsilon_2 m)^{-1} p_w]_w = \epsilon_1[(\epsilon_1 p - \epsilon_2 m)^{-1} m_x]_x. \tag{4.2}$$

In order to exploit this condition and the field equations

$$\Phi_{00} = \Phi_{01} = \Phi_{02} = 0, \tag{4.3}$$

to determine the tetrad functions m and p and the function T appearing in the conformal transformation (3.39) it is convenient to consider the same possibilities as those used in the proof of Theorem 1.

Case A: $\epsilon_1\epsilon_2 \neq 0$: Since ϵ_1 and ϵ_2 can be given any nonzero values by form preserving coordinate transformations there is no loss of generality in assuming

$$\epsilon_1 = \epsilon_2 = 1, \tag{4.4}$$

in which case Eq. (4.2) can be rewritten as

$$(p - m)p_{ww} - p_w^2 = (p - m)m_{xx} + m_x^2, \tag{4.5}$$

where by (3.34)

$$p_w m_x \neq 0. \tag{4.6}$$

Repeated differentiation of the Eq. (4.5) yields

$$p_{www}/p_w = -m_{xxx}/m_x = c, \tag{4.7}$$

where c is a separation constant. There are three possibilities to consider.

Subcase A1: $c = 0$. Integration of the differential equations (4.7) yields

$$p(w) = c_1 w^2 + c_2 w + c_3, \tag{4.8a}$$

$$m(w) = d_1 x^2 + d_2 x + d_3, \tag{4.8b}$$

where c_i and d_i, $i = 1,2,3$ are constant. Substitution of these expressions into (4.5) gives

$$d_1 = -c_1,\ c_2^2 + d_2^2 = 4c_1(c_3 - d_3), \tag{4.8c}$$

which together with (4.6) imply that $c_1 \neq 0$. When the above expressions for p and m are introduced into the tetrad (3.31), we may set $c_1 = -d_1 = 1$, $c_2 = d_2 = c_3 = d_3 = 0$ by an appropriate coordinate transformation.

We conclude that in the Subcase A1 a coordinate system exists in which the parameters ϵ_1 and ϵ_2 and the functions p and m in the tetrad (3.37) are given as follows:

$$\epsilon_1 = \epsilon_2 = 1,\quad p(w) = w^2,\quad m(x) = -x^2. \tag{4.9}$$

Subcase A2: $c = \omega^2 \neq 0$. Integration of the differential equations (4.7) yields

$$p(w) = c_1 \sinh(\omega w) + c_2 \cosh(\omega w) + c_3, \tag{4.10a}$$

$$m(x) = d_1 \sin(\omega x) + d_2 \cos(\omega x) + d_3, \tag{4.10b}$$

where c_i and d_i, $i = 1,2,3$ are constant. Because of Eq. (4.5) these constants must satisfy

$$d_3 = c_3,\quad c_2^2 - c_1^2 = d_1^2 + d_2^2. \tag{4.11}$$

By means of the translations

$$w \to w + w_0,\quad x \to x + x_0 \tag{4.12}$$

we may set $c_1 = d_1 = 0$, which implies $c_2{}^2 = d_2{}^2$.

There are two possibilities to consider.

(a): $c_2 = d_2$. In this case the functions p and m have the form

$$p(w) = 2c_2 \sinh^2(\tfrac{1}{2}\omega w) + c_4, \tag{4.13a}$$

$$m(x) = -2c_2 \sin^2(\tfrac{1}{2}\omega x) + c_4, \tag{4.13b}$$

where $c_4 = c_2 + c_3$. When these expressions are substituted into the tetrad (3.37) and the coordinate transformation

$$u \to \tfrac{1}{2}\omega[(2c_2 - c_4)u - c_4 v],\quad v \to \tfrac{1}{2}\epsilon_1\omega(v + u) \tag{4.14a}$$

$$w \to 2\omega^{-1}\tanh^{-1} w,\quad x \to 2\omega^{-1}\tan^{-1} x \tag{4.14b}$$

is made one obtains a tetrad of the form (3.37) in which p and m have the form (4.9).

(b): $c_2 = -d_2$. In this case p and m may be expressed as

$$p(w) = 2c_2 \sinh^2(\tfrac{1}{2}\omega w) + c_4, \tag{4.15a}$$

$$m(x) = -2c_2 \cos^2(\tfrac{1}{2}\omega x) + c_4. \tag{4.15b}$$

When these expressions are substituted into the tetrad (3.37) and the coordinate transformation (4.14) is made where the second member in (4.14b) is replaced by $x \to (2/\omega)\mathrm{arccot}(-x)$, one again obtains a tetrad of the form (3.37) in which p and m have the form (4.19).

We conclude that in the Subcase A2 a coordinate system exists in which the parameters ϵ_1 and ϵ_2 and the functions p and m in the tetrad (3.37) are given by Eq. (4.9).

Subcase A3: $c = -\omega^2 \neq 0$. Integration of the differential equations (4.7) and substitution of the solutions into Eq. (4.5) implies that p and m have the form

$$p(w) = d_1 \sin(\omega w) + d_2 \cos(\omega w) + c_3, \tag{4.16a}$$

$$m(x) = c_1 \sinh(\omega x) + c_2 \cosh(\omega x) + c_3, \tag{4.16b}$$

where c_i, $i = 1,2,3$ and d_i, $i = 1,2$ are constants satisfying

$$c_2^2 - c_1^2 = d_1^2 + d_2^2. \tag{4.16c}$$

We note that the function p and m above are identical to the functions m and p, respectively, given by the Eqs. (4.10). In view of this fact one arrives at the same conclusion in this subcase as in the Subcase A2 by an argument identical to that already given.

It follows from the results of the above subcases that in the Case A a coordinate system exists in which the functions

m and p appearing in the canonical tetrad (3.37) have the form (4.9). This result agrees with that previously obtained by Debever[72] in the case $f = 1$.

It remains to determine the form of the function T in this case from the field equations (4.3) which, in view of the Eqs. (3.41d), (3.41e), (3.41f), the type D condition (4.5), the explicit forms for p and m, and the values for ϵ_1 and ϵ_2 given in (4.9), take the form

$$T_{ww} = 0, \quad T_{xx} = 0, \quad wT_w - xT_x = 0, \tag{4.17a}$$

$$(w^2 + x^2)T_{wx} = xT_w + wT_x. \tag{4.17b}$$

The Eqs. (4.17a) imply that T has the form

$$T(w,x) = b_1 wx + b_4, \tag{4.18}$$

where b_1 and b_4 are arbitrary constants not both zero, which identically satisfies the Eq. (4.17b). When the expressions for m and p given by (4.9) and the above expression for T, assuming $b_4 \neq 0$, are introduced into the tetrad (3.39) where ψ is given by (3.38) we may set $b_4 = 1$, by an appropriate rescaling of the coordinates. If in addition $b_1 \neq 0$, we may set $b_1 = 1$, by a further rescaling of the coordinates. We note that there is no loss of generality in the assumption $b_4 \neq 0$, since the coordinate transformation $u \to -v$, $v \to -u$, $w \to 1/w$, $x \to 1/x$, which preserves the form of the tetrad, induces the transformation $b_1 \to b_4$, $b_4 \to b_1$.

The results of this case may be summarized as follows: In the Case A there exists a system of coordinates in which the parameters ϵ_1 and ϵ_2 and the functions p and m appearing in the canonical tetrad (3.37) are given by the Eqs. (4.9) namely

$$\epsilon_1 = \epsilon_2 = 1, \quad p(w) = w^2, \quad m(x) = -x^2, \tag{4.19a}$$

and the function T in the conformal transformation (3.39) has the form

$$T(w,x) = b_1 wx + 1, \tag{4.19b}$$

where b_1 is constant.

The form of the conformal transformation (3.39) implied by the Eq. (4.19b) is equivalent to that given earlier by Debever.[73] It's also worth noting that the metric (2.5a) with $f = 1$ and $\epsilon_1, \epsilon_2, p, m$, and T given by the Eqs. (4.19) is equivalent to that used by Plebański and Demiański[74] as the starting point of their integration procedure.

Case B_-: $\epsilon_1 = 0$, $\epsilon_2 \neq 0$. Under these assumptions the Eq. (4.2) reduces to

$$p_{ww} = 0, \tag{4.20}$$

which has the solution

$$p(w) = c_2 w + c_3, \tag{4.21}$$

where c_2 and c_3 are constants which because of the inequality (3.33b) satisfy $c_2 \neq 0$. By virtue of this inequality we may set $c_3 = 0$, by the form preserving translation $w \to w + w_0$. It thus follows that p thas the form

$$p(w) = c_2 w. \tag{4.22}$$

The functions m and T are determined by the field equations (4.3) which, because of the Eqs. (3.41d), (3.41e), (3.41f), (4.22), reduce to

$$T_w = 0, \tag{4.23a}$$

$$4T^{-1}T_{xx} - (m^{-1}m_x)^2 + c_2^2 m^{-2} - 2(m^{-1}m_x)_x = 0. \tag{4.23b}$$

In order to obtain suitable explicit solutions to the above equations we use that the fact that nonzero function m appearing in the tetrad 1-form (3.37c), namely,

$$\tilde{\theta}^3 = (1/\sqrt{2})[-\epsilon_2^{-1} V^{1/2}(x)(m(x))^{-1}(\epsilon_2\, du + p(w)dv) + iV^{-1/2}(x)dx], \tag{4.24}$$

is arbitrary in the sense that independently of its form it may be transformed to a constant function by an appropriate form preserving coordinate transformation of the form $x \to k(x)$. We use this freedom to choose m to be a solution of the differential equation

$$2(m^{-1}m_x)_x + (m^{-1}m_x)^2 - c_2^2 m^{-2} = 0. \tag{4.25}$$

The general solution of this equation is

$$m(x) = d_1 x^2 + d_2 x + d_3, \tag{4.26a}$$

where the d_i, $i = 1,2,3$ are constants satisfying

$$d_2^2 - 4d_1 d_3 = -c_2^2. \tag{4.26b}$$

This equation and the inequality $c_2 \neq 0$ imply that $d_1 d_3 \neq 0$, from which it follows that we may set $d_2 = 0$, by the translation $x \to x + x_0$. Thus m has the form

$$m(x) = d_1 x^2 + \tfrac{1}{4} d_1^{-1} c_2^2, \tag{4.27}$$

where (4.26) has been used. When this form of m and p given by Eq. (4.22) are substituted into the tetrad 1-form (4.24) we may set $d_1 = 1$, by rescaling the coordinate u. It follows that in the Case B_- a coordinate system exists in which the functions p and m in the tetrad (3.57) have the form

$$p(w) = 2kw, \quad m(x) = x^2 + k^2, \tag{4.28}$$

where $k = \tfrac{1}{2}c_2$ is constant.

It remains to determine the function T which satisfies the Eq. (4.23a) and by virtue of (4.23b) and (4.25a) the equation $T_{xx} = 0$.

It follows from these equations that T must have the form

$$T(w,x) = b_3 x + b_4, \tag{4.29}$$

where b_3 and b_4 are constants not both zero. The Eqs. (4.28) and (4.29) imply that the function ψ given by Eq. (3.39a) is independent of w. It thus follows from the first transformation in (3.30) and the operator (3.40a) that in the Case B_- the spin coefficient[75]

$$\rho = 0. \tag{4.30}$$

When the functions p and m given by (4.28) and the function T given by (4.29) are substituted into the tetrad (3.38) where ψ is given by (3.39) we may set $b_3 = 0$, if $b_4 \neq 0$, by an appropriate bilinear transformation $x \to (ax + b)/(cx + d)$. We may further set $b_4 = 1$, by rescaling the coordinates u and v. There is no loss in generality in assuming $b_4 \neq 0$ since the inversion $x \to 1/x$, which preserves the form of the tetrad after a suitable rescaling of u and v, induces the transformation $b_3 \to b_4$, $b_4 \to b_3$.

The above results may be summarized as follows: In the Case B_- there exists a system of coordinates in which the parameters ϵ_1 and ϵ_2 and the functions p and m appearing in the canonical tetrad (3.58) are given by

$$\epsilon_1 = 0, \quad \epsilon_2 = -1, \quad p(w) = 2kw, \quad m(x) = x^2 + k^2, \tag{4.31a}$$

where k is constant and the function T in the conformal transformation (3.38b) has the form

$$T(w,x) = 1. \tag{4.31b}$$

Case B_+: $\epsilon_2 = 0$, $\epsilon_1 \neq 0$. By arguments similar to those employed in the Case B_- it may be shown that [76]

$$\tau = 0, \tag{4.32}$$

and that there exists a system of coordinates in which the parameters ϵ_1 and ϵ_2 and the functions p and m appearing in the conical tetrad (3.38) are given by

$$\epsilon_1 = 1, \quad \epsilon_2 = 0, \quad p(w) = x^2 + l^2, \quad m(x) = 2lx, \tag{4.33a}$$

where l is constant and the function T in the conformal transformation (3.59b) has the form

$$T(w,x) = 1. \tag{4.33b}$$

Case C: $\epsilon_1 = 1$, $\epsilon_2 = 0$, $p(w) = 1$, $m(x) = 0$. In this case the type D condition is satisfied identically and the field equations (4.3) reduce to

$$T_{ww} = T_{wx} = T_{xx} = 0, \tag{4.34}$$

which has the general solution

$$T(w,x) = b_2 w + b_3 x + b_4, \tag{4.35}$$

where b_i, $i = 2,3,4$ are constant. The translations (4.12) induce the transformations

$$b_2 \to b_2, \quad b_3 \to b_3, \quad b_4 \to b_4 + w_0 b_2 + x_0 b_3, \tag{4.36}$$

which give rise to the following subclassification[77]:

Subcase C^*: $b_2 b_3 \neq 0$. It follows from (4.36) that we may set $b_4 = 0$ which implies by (4.35) that

$$T(w,x) = b_2 w + b_3 x. \tag{4.37}$$

The transformation formulas (3.30) by virtue of the above form of T imply that in the Subcase C*

$$\rho = \bar{\rho} \neq 0, \quad \tau = -\bar{\tau} \neq 0, \tag{4.38}$$

which equations provide an invariant characterization of this subcase. When the function T given by (4.37) is introduced in the tetrad (3.38) via the Eq. (3.39b) we may set $b_2 = b_3 = 1$ by rescaling the coordinates. It thus follows that we have

$$T(w,x) = w + x. \tag{4.39}$$

Subcase C_-: $b_2 = 0$, $b_3 \neq 0$. As in the previous subcase we may set $b_4 = 0$ which implies that

$$T(w,x) = b_3 x, \tag{4.40}$$

where we may further set $b_3 = 1$, by a coordinate rescaling. By an argument similar to that in the Subcase C^* we find that in the Subcase C_-

$$\rho = 0, \quad \tau = -\bar{\tau} \neq 0. \tag{4.41}$$

Subcase C_+: $b_2 \neq 0$, $b_3 = 0$. In this subcase we may also set $b_4 = 0$ yielding

$$T(w,x) = b_2 w, \tag{4.42}$$

where in addition we may set $b_2 = 1$ by a coordinate rescaling. By an argument similar to that given previously we have in the Subcase C_+

$$\rho = \rho \neq 0, \quad \tau = 0. \tag{4.43}$$

Subcase C^{00}: $b_2 = b_3 = 0$, $b_4 \neq 0$. It follows that

$$T(w,x) = b_4, \tag{4.44a}$$

where we may set $b_4 = 1$ by rescaling. It also follows that in the subcase C^{00} we have

$$\rho = \tau = 0. \tag{4.44b}$$

The above results may be summarized as follows: In the Case C there exists a system of coordinates in which the parameters ϵ_1 and ϵ_2 and the functions p and m in the canonical tetrad (3.58) are given by

$$\epsilon_1 = 1, \quad \epsilon_2 = 0, \quad p(w) = 1, \quad m(x) = 0. \tag{4.45a}$$

The function T in the conformal transformation (3.59b) has the following forms:

$$\text{Subcase C*:} \quad T(w,x) = w + x, \tag{4.45b}$$

$$\text{Subcase C}_-: \quad T(w,x) = x, \tag{4.45c}$$

$$\text{Subcase C}_+: \quad T(w,x) = w, \tag{4.45d}$$

$$\text{Subcase C}^{00}: \quad T(w,x) = 1. \tag{4.45e}$$

The next step in the integration involves the solution of Maxwell's Eqs. (2.1b) which may be written as

$$d\overset{+}{F} = 0, \tag{4.46a}$$

where by the Eqs. (3.8) and (3.9) the self-dual Maxwell field

$$\overset{+}{F} = \widetilde{B}\widetilde{Z}^2 \tag{4.46b}$$

is expressed in terms of the conformally related tetrad defined by the Eqs. (3.10) and (3.12). Since

$$d\widetilde{Z}^2 = \tilde{\theta} \wedge \widetilde{Z}^2, \tag{4.47}$$

the Eq. (4.46a) for an $\overset{+}{F}$ of the form (4.46b) is equivalent to

$$d\widetilde{B} = -\widetilde{B}\tilde{\theta}, \tag{4.48}$$

which implies by virtue of (3.13) that $\widetilde{B}$ has the form

$$\widetilde{B} = B_0 e^{i\mathscr{B}}, \tag{4.49a}$$

where B_0 is a complex constant and $\mathscr{B}$ is a real-valued function satisfying

$$d\mathscr{B} = i\tilde{\theta}. \tag{4.49b}$$

Because of the Eqs. (3.3), (3.5), (3.25), and (3.37), the last equation may be written as

$$d\mathscr{B} = Z^{-1}(\epsilon_1 m_x \, dw + \epsilon_2 p_w \, dx), \tag{4.50}$$

from which it follows that

$$\mathscr{B}_u = \mathscr{B}_v = 0. \tag{4.51}$$

In order to complete the integration of Eq. (4.50) it is again necessary to treat separately the cases delineated in the proof of Theorem 1 and used earlier in this section.

Case A: Because of the Eqs. (4.19a) the Eq. (4.50) reduces to

$$d\mathscr{B} = 2(w^2 + x^2)^{-1}(w \, dx - x \, dw), \tag{4.52}$$

which in view of (4.51) has the general solution

$$\mathscr{B} = i[\ln(w - ix) - \ln(w + ix)] + b_0, \tag{4.53}$$

where b_0 is a real constant. It follows from this equation and from Eq. (4.49a) that in the Case A the $\widetilde{B}$ appearing in the self-dual Maxwell field (4.46b) has the form[78]

$$\widetilde{B} = B_0(w + ix)/(w - ix), \qquad (4.54)$$

where B_0 is a complex constant.

Case B_-: On account of the Eqs. (4.31a) the Eq. (4.50) may be written as

$$d\mathscr{B} = -2k(x^2 + k^2)^{-1}\,dx, \qquad (4.55)$$

which has the general solution

$$\mathscr{B} = i[\ln(x - ik) - \ln(x + ik)] + b_0, \qquad (4.56)$$

where b_0 is a real constant. This equation implies that in the Case B_- the function $\widetilde{B}$ appearing in the self-dual Maxwell field (4.46b) has the form

$$\widetilde{B} = B_0(x + ik)/(x - ik), \qquad (4.57)$$

where B_0 is a complex constant.

Case B_+: The Eqs. (4.33a) imply that the Eq. (4.50) may be expressed as

$$d\mathscr{B} = 2l(w^2 + l^2)^{-1}\,dw, \qquad (4.58)$$

which has the general solution

$$\mathscr{B} = i[\ln(w + il) - \ln(w - il)] + b_0,$$

where b_0 is constant. It results from this equation that in the Case B_+ the $\widetilde{B}$ appearing in the self-dual Maxwell field (4.46b) has the form

$$\widetilde{B} = B_0(w - il)/(w + il), \qquad (4.59)$$

where B_0 is a complex constant.

Case C: By the Eq. (4.45a) the Eq.(4.50) reduces to

$$d\mathscr{B} = 0, \qquad (4.60)$$

which implies $\mathscr{B}$ is constant. It thus follows that in the Case C the function $\widetilde{B}$ appearing in the self-dual Maxwell field (4.63b) has the form

$$\widetilde{B} = B_0, \qquad (4.61)$$

where B_0 is a complex constant.

The final step in the integration procedure is the solution of the remaining field equations which are

$$\Phi_{11} = B\overline{B}, \qquad (4.62)$$

$$\Lambda = \tfrac{1}{6}\lambda, \qquad (4.63)$$

where B (which may be zero) is the function appearing in the self-dual Maxwell field (3.8) and λ is the cosmological constant, in order to determine the functions U and V. The study of the first of these equations requires an explicit form for the function B appearing therein which may be obtained from the Eq.(4.49a) using the conformal invariance of the Maxwell 2-form $\overset{+}{F}$ namely the fact that

$$\overset{+}{F} = BZ^2 = \widetilde{B}\widetilde{Z}^2. \qquad (4.64)$$

This equation implies that

$$B = e^{2\psi}\widetilde{B}, \qquad (4.65)$$

on account of the transformation law

$$\widetilde{Z}^2 = e^{2\psi}Z^2 \qquad (4.66)$$

induced by (3.10). If we now replace $\widetilde{B}$ in (4.65) using (4.49a) and ψ employing (3.38), the Eq. (4.62) takes the form

$$\begin{aligned} 8B_0\overline{B}_0T^3 &= f(TZU_{ww} - 2\epsilon_1 Tp_w U_w - 2ZT_w U_w + 2\epsilon_1 TUp_{ww} \\ &\quad + 4\epsilon_1 Up_w T_w) - TZV_{xx} - 2\epsilon_2 Tm_x V_x \\ &\quad + 2ZT_x V_x + 2\epsilon_2 TVm_{xx} + 4\epsilon_2 Vm_x T_x, \end{aligned} \qquad (4.67)$$

where the terms involving T_{ww} and T_{xx} in (3.41g) have been replaced using the first two of the Eqs. (4.3) and the corresponding expressions for Φ_{00} and Φ_{02} given by the Eqs. (3.41d) and (3.41e).

We now turn our attention to the Eq. (4.63) which may be written as

$$\begin{aligned} 24Z\Lambda &= f[-T^2U_{ww} + 6TT_wU_w + 4U(TT_{ww} - 3T_w{}^2)] \\ &\quad - T^2V_{xx} + 6TT_xV_x + 4V(TT_{xx} - 3T_x{}^2), \end{aligned} \qquad (4.68)$$

where the first two of the Eqs. (4.3) and the corresponding Eqs. (3.41d) and (3.41e) have been used to eliminate the terms involving the derivatives of p and m in (3.41j). It follows from this equation and the fact that there exists a coordinate system in which the functions p, m, and T have the following forms:

$$p(w) = c_1w^2 + c_2w + c_3, \qquad (4.69a)$$

$$m(x) = d_1x^2 + d_2x + d_3, \qquad (4.69b)$$

$$T(w,x) = b_1wx + b_2w + b_3x + b_4, \qquad (4.69c)$$

where c_i, d_i, $i = 1,2,3$, and b_i, $i = 1,2,3,4$ are constant, shown earlier in this section for every solution in $\mathfrak{D}$, that the functions fU and V are polynomials of the fourth degree in the variables w and x, respectively. The property that coordinates can be chosen such that all the metric functions are rational functions thereof has been previously noted and exploited by Carter,[79] Debever,[80] Plebański and Demiański,[81] and Weir and Kerr[82] in their respective integration procedures. We may establish it for the functions fU and V by taking third derivatives of the Eq. (4.68) first with respect to w and then with respect to x which procedure in view of (4.69) yields the equations

$$\frac{\partial^5(fU)}{\partial w^5} = 0, \quad \frac{\partial^5 V}{\partial x^5} = 0 \qquad (4.70)$$

from which the result follows. We thus may write

$$fU(w) = f_4w^4 + f_3w^3 + f_2w^2 + f_1w + f_0, \qquad (4.71)$$

$$V(x) = g_4x^4 + g_3x^3 + g_2x^2 + g_1x + g_0, \qquad (4.72)$$

where f_i and g_i, $i = 0,1,2,3,4$, are constant.

The integration is completed by substituting the above expressions for fU and V into the Eqs. (4.67) and (4.68) in order to determine the relations satisfied by the constants appearing therein. Again and for the last time it is necessary to treat separately the cases considered in the proof of Theorem 1 and earlier in this section.

Case A: Because of the Eqs. (4.19), (4.71), and (4.72) the Eqs. (4.67) and (4.68) imply

$$f_3 = b_1g_1, \quad g_2 = -f_2, \quad g_3 = b_1f_1, \qquad (4.73a)$$

$$f_4 - g_4 = 2b_1{}^2B_0\overline{B}_0, \quad f_0 - g_0 = 2B_0\overline{B}_0, \qquad (4.73b)$$

$$f_4 + b_1{}^2g_0 = -\tfrac{1}{3}\lambda, \quad b_1{}^2f_0 + g_4 = -\tfrac{1}{3}\lambda. \qquad (4.73c)$$

The Eqs. (4.73b) and (4.73c) form a consistent system of linear equations in the quantities f_0, f_4, g_0, and g_4 whose solution may be written as

$$f_4 = -b_1{}^2f_0 + 2(b_1{}^2B_0\overline{B}_0 - \tfrac{1}{6}\lambda), \quad g_0 = f_0 - 2B_0\overline{B}_0,$$

$$g_4 = -(b_1^2 f_0 + \tfrac{1}{3}\lambda), \tag{4.74}$$

where f_0 is an arbitrary parameter. When the Eqs. (4.73a) and (4.74) are taken into account in the Eqs. (4.77) and (4.72) and the substitution $b_1 \to a$ is made, we obtain precisely the expressions for fU and V given by the Eqs. (2.7c) and (2.7d), respectively. Furthermore that the expressions for ϵ_1, ϵ_2, m, p, T, and $\widetilde{B}$ given by the Eqs.(4.19) and (4.54) when substituted in (2.5) yield the same expressions for the metric and self-dual Maxwell field after the transformation $u \to \frac{1}{2}u$, as that which is obtained from the Eqs. (2.7). We thus conclude that the general solution in the Case A is given by a metric and Maxwell field of the form (2.5), where the functions appearing therein are given by the Eqs. (2.7).

Case B_-: A procedure similar to that employed in the preceeding case yields the relations

$$f_3 = f_4 = g_3 = 0, \quad g_0 = k^2 f_2 + k^4 \lambda - 2B_0\overline{B}_0, \tag{4.75a}$$

$$g_2 = -(f_2 + 2k^2\lambda), \quad g_4 = -\tfrac{1}{3}\lambda, \tag{4.75b}$$

where f_0, f_1, f_2, and g_1 are arbitrary parameters. The resulting expressions for fU and V have exactly the same form as those given by the Eqs. (2.8c) and (2.8d), respectively. We also remark that the expressions for ϵ_1, ϵ_2, m, p, T, and $\widetilde{B}$ given by (4.31) and (4.57) yield the same expressions for the metric and self-dual Maxwell field as the corresponding quantities given by (2.8b) and (2.8e) after the substitutions $v \to -v$ and $x \to -x$ and appropriate parameter sign changes have been made. We conclude that the general solution in the Case B_- is given by a metric and Maxwell field of the form (2.5), where the functions appearing therein are given by the Eqs. (2.8).

Case B_+: In this case we obtain the following relations between the parameters:

$$f_3 = g_3 = g_4 = 0, \quad f_0 = l^2 g_2 + l^2\lambda + 2B_0\overline{B}_0, \tag{4.76a}$$

$$f_2 = -(g_2 + 2l^2\lambda), \quad f_4 = -\tfrac{1}{3}\lambda, \tag{4.76b}$$

where f_1, g_0, g_1, and g_2 are arbitrary parameters. These relations yield expressions for fU and V which have the same form as those given by (2.9c) and (2.9d). Furthermore, one obtains the same expressions for the metric and self-dual Maxwell field by substituting in (2.5a) and (2.5b) for ϵ_1, ϵ_2, m, p, T, and B from the Eqs. (4.33) and (4.59) as one does by using the Eqs. (2.9) and making the substitution $l \to -l$. We conclude that the general solution in the Case B_+ is given by a metric and Maxwell field of the form (2.5) where the functions appearing therein are given by the Eqs. (2.9).

Case C: In this case we have a coordinate system in which the Eqs. (4.45a) hold. However, we have to consider separately the four subcases already determined corresponding to the different possible forms of T given in the Eqs. (4.45).

Subcase C^*: The relations obtained in this subcase are

$$f_4 = 2B_0\overline{B}_0, \quad g_0 = -(f_0 + \tfrac{1}{3}\lambda), \tag{4.77a}$$

$$g_1 = f_1, \quad g_2 = -f_2, \quad g_3 = f_3, \quad g_4 = -2B_0\overline{B}_0, \tag{4.77b}$$

where f_0, f_1, f_2, and f_3 are arbitrary parameters. The functions fU and V thus have the form

$$fU(w) = 2B_0\overline{B}_0 w^4 + f_3 w^3 + f_2 w^2 + f_1 w + f_0, \tag{4.78}$$

$$V(x) = -2B_0\overline{B}_0 x^4 + f_3 x^3 - f_2 x^2 + f_1 x - f_0 - \tfrac{1}{3}\lambda. \tag{4.79}$$

When one substitutes for $\epsilon_1, \epsilon_2, m, p, T, fU, V$, and B from the Eqs. (4.45a), (4.45b), (4.78), (4.79), and (4.61) in the expressions (2.5) for the metric and Maxwell field and performs the substitutions $u \to -\sqrt{2}v$, $v = -u/\sqrt{2}$, $x \to 1/x$, and $B_0 \to -B_0$, one obtains the same solution as that which arises from the Eqs. (2.10). We conclude that the general solution in the Case C^* is given by a metric and Maxwell field of the form (2.5) where the functions appearing therein are given by the Eqs. (2.10).

Subcase C_-: We obtain the following relations between the parameters:

$$f_3 = f_4 = g_1 = 0, \quad g_0 = -\tfrac{1}{3}\lambda, \quad g_2 = -f_2,$$
$$g_4 = -2B_0\overline{B}_0, \tag{4.80}$$

where f_0, f_1, f_2, and g_3 are arbitrary parameters. It follows that the functions fU and V have the form

$$fU(w) = f_2 w^2 + f_1 w + f_0, \tag{4.81}$$

$$V(x) = -2B_0\overline{B}_0 x^4 + g_3 x^3 - f_2 x^2 - \tfrac{1}{3}\lambda. \tag{4.82}$$

When substitutions are made for ϵ_1, ϵ_2, m, p, T, fU, V, and B from the Eqs. (4.45a), (4.45c), (4.81), (4.82), and (4.61) in the expressions (2.5) for the metric and the Maxwell field and the substitutions $u \to -v$, $v \to -u$, $x \to 1/x$ are made one obtains the solution B^0_- given by the Eqs. (2.8) with $k = 0$. We conclude that the solution C^0_- is a limiting case of the solution B^0_- when $k = 0$, a fact previously noted by Carter in the case $f \neq 0$.

Subcase C_+: The relations in this subcase are

$$f_1 = g_3 = g_4 = 0, \quad f_0 = -\tfrac{1}{3}\lambda, \quad f_2 = -g_2,$$
$$f_4 = 2B_0\overline{B}_0, \tag{4.83}$$

where g_0, g_1, and g_0 are arbitrary parameters. It follows that the functions fU and V have the form

$$fU(w) = 2B_0\overline{B}_0 w^4 + f_3 w^3 - g_2 w^2 - \tfrac{1}{3}\lambda, \tag{4.84}$$

$$V(x) = g_2 x^2 + g_1 x + g_0. \tag{4.85}$$

The solution which results from the substitution of the above expressions in the Eqs. (2.5) and the use of the Eqs. (4.45a), (4.45d), and (4.61) followed by the substitutions $u \to -u$, $w \to 1/w, f_3 \to f_1$ is the solution that arises from the Eqs. (2.5) and (2.9) when $l = 0$. We conclude that the solution C^0_+ is a limiting case of the solution B^0_+ when $l = 0$.

Subcase C^{00}: We have the relations

$$f_3 = f_4 = g_3 = g_4 = 0, \tag{4.86}$$

$$f_2 = 2B_0\overline{B}_0 - \lambda, \quad g_2 = -(2B_0\overline{B}_0 + \lambda), \tag{4.87}$$

where f_0, f_1, g_0, and g_1 are arbitrary parameters. It follows that fU and V have the form of (2.11c) and (2.11d), respectively. Moreover the expressions for ϵ_1, ϵ_2, m, p, T, and B given by the Eqs. (4.45a), (4.45e), and (4.61) agree with the corresponding expressions (2.11) except for the sign of B_0. We thus conclude that the general solution in the Subcase C^{00} is given by a metric and Maxwell field of the form (2.5) where the functions appearing therein are given by the Eqs. (2.11).

The classes of solutions enumerated above exhaust the solutions in the class $\mathfrak{D}$.

The solution $\widetilde{A}^*$ of Theorem 2 is obtained from the solution A^* defined by the Eqs. (2.7) by the coordinate transformation

$$u = bc^{-1}[\tilde{u} + \sec\gamma\csc\gamma(l^2\cos^2\gamma - k^2\sin^2\gamma)\tilde{v}], \tag{4.88a}$$

$$v = bc^{-1}\sec\gamma\csc\gamma\tilde{v}, \tag{4.88b}$$

$$w = b(c\tilde{w}\cos\gamma + k), \tag{4.88c}$$

$$x = b^{-1}(c\tilde{x}\sin\gamma + l), \tag{4.88d}$$

where b, c, k, l, and γ are arbitrary parameters restricted for the present by the requirement that the transformation be nonsingular. The relations between the new parameters and the old are given by

$$\tilde{f}_4 = b^4[a^2(2B_0\overline{B}_0 - f_0) - \tfrac{1}{3}\lambda], \tag{4.89a}$$

$$\tilde{f}_3 = ab^3g_1 + 4k\tilde{f}_4, \tag{4.89b}$$

$$\tilde{f}_2 = b^2f_2 + 3ab^3kg_1 + 6k^2\tilde{f}_4, \tag{4.89c}$$

$$c\tilde{f}_1\cos\gamma = bf_1 + 2b^2kf_2 + 3ab^3k^2g_1 + 4k^3\tilde{f}_4, \tag{4.89d}$$

$$c^2\tilde{f}_0\cos^2\gamma = f_0 + bkf_1 + b^2kf_2 + ab^3k^3g_1 + k^4\tilde{f}_4, \tag{4.89e}$$

$$\tilde{g}_4 = -(a^2f_0 + \tfrac{1}{3}\lambda), \tag{4.89f}$$

$$\tilde{g}_3 = abf_1 + 4l\tilde{g}_4, \tag{4.89g}$$

$$\tilde{g}_2 = -b^2f_2 + 3ablf_1 + 6l^2\tilde{g}_4, \tag{4.89h}$$

$$c\tilde{g}_1\sin\gamma = b^3g_1 - 2b^2lf_2 + 3abl^2f_1 + 4l^3\tilde{g}_4, \tag{4.89i}$$

$$c^2\tilde{g}_0\sin^2\gamma = b^4f_0 - 2b^4B_0\overline{B}_0 + b^3lg_1 - b^2l^2f_2 + abl^3f_1 + l^4\tilde{g}_4, \tag{4.89j}$$

$$\widetilde{B}_0 = b^2B_0. \tag{4.89k}$$

When one eliminates the quantities f_i and g_i from the above equations one obtains the relations (2.6h) and (2.6i) and the expression (2.6j) where the tildes have been dropped. The $\widetilde{A}^*$ metric and self-dual 2-form will clearly be a solution of the Einstein–Maxwell field equations for all values of the parameters for which the transformation (4.88) is nonsingular. Moreover, by the analyticity of the metric and Maxwell tensor components in the parameters the field equations will continue to hold even for parameter values for which the transformation (4.88) is singular. Thus the $\widetilde{A}^*$ metric and 2-form will be a solution for the range of parameter values indicated in Theorem 2.

5. NULL ORBIT SOLUTIONS

These solutions may be determined by setting $f = 0$ in each of the main special cases listed after Theorem 2, the results of which are given below.

Case A^*: The Eq. (2.7c) implies the following relations when $f = 0$:

$$ag_1 = 0, \quad f_0 = f_1 = f_2 = 0, \quad \lambda = 3a^2(2B_0\overline{B}_0 - f_0). \tag{5.1}$$

Subcase (a): $a = 0$. In this case, it is easily seen that one obtains the solution (2.14), for which by (2.79)

$$\Psi_2 = \tfrac{1}{2}(w - ix)^{-3}[ig_1 + B_0\overline{B}_0(w + ix)^{-1}]. \tag{5.2}$$

This quantity is nonzero implying the solution is of type D if $g_1 \neq 0$ or $B_0 \neq 0$.

Subcase (b): $a \neq 0$. The Eqs. (5.1) imply

$$g_1 = 0, \quad \lambda = 6a^2B_0\overline{B}_0, \tag{5.3}$$

from which it follows by (2.7d) that

$$V(x) = -2B_0\overline{B}_0(a^2x^4 + 1) < 0. \tag{5.4}$$

This is an impossibility[83] since by Theorem 2 the function V must be positive in order that the metric have the assumed minus two signature. Moreover, it may be verified by direct calculation the null orbit metric and Maxwell field (2.5) defined by the Eqs. (2.7b), (2.7e), and (5.4) do not satisfy the Einstein–Maxwell Eqs. (2.1).

Case B^0_- : The Eq. (2.8c) implies

$$f_0 = f_1 = f_2 = 0, \tag{5.5}$$

from which it follows by (2.8d) that V has the form (2.15c). We thus have obtained the null orbit solution (2.15) where by (2.8g)

$$\Psi_2 = \tfrac{1}{6}(k - ix)^{-3}[-8k^3\lambda + 3ig_1 + 12B_0\overline{B}_0(k + ix)^{-1}], \tag{5.6}$$

is nonzero implying a type D solution if $k^3\lambda \neq 0$ or $g_1 \neq 0$ or $B_0 \neq 0$. By setting $k = 0$ in (2.15) we obtain the null solution denoted by $C^0_{-,0}$.

Case B^0_+ : The Eq. (2.9c) implies

$$f_1 = g_2 = \lambda = B_0 = 0, \tag{5.7}$$

from which it follows by (2.9g) that

$$\Psi_2 = 0. \tag{5.8}$$

The Eqs. (5.7) and (5.8) imply that the B^0_+ null orbit solution is flat.

Case C^*: The Eqs. (2.10c) yield the relations

$$f_0 = f_1 = f_2 = f_3 = B_0 = 0, \tag{5.9}$$

which by Eq. (2.8g) implies

$$\Psi_2 = 0, \tag{5.10}$$

which in turn implies that the solution is conformally flat. It follows from the above equations that the only nonvanishing curvature component is the cosmological constant λ. Thus the C^* null orbit solution is a space of constant curvature.

Case C^{00}: The Eqs. (2.11c) imply

$$f_0 = f_1 = 0, \quad \lambda = 2B_0\overline{B}_0, \tag{5.11}$$

which equations are easily seen to yield the solution (2.17). By Eq. (2.11g) this solution is type D if $\lambda \neq 0$.

Since the classes of solutions A^*, B^0_-, B^0_+, C^*, and C^{00} exhaust the class $\mathfrak{D}$ the solution A_0 given by (2.14), $B^0_{-,0}$ and $C^0_{-,0}$ given by (2.15) and C^{00} given by (2.17) constitute the only null orbit solutions in $\mathfrak{D}$.

6. PROOF OF THEOREM 4

When one substitutes in the Eq. (2.29) with $m = 0$ for the metric tensor, the vector potential, the curvature scalar, and the function ψ from the Eqs. (2.5a), (2.5c), (2.31), and (3.41j), respectively, one finds after considerable algebra that the nonseparable terms have the form

$$N(w,x) = \tfrac{1}{12}e^{i(\alpha u + \beta v)}\psi_1(w)\psi_2(x)T^3(w,x)Z^{-1}(w,x)O(w,x), \tag{6.1a}$$

where

$$O(w,x) = Z^2(w,x)\{ fU(w)[(\epsilon_1 p'(w))^2 + (\epsilon_1 m'(x))^2 - 2\epsilon_1 Z(w,x)p''(w)] + V(x)[(\epsilon_2 m'(x))^2 + (\epsilon_2 p'(w))^2 + 2\epsilon_2 Z(w,x)m''(x)]\}. \quad (6.1b)$$

In the above expression the factor T^3Z^{-1} is not of concern since it appears in every term of the equation and hence can be removed by division. Thus the real obstruction to separability is the last factor O in the expression. A sufficient condition (which may not be necessary) for the existence of coordinates w and x such that this vanishes is that the type D condition (2.32) or (4.2) is satisfied by the metric (2.5a). To see this, it is required to consider the same cases used in the proof of Theorem 2.

Case A: It follows from the type D condition (4.5) that a coordinate system exists in which $\epsilon_1, \epsilon_2, p$ and m are given by (4.9) from which it follows easily that

$$0(w,x) = 0, \quad (6.2)$$

proving the result in this case.

Case B_-: It follows from the type D condition (4.25) and the remaining coordinate freedom in x that a coordinate system exists in which $\epsilon_1 = 0, \epsilon_2 = -1$, and p and m are given by (4.36). The separability condition (6.2) follows immediately.

Case B_+: By an argument similar to the preceding case it may be shown that the condition (6.2) is satisfied on account of the Eqs. (4.33a).

Case C: A coordinate system exists for which

$$\epsilon_1 = 1, \quad \epsilon_2 = 0, \quad p(w) = 1, \quad m(x) = 0, \quad (6.3)$$

from which (6.2) follows at once.

This completes the proof of Theorem 4 in the case $m = 0$ since, because of the type D condition, the Eq. (2.29) has the separable form

$$\begin{aligned} e^{i(\alpha u + \beta v)}T^3Z^{-1}\{ & fU\psi_2\psi_1'' + V\psi_1\psi_2'' \\ & + [fU' - i(1-f^2)g^{-2}(\alpha p - \epsilon_2\beta)]\psi_2\psi_1' \\ & + V\psi_1\psi_2' + [g^{-4}U^{-1}(\alpha p - \epsilon_2\beta)(f(\alpha\beta - \epsilon_2\beta) \\ & - g^2 eH) - \tfrac{1}{2}ig^{-2}(1-f^2)\alpha p' \\ & + eHU^{-1}(fg^{-4}eH - g^{-2}(\alpha\beta - \epsilon_2\beta)) \\ & + \tfrac{1}{6}U'' - V^{-1}(\epsilon_1\beta - \alpha m)(\epsilon_1\beta - \alpha m + eG) \\ & - eGV^{-1}(\epsilon_1\beta - \alpha m + eG) + \tfrac{1}{6}V'']\psi_1\psi_2\} = 0, \end{aligned} \quad (6.4)$$

where g is given by (3.27). If $m \neq 0$ it is clear from the above and the form (2.31) that the equation is separable iff the function ZT^{-2} has the form (2.22) that is iff Eq. (2.19) is satisfied, where we recall that Z is given by Eq. (3.57d).

This completes the proof of Theorem 4.

ACKNOWLEDGMENTS

Two of the authors (R.D. and R.G.M.) would like to express appreciation to their reciprocal departments for hospitality during recent visits in which this paper was being prepared. One of the authors (R.G.M.) also wishes to thank the Instituts Internationaux de Physique et Chimie, E. Solvay for financial support for his visit to Brussels.

After this paper was concluded, we learned of the work of A. Garcia Diaz on closely related subjects.[84]

[1]B. Carter, Phys. Lett. A **26**, 399 (1968).
[2]B. Carter, Commun. Math. Phys. **10**, 280 (1968).
[3]B. Carter, in *Black Holes*, edited by B. DeWitt and C. DeWitt (Gordon and Breach, New York, 1973), p. 89.
[4]These solutions have been rediscovered by J. F. Plebański, Ann. Phys. (N.Y.) **90**, 196 (1975).
[5]W. Kinnersley, J. Math. Phys. **10**, 1195 (1969).
[6]J. N. Goldberg and R. K. Sachs, Acta Phys. Pol. **22**, 13 (1962).
[7]L. P. Hughston and P. Sommers, Commun. Math. Phys. **33**, 129 (1973).
[8]G. J. Weir and R. P. Kerr, Proc. R. Soc. London Ser. A **355**, 31 (1977).
[9]S. R. Czapor and R. G. McLenaghan, J. Math. Phys. **23**, 2159 (1982).
[10]D. Matravers, Utilitas Math. **2**, 55 (1972).
[11]J. F. Plebański and M. Demiański, Ann. Phys. (N.Y.) **98**, 98 (1976).
[12]Reference 2.
[13]R. Debever, Bull. Cl. Sci. Acad. R. Belg. **LV**, 8 (1969).
[14]R. Debever and N. Kamran, Bull. Cl. Sci., Acad. R. Belg. **LXVI**, 585 (1980).
[15]R. Debever, Bull. Soc. Math. Belg. **XXIII**, 360 (1971).
[16]K. Ishikawa and T. Miyashita, Prog. Theor. Phys. **67**, 828 (1982).
[17]R. Debever and R. G. McLenaghan, Bull. Cl. Sci., Acad. R. Belg. **LXVI**, 204 (1980).
[18]R. Debever and R. G. McLenaghan, J. Math. Phys. **22**, 1711 (1981).
[19]Reference 14.
[20]R. Debever, N. Kamran, and R. G. McLenaghan, Phys. Lett. A **93**, 399 (1983).
[21]See also J. F. Plebański, J. Math. Phys. **20**, 1946 (1979) for a treatment of the nondiverging and nontwisting case.
[22]R. Debever, N. Kamran, and R. G. McLenaghan, Bull. Cl. Sci., Acad. R. Belg. **LXVIII**, 592 (1982).
[23]Reference 13.
[24]Reference 18.
[25]R. Debever, R. G. McLenaghan, and N. Tariq, Gen. Relativ. Gravit. **10**, 853 (1979).
[26]E. Newman and R. Penrose, J. Math. Phys. **3**, 566 (1962).
[27]R. Debever, Cah. Phys. **168–169**, 303 (1964); M. Cahen, R. Debever, and L. Defrise, J. Math. Mech. **16**, 761 (1967).
[28]Reference 25.
[29]I. Robinson and A. Schild, J. Math. Phys. **4**, 484 (1962).
[30]W. Kundt and A. H. Thompson, C. R. Acad. Sci. **254**, 4257 (1962).
[31]J. F. Plebański and S. Hacyan, J. Math. Phys. **20**, 1004 (1979).
[32]A. Garcia and J. F. Plebański, J. Math. Phys. **23**, 123 (1982).
[33]To obtain the forms (1) and (2) of Ref. 20 requires the substitution $u \to (2/(1+f^2))^{1/2}u$, $v \to (2/(1+f^2))^{1/2}v$.
[34]Reference 2.
[35]References 1 and 2.
[36]Reference 4.
[37]M. Demiański and E. T. Newman, Bull. Acad. Pol. Sci. Ser. Sci. Math. Astron. Phys. **XIV**, 653 (1966).
[38]E. T. Newman, E. Couch, R. Chinnapared, A. Exton, A. Prakash, and R. Torrence, J. Math. Phys. **6**, 918 (1965).
[39]R. P. Kerr, Phys. Rev. Lett. **11**, 237 (1963).
[40]A. Taub, Ann. Math. **53**, 472 (1951).
[41]E. T. Newman, L. Tamburino, and T. Unti, J. Math. Phys. **4**, 915 (1963).
[42]D. Brill, Phys. Rev. B **133**, 845 (1964).
[43]B. Bertotti, Phys. Rev. **116**, 1331 (1959).
[44]I. Robinson, Bull. Acad. Pol Sci. Ser. Sci. Math. Astron. Phys. **7**, 351 (1959).
[45]Reference 5.
[46]Reference 11.
[47]Reference 8.
[48]Reference 14.
[49]Reference 16.
[50]Reference 15.
[51]T. Levi-Cività, Atti. Accad. Naz. Lincei Cl. Sci. Fis. Mat. Nat. Rend. **27**, 343 (1918).
[52]E. T. Newman and L. Tamburino, J. Math. Phys. **2**, 667 (1961).
[53]I. Robinson and A. Trautman, Proc. R. Soc. London Ser. A **265**, 463 (1962).

[54]J. Ehlers and W. Kundt in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962), p. 73.
[55]Reference 9.
[56]A. L. Dudley and J. D. Finley, III, J. Math. Phys. **20**, 311 (1979).
[57]References 17 and 18.
[58]Reference 15.
[59]J. Leroy, Bull. Cl. Sci., Acad. R. Belg. **LXIV**, 130 (1978).
[60]A. Garcia and J. F. Plebański, J. Math. Phys. **23**, 1463 (1982), show that the null orbit solution A_0 given by (2.14) is a "limiting contraction" of the seven-parameter solution. They also attribute the A_0 solution to Leroy (Ref. 59). However, it was apparently unknown that Leroy's class of solutions contained the A_0 solution as a subcase until it was pointed out by two of the authors in Refs. 17 and 18.
[61]M. A. Melvin, Phys. Lett. **8**, 55 (1964); his solution may be obtained from the $C^0_{-,0}$ solution in a different coordinate system by setting $\lambda = 0$; see DKM 1 and Ref. 60.
[62]Reference 3, p. 118.
[63]B. Carter, Phys. Rev. D **16**, 3395 (1977).
[64]L. P. Hughston and P. Sommers, Commun. Math. Phys. **32**, 147 (1973).
[65]Reference 63.
[66]P. Moon and D. E. Spencer, *Field Theory Handbook* (Springer-Verlag, New York, 1971), 2nd ed., p. 96.
[67]Reference 9.
[68]Reference 18.
[69]Reference 2.
[70]Reference 2.
[71]Reference 18.
[72]Reference 13, Eq. (3.26).
[73]Reference 13, Eq. (4.5).
[74]Reference 11, Eq. (3.30).
[75]See Theorem II of Ref. 20.
[76]Reference 75.
[77]Reference 20, Eqs. (5.2).
[78]Reference 13, Eq. (3.27).
[79]Reference 2.
[80]Reference 13.
[81]Reference 11.
[82]Reference 8.
[83]Attention was drawn to this difficulty in Ref. 60.
[84]A. García Díaz, J. Math. Phys. **25**, 1951 (1984).

Exact solutions of plane symmetric cosmological models

Roberto Hojman
Departmento de Física, Universidad de Santiago de Chile, Casilla 5659-Correo 2, Santiago, Chile
Jorge Santamarina
Instituto de Física, Universidad Austral de Chile, Casilla 567, Valdivia, Chile

(Received 14 September 1983; accepted for publication 10 February 1984)

The exact solutions for perfect fluid plane symmetric cosmological models with cosmological constant are derived. It is also shown that under an appropriate transformation the $\Lambda = 0$ and $\Lambda \neq 0$ cases are mathematically equivalent.

PACS numbers: 04.20.Jb, 98.80.Bp

I. INTRODUCTION

For certain purposes spherical, plane, and hyperbolic static symmetries can be handled in the same foot if a generalized line element is considered (see, for instance, Ref. 1).

The aim of this article is to study Einstein gravitational field equations with cosmological constant for the mentioned symmetries when a perfect fluid is assumed to be the source of the gravitational field. The field equations for the generalized metric are obtained in Sec. II. In Sec. III we convert by an appropriate transformation the usual three-dimensional phase space[2] into a two-dimensional one that is known to be more suitable for qualitative method analysis.[2,3] The exact solutions for plane symmetry are derived in the fourth section. Finally a brief discussion of the results is given in the fifth section.

II. FIELD EQUATIONS

The line element considered here is

$$ds^2 = g^2(x)dw^2 - dx^2 - f^2(x)[dy^2 + F(y)dz^2], \tag{2.1}$$

which represents the most general static line element admitting spherical, plane, or hyperbolic transformations depending on whether $F(y) = \sin^2 y$, 1 or e^{2y}, respectively.

The field equations to be studied are

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = T_{\mu\nu}, \tag{2.2}$$

with

$$G_{\mu\nu} = R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R, \tag{2.3}$$

$$T_{\mu\nu} = (p+\rho)u_\mu u_\nu - p g_{\mu\nu}, \tag{2.4}$$

and Λ is the cosmological constant. In expression (2.4) ρ and p are, respectively, the energy density and pressure of the fluid whose flow lines are tangent to the unit vector $u^\mu = (1/g)\delta^\mu_0$.

The nonzero components of the curvature are

$$\begin{aligned} R_{00} &= gg'' + 2(g/f)g'f', \\ R_{11} &= -g''/g - 2f''/f, \\ R_{22} &= (R_{33}/F) = \epsilon(F) - f'^2 - ff'' - (f/g)f'g', \end{aligned} \tag{2.5}$$

with

$$\epsilon(F) \equiv \left(\frac{\dot F}{2F}\right)^2 - \frac{\ddot F}{2F} = \begin{cases} 1, & \text{for } F = \sin^2 y, \\ 0, & \text{for } F = 1, \\ -1, & \text{for } F = e^{2y}, \end{cases} \tag{2.6}$$

prime and dot meaning x and y differentiation, respectively. Also we have

$$R = -\frac{2}{f^2}\epsilon(F) + 2\frac{g''}{g} + 4\frac{f''}{f} + 2\frac{f'^2}{f^2} + 4\frac{f'g'}{fg}. \tag{2.7}$$

The field equations (2.2) for the metric given by (2.1) become

$$\begin{aligned} &2\frac{f''}{f} + \frac{f'^2}{f^2} - \frac{1}{f^2}\epsilon = -\rho - \Lambda, \\ &\frac{f'^2}{f^2} + 2\frac{f'g'}{fg} - \frac{1}{f^2}\epsilon = p - \Lambda, \\ &\frac{f''}{f} + \frac{g''}{g} + \frac{f'g'}{fg} = p - \Lambda. \end{aligned} \tag{2.8}$$

The compatibility of the system (2.8) is ensured if the Bianchi identities $G^{\mu\nu}{}_{;\nu} = 0$ are satisfied, i.e.,

$$(\rho + p)(g'/g) + p' = 0. \tag{2.9}$$

Equation (2.9) represents the equation of hydrostatic support.

III. DIMENSIONAL REDUCTION OF PHASE SPACE

In a recent publication[2] Collins devised a sophisticated change of variables under which the field equations (2.8) for $\epsilon = 1$ are converted into a three-dimensional autonomous system. The extension to $\epsilon = 0, -1$ is trivial and is reproduced in what follows.

By multiplying the first of the equations (2.8) times $f^2 f'$ and integrating over x it is found that

$$f'^2 = \epsilon - m(f)/f, \tag{3.1}$$

where

$$\frac{dm(f)}{df} = (\rho + \Lambda)f^2. \tag{3.2}$$

Define

$$M = m/2f, \quad D = \tfrac{1}{2}\rho f^2, \quad P = \tfrac{1}{2}pf^2,$$

$$\lambda = \tfrac{1}{2}\Lambda f^2, \quad t = \ln f.$$

When the equation of state is

$$p = (\gamma - 1)\rho, \tag{3.3}$$

the system (2.8) becomes

$$\frac{dD}{dt} = \frac{D}{\epsilon - 2M}\left[2\epsilon - M\left(\frac{5\gamma - 4}{\gamma - 1}\right) - \gamma D + \frac{\lambda\gamma}{\gamma - 1}\right],$$

$$\frac{dM}{dt} = D + \lambda - M, \tag{3.4}$$

$$\frac{d\lambda}{dt} = 2\lambda,$$

from which a qualitative analysis on the footsteps of Ref. 2 follows.

As it is asserted in Ref. 2 the system (3.4) (with $\epsilon = 1$)"...generalizes that considered previously, wherein $\Lambda = 0$ (so $\lambda = 0$)."[3]

However, without loss of generality, Λ can be put always equal to zero.

In fact the redefinition of the pressure and the energy density as

$$\tilde{p} = p - \Lambda, \quad \tilde{\rho} = \rho + \Lambda \tag{3.5}$$

transforms Eq. (2.8) and (2.9) into an equivalent system with $\Lambda = 0$.

Now twirls can be dropped keeping in mind that the $\Lambda \neq 0$ case is already included. The system is now treatable as it was done in Ref. 3 for $\epsilon = 1$.

The property herein demonstrated is not a peculiarity of the symmetries but of the $T_{\mu\nu}$ given by (2.4) as can be easily proved.

IV. EXACT PLANE SYMMETRIC SOLUTIONS ($\epsilon = 0$)

When $\epsilon = 0$ Eq. (3.1) reads

$$f'^2 = -m(f)/f. \tag{4.1}$$

Combining the last expression with the second Eq. (2.8) (for $\epsilon = 0$) and with Eq. (2.9) one obtains

$$-\frac{m}{f^3} + \frac{2m}{f^2}\frac{dp}{df}\frac{1}{\rho + p} = p. \tag{4.2}$$

As $m = m(f)$ and

$$\rho = \frac{1}{f^2}\frac{dm(f)}{df}, \tag{4.3}$$

Eq. (4.2) is a differential equation for $p = p(f)$. Moreover it is always possible to define a function $G = G(f)$ as follows:

$$G(f) \equiv m(f)/p(f). \tag{4.4}$$

Then (4.2)–(4.4) yield

$$\frac{1}{p}\frac{dp}{df} = \frac{(f^2 + dG/df)(f^3 + G)}{G(f^3 - G)}, \tag{4.5}$$

which can be integrated at once if G is given.

In fact,

$$p(f) = p_0 \exp\left(\int \frac{(f^2 + dG/df)(f^3 + G)}{G(f^3 - G)}\,df\right), \tag{4.6}$$

where p_0 is an integration constant.

From $p = p(f)$ thus determined $\rho(f)$ can be readily obtained if Eq. (4.3) and (4.4) are used:

$$\rho(f) = \frac{1}{f^2}\left(p\frac{dG}{df} + \frac{dp}{df}G\right). \tag{4.7}$$

The expressions (4.6) and (4.7) for $p(f)$ and $\rho(f)$ constitute the parametric equation of state.

Equation (2.9) can be directly integrated for g after $p(f)$ and $\rho(f)$ have been found:

$$g(f) = g_0 \exp\left(-\int \frac{dp/df}{\rho + p}\,df\right), \tag{4.8}$$

g_0 being an integration constant.

Finally to recover the original variable x we proceed as follows: when $G(f)$ is given and consequently $p(f)$ is determined through (4.6), $m(f)$ can be constructed from the relation (4.4). Besides, Eq. (4.1) can be integrated giving

$$x = \int \sqrt{-\frac{f}{m(f)}}\,df, \tag{4.9}$$

which determines implicitly the metric coefficient f in terms of x.

V. DISCUSSION

A crucial step in the integration of the field equations is constituted by the definition of the function $G(f)$. As it was demonstrated, once G is fixed, p and ρ can be determined so that G contains the equation of state. For instance if G is selected to be proportional to f^3 (constant of proportionality α different from 0, 1) $p(f)$ and $\rho(f)$ are found to be proportional to f^r (varying α, r can take any value). Consequently the equation of state is of the form $p = (\gamma - 1)\rho$. It can also be found that in this case f (and therefore p and ρ) decreases as x^{-2}.

It is possible to argue that other functions rather than G can be accommodated to make Eq. (2.8) and (2.9) solvable. However this does not happen to be the case. The other functions we have at hand are ρ and p. If $\rho = \rho(f)$ is fixed, then Eq. (4.2) becomes a generalized Riccati equation.[4] Conversely, when $p = p(f)$ is given, Eq. (4.2) can be identified as an Abel equation of the first kind.[4] Both types of equations are solvable only in few special cases (i.e., for few ρ or p choices) so integration can be carried out just for exceptional selections.[4]

On the other hand the procedure exhibited here allows one to solve the field equations (2.8) and their compatibility condition (2.9) with three quadratures [see (4.6), (4.8), and (4.9)].

ACKNOWLEDGMENTS

One of us (R. H.) is indebted to Dr. Flavio Ojeda for hospitality at the Instituto de Física, Universidad Austral de Chile, where this work was done.

This work was supported by the Dirección de Investigaciones Científicas y Tecnológicas, Universidad de Santiago de Chile, and by the Dirección de Investigación y Desarrollo, Universidad Austral de Chile, under Grant No. S8235.

[1] R. Hojman and A. Smailagic, J. Math. Phys. **24**, 125 (1983).
[2] C. B. Collins, J. Math. Phys. **24**, 215 (1983).
[3] C. B. Collins, J. Math. Phys. **18**, 1374 (1977).
[4] G. M. Murphy, *Ordinary Differential Equations and their Solutions* (Van Nostrand, Princeton, 1960).

Cosmological solutions of the Einstein equation with heat flow

Abhik Kumar Sanyal
Department of Physics, Calcutta University, 92 A.P.C. Road, Calcutta-700009, India
Dipankar Ray
Department of Applied Mathematics, Calcutta University, 92 A.P.C. Road, Calcutta-700009, India

(Received 27 January 1983; accepted for publication 2 December 1983)

Solutions of Einstein equation for fluids with heat flow are obtained, generalizing the results of Bergmann.

PACS numbers: 04.20.Jb

1. INTRODUCTION

Consider Einstein's equation of general relativity for a fluid with heat flow having energy–momentum tensor

$$T^{\alpha\beta} = (\rho + p)v^\alpha v^\beta - pg^{\alpha\beta} + q^\alpha v^\beta + q^\beta v^\alpha, \tag{1}$$

where $q_\alpha v^\alpha = 0$, with the line element

$$dS^2 = A^2 dt^2 - B^2(dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\, d\phi^2) \tag{2}$$

where A and B can be functions of t and r.

Einstein's equation with (1) and (2) have been reduced by Bergmann[1] to

$$A'' + 2(F'/F)A' - (F''/F)A = 0, \tag{3}$$

where $F = B^{-1}$ and dots and primes denote differentiation with respect to t and $x = r^2$, respectively. Bergmann found a simple solution with $A = 1$. We are looking here for general solutions. The radial component of the heat flow is given by

$$q = \left(\frac{4r}{GB^2}\right)\left(\frac{\dot B}{AB}\right)', \tag{4}$$

where G is the constant of gravitation.

2. SOLUTIONS

Case 1: $A'' = 0$:

$$\therefore \quad A' = Q(t) \quad \text{and} \quad A = Q(t)x + P(t). \tag{5}$$

Equation (3) becomes

$$2QF' - QxF'' - PF'' = 0.$$

Integrating and dividing throughout by $(Qx + P)^4$, we get

$$\left(\frac{F}{Qx+P}\right)' + \frac{h(t)}{(Qx+P)^4} = 0.$$

Further integration yields

$$F = h/3Q + (Qx+P)^3 L,$$

i.e.,

$$B = F^{-1} = [h/3Q + (Qx+P)^3 L]^{-1}, \tag{6}$$

where h, Q, P, L are all functions of t.

Equations (5) and (6) can be used to find the value of q.

Case 2: $A'' \neq 0$: By Eq. (3) $F' \neq 0$ so that one can express A as

$$A = A(F,t); \quad \therefore \quad A' = A_F F', \quad A'' = A_{FF}F'^2 + A_F F''. \tag{7}$$

So Eq. (3) for this case reduces to

$$\frac{A_{FF} + 2A_F/F}{A_F - A/F}\, dF + \frac{dF'}{F'} = 0.$$

Integrating

$$\int \left[\frac{A_{FF} + 2A_F/F}{A_F - A/F}\right] dF + \log F' = e^{\alpha(t)}$$

or

$$\exp \int \left[\frac{A_{FF} + 2A_F/F}{A_F - A/F}\, dF\right] = \alpha(t)\frac{dx}{dF}.$$

Integrating once again, we get

$$\int \left[\exp \int \left(\frac{A_{FF} + 2A_F/F}{A_F - A/F}\, dF\right)\right] dF = \alpha(t)x + \beta(t). \tag{8}$$

If A is given as a function of F and t, then the integral can be evaluated and hence the solution can be obtained. Nevertheless, for a particular case simple solutions are obtained as follows.

Consider $F'' = mF$, where m is a function of time alone. So Eq. (3) can be written as

$$\frac{U''}{U} = 2\frac{F''}{F} = \pm k^2 \text{ (function of time)}, \tag{9}$$

where

$$U = AF, \tag{10}$$

$$\therefore \quad U'' = \pm k^2 U. \tag{11}$$

Solutions of Eq. (11) are given below:

(i) $U = C_1 e^{kx} + D_1 e^{-kx}$, when m is positive ($m = k^2$);

(ii) $U = C_1 \cos(kx) + D_1 \sin(kx)$, when m is negative ($-m = k^2$);

(iii $U = qx + r$, when m is zero.

(i) When m is positive,

$$\therefore \quad 2F''/F = k^2 \quad \text{from Eq. (9);}$$

$$\therefore \quad F = C_2 e^{kx/\sqrt{2}} + D_2 e^{-kx/\sqrt{2}}. \tag{12}$$

Again, from (10) and (i)

$$AF = U = C_1 e^{kx} + D_1 e^{-kx},$$

$$\therefore \quad A = \frac{C_1 e^{kx} + D_1 e^{-kx}}{C_2 e^{kx/\sqrt{2}} + D_2 e^{-kx/\sqrt{2}}}, \tag{13}$$

and

$$B = F^{-1} = \frac{1}{C_2 e^{kx/\sqrt{2}} + D_2 e^{-kx/\sqrt{2}}},$$

where C_1, D_1, C_2, D_2, and k are all functions of time.

Equation (13) can be used to evaluate q.

(ii) When m is negative ($-m = k^2$),

$$\therefore \quad 2F''/F = -k^2 \quad [\text{from Eq. (9)}],$$

$$\therefore \quad F = C_3 \cos(kx/\sqrt{2}) + D_3 \sin(kx/\sqrt{2}). \tag{14}$$

Again,

$AF = U = C_1 \cos(kx) + D_1 \sin(kx)$ from Eq. (10) and (ii),

$$\therefore \quad A = \frac{C_1 \cos(kx) + D_1 \sin(kx)}{C_3 \cos(kx/\sqrt{2}) + D_3 \sin(kx/\sqrt{2})}$$

and

$$B = F^{-1} = \frac{1}{C_3 \cos(kx/\sqrt{2}) + D_3 \sin(kx/\sqrt{2})}, \tag{15}$$

where C_1, D_1, C_3, D_3, and k are all functions of time.

Equation (15) can be used to evaluate q.

(iii) When m is zero (i.e., $k^2 = 0$)

$$\therefore \quad 2F''/F = 0 \quad \text{from Eq. (9)},$$

i.e.,

$$F'' = 0, \quad F' = k(t),$$

$$F = k(t)x + C(t). \tag{16}$$

Substituting this in Eq. (3), we get

$$kx \frac{d}{dx}\left(\frac{dA}{dx}\right) + C \frac{d}{dx}\left(\frac{dA}{dx}\right) + 2k \frac{dA}{dx} = 0.$$

Integration yields

$$kx\, dA + kA\, dx + C\, dA = f(t)\, dx.$$

Integrating again, we get

$$A = \frac{f(t)x + g(t)}{k(t)x + C(t)} + B = F^{-1} = \frac{1}{k(t)x + C(t)}. \tag{17}$$

Equation (17) can be used to find the value of q.

3. CONCLUSION

Summarily, the present paper gives the complete set of solutions of Eq. (3) either explicitly or implicitly. For $A'' = 0$, solutions have been obtained explicitly and are given by (5) and (6). For $A'' \neq 0$ solutions are given implicitly by (8). However, some explicit solutions can be obtained for $F'' = \pm \frac{1}{2}k^2F$ [Eq. (13)], for $F'' = \frac{1}{2}k^2F$ [Eq. (15)], and for $F'' = 0$ [Eq. (17)].

It has already been stated that the solution of (3) give the solutions of Einstein equation for the metric (2) and energy–momentum tensor (1), where $B = F^{-1}$ and q is the heat flow given by (4). Having obtained these solutions, it remains to be shown that they are physically acceptable. Certain energy conditions have to be satisfied, especially, that the energy density be positive everywhere.

ACKNOWLEDGMENT

The authors would like to thank Dr. A. Banerjee for bringing to their attention the work of Bergmann.

[1] O. Bergmann, Phys. Lett. A **82**, 383 (1981).

Scalar meson field in a conformally flat space

Abhik Kumar Sanyal
Department of Physics, University College of Science, 92 A.P.C. Road, Calcutta-700009, India
Dipankar Ray
Department of Applied Mathematics, University College of Science, 92 A.P.C. Road, Calcutta-700009, India

(Received 10 January 1983; accepted for publication 19 August 1983)

Among several authors, who studied massive and massless scalar meson fields in general relativity, attempts to obtain a complete set of solutions for a conformally flat metric $e^{\psi}(dx^{1^2} + dx^{2^2} + dx^{3^2} - dx^{4^2})$ were made by Ray for massive and massless mesons and Gursay for massless mesons. Both of them concluded that ψ must be a function of $K_0(x^{1^2} + x^{2^2} + x^{3^2} - x^{4^2}) + K_1x^1 + K_2x^2 + K_3x^3 + K_4x^4$ where K_0, K_1, K_2, K_3, K_4 are all constants. Both Ray and Gursay, however, overlooked an important particular case, which is studied here. As a by-product certain equations obtained by Auria and Regge in connection with "Gravitational theories with asymptotic flat instantons," are solved under less restrictive assumptions.

PACS numbers: 04.20.Jb

I. INTRODUCTION

In general relativity the field equations for a scalar meson field are given by

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = -\,[\phi_{,\mu}\phi_{,\nu} - \tfrac{1}{2} g_{\mu\nu}(\phi_{,\alpha}\phi^{,\alpha} - m^2\phi^2)], \tag{1.1}$$

where $K = 1$ and ϕ is the meson field and m is the meson mass and $\phi_{,\mu} = \partial\phi/\partial x^{\mu}$, and so on. Equation (1.1) can be rewritten as

$$R_{\mu\nu} = -\phi_{,\mu}\phi_{,\nu} + \tfrac{1}{2} g_{\mu\nu} m^2 \phi^2. \tag{1.2}$$

For a conformally flat metric

$$g_{\mu\nu} = e^{\psi}\eta_{\mu\nu},$$

where

$$\begin{aligned} \eta_{\mu\nu} &= 0 \quad \text{for } \mu \neq \nu, \\ &= 1 \quad \text{for } \mu = \nu = 1, 2, 3, \\ &= -1 \quad \text{for } \mu = \nu = 4. \end{aligned}$$

$R_{\mu\nu}$ takes the form

$$R_{\mu\nu} = \psi_{,\mu\nu} - \tfrac{1}{2}\psi_{,\mu}\psi_{,\nu} + \tfrac{1}{2}\eta_{\mu\nu}x, \tag{1.3}$$

where

$$\chi = \psi_{,11} + \psi_{,22} + \psi_{,33} - \psi_{,44} + \psi_{,1}^2 + \psi_{,2}^2 + \psi_{,3}^2 - \psi_{,4}^2. \tag{1.4}$$

The coupled equations (1.2)–(1.4) form the complete set of equations for a scalar meson field in a conformally flat space in general relativity. In order to simplify these equations and to study the properties of their solutions Ray[1] noted that for $\phi = \text{const}$ the solution of the above-mentioned set of coupled equations is found quite simply, while for $\phi \neq \text{const}$, the coupled equations lead to

$$(e^{-\psi/2})_{,\mu} = \eta P_{,\mu} + \rho Q_{,\mu}, \tag{1.5a}$$

$$\tfrac{1}{2} e^{-\psi/2}\phi_{,\mu} = \eta_{,\phi} P_{,\mu} + \rho_{,\phi} Q_{,\mu}, \tag{1.5b}$$

$$2\eta_{,\phi\phi} + \eta = \xi(\phi)\eta_{,\phi}, \tag{1.6a}$$

$$2\rho_{,\phi\phi} + \rho = \xi(\phi)\rho_{,\phi}, \tag{1.6b}$$

where $\eta = \eta(\phi)$ and $\rho = \rho(\phi)$ and[2]

$$P = Ax^1 + A'x^4, \tag{1.7a}$$

$$Q = \tfrac{1}{2} B(x^{1^2} + x^{2^2} + x^{3^2} - x^{4^2}) + B_1x^1 + B_2x^2 + B_3x^3 + B_4x^4, \tag{1.7b}$$

where $A, A', B, B_1, B_2, B_3, B_4$ are constants. From (1.5) Ray[1] obtained

$$2\exp\left(-\frac{\epsilon}{2}\right)\left(\frac{\rho_{,\phi}\eta}{\eta_{\phi}} - \rho\right) = \frac{f'(Q)}{\exp(f(Q)/2)}, \tag{1.8}$$

where $\epsilon = \int \eta \, d\phi/\eta_{,\phi}$ and f is a function of Q.

Since the left-hand side of Eq. (1.8) is a function of ϕ and the right-hand side of Eq. (1.8) is a function of Q, it was argued by Ray[1] that from Eq. (1.8) one gets $\phi = \phi(Q)$. However, it is obvious from Eq. (1.8) that there arises another possibility which is

$$2\exp\left(-\frac{\epsilon}{2}\right)\left(\frac{\rho_{\phi}\eta}{\eta_{\phi}} - \rho\right) = \frac{f'(Q)}{\exp(f(Q)/2)} = K', \tag{1.9}$$

where K' is a constant. In this case ϕ need not be a function of Q. In the present paper we shall see whether this previously left out possibility of ϕ not being a function of Q leads to some new solutions. Also we explore whether the results obtained here can be suitably modified so as to be applied in the case of gravitational instantons (Sec. III).

II. NEW SOLUTIONS FOR SCALAR MESON FIELD

In search of solutions that may have been left out by Ray[1] we require that

$$\eta_{,\phi} \neq 0 \quad \text{and} \quad \rho_{,\phi}/\eta_{,\phi} \neq \text{const}. \tag{2.1}$$

Also

$$Q_{,\mu} \neq 0 \quad \text{and} \quad P_{,\mu}/Q_{,\mu} \neq \text{const}. \tag{2.2}$$

The reason for making these requirements is that, if Eq. (2.1) is not satisfied from Eq. (1.5b), one gets

$$\phi = \phi(KP + K'Q),$$

where K and K' are constants.

In view of Eq. (1.7), the above equation is essentially of the same form as $\phi = \phi(Q)$. Similarly if Eq. (2.2) holds from

Eq. (1.5b) one gets $\phi = \phi(Q)$. Since all the solutions of the form $\phi = \phi(Q)$ are already given by Ray,[1] there is no need to look for such solutions. Now from Eq. (1.5) one can write

$$\begin{aligned} &\psi = \psi(P,Q), \qquad \phi = \phi(P,Q), \\ &(e^{-\psi/2})_{,P} = \eta, \qquad (e^{-\psi/2})_{,Q} = \rho, \\ &\tfrac{1}{2}e^{-\psi/2}\phi_{,P} = \eta_{,\phi}, \quad \tfrac{1}{2}e^{-\psi/2}\phi_{,Q} = \rho_{,Q}. \end{aligned} \tag{2.3}$$

$\phi = $ const surfaces are given by

$$\eta_\phi\, dP + \rho_\phi\, dQ = 0.$$

Integrating, treating ϕ as constant,

$$\eta_\phi P + \rho_\phi Q = \lambda_\phi. \tag{2.4}$$

From Eqs. (2.1) and (2.2) we get

$$(\eta P + \rho Q)_\mu = \lambda_\phi \phi_\mu + (e^{-\psi/2})_\mu.$$

Integrating and absorbing the constant of integration we get

$$\eta P + \rho Q = \lambda + e^{-\psi/2}. \tag{2.5}$$

Differentiating Eq. (2.4) with respect to P, we get

$$(\eta_{\phi\phi}P + \rho_{\phi\phi}Q - \lambda_{\phi\phi})\phi_P + \eta_\phi = 0.$$

From Eq. (2.3) we obtain

$$(2\lambda_{\phi\phi} + \lambda) - (2\eta_{\phi\phi} + \eta)P - (2\rho_{\phi\phi} + \rho)Q = 0.$$

From Eqs. (1.6) and (2.2) we get

$$2\lambda_{\phi\phi} + \lambda = \xi(\phi)\lambda_\phi. \tag{2.6}$$

From Eqs. (1.6) and (2.6), η, ρ, and λ are found to be the solutions of the same ordinary linear second-order differential equation, and since in view of Eq. (2.1) η and ρ are linearly independent, there exists D and D' such that $\lambda = \eta D + \rho D'$ where D and D' are constants.

From Eqs. (2.4) and (2.6)

$$\eta_\phi(P - D) + \rho_\phi(Q - D') = 0, \tag{2.7a}$$

$$\eta(P - D) + \rho(Q - D') = e^{-\psi/2}. \tag{2.7b}$$

Now Eq. (1.9) gives

$$2e^{-\epsilon/2}(\rho_\phi \eta/\eta_\phi - \rho) = K'.$$

Substituting η and η_ϕ from Eq. (2.7) and simplifying we get

$$e^{\psi/2} = (2/K'(D' - Q))e^{-\epsilon/2}. \tag{2.8}$$

Now differentiating Eq. (2.7a) with respect to ϕ and using (1.5b) we get

$$\tfrac{1}{2}e^{-\psi/2} + (P - D)\eta_{\phi\phi} = (D' - Q)\rho_{\phi\phi}.$$

Now substituting $e^{-\psi/2}$ from Eq. (2.8) we have

$$(P - D)\eta_{\phi\phi} = (D' - Q)\left[\rho_{\phi\phi} - \frac{K'e^{\epsilon/2}}{4}\right]. \tag{2.9}$$

From Eqs. (2.9) and (2.7) we get

$$\eta_{\phi\phi}\rho_\phi = \eta_\phi \rho_{\phi\phi} - \eta_\phi K' e^{\epsilon/2}/4. \tag{2.10}$$

Let

$$P - D = P' \quad \text{and} \quad Q - D' = Q' \tag{2.11a}$$

so that

$$\alpha = P'/Q', \tag{2.11b}$$

where

$$P' = Ax^1 + A'x^4 - D \tag{2.11c}$$

and

$$\begin{aligned} Q' = \tfrac{1}{2}B(x^{1^2} + x^{2^2} + x^{3^2} - x^{4^2}) \\ + B_1x^1 + B_2x^2 + B_3x^3 + B_4x^4 - D'. \end{aligned} \tag{2.11d}$$

Also

$$\phi = \phi(\alpha). \tag{2.11e}$$

Now,

$$\phi_{,\mu} = \frac{\phi_\alpha P'_\mu}{Q'} - \frac{P'}{Q'^2}\phi_\alpha Q'_\mu.$$

Comparing the above equation with Eq. (1.5b) and noting that P and P' differ by a constant and likewise Q and Q', we get

$$2e^{\psi/2}\eta_\phi = \phi_\alpha/Q', \tag{2.12a}$$

$$2e^{\psi/2}\rho_\phi = -(P'/Q'^2)\phi_\alpha. \tag{2.12b}$$

The general scalar meson field equation is

$$-2e^{\psi/2}(e^{-\psi/2})_{\mu\nu} = -\phi_\mu\phi_\nu + \tfrac{1}{2}g_{\mu\nu}m^2\phi^2. \tag{2.13}$$

From Eq. (2.12a) we get

$$\begin{aligned} (e^{-\psi/2})_{\mu\nu} = \frac{Q'_{\mu\nu}}{F} - \frac{F_\alpha}{F^2}(Q'_\mu\alpha_\nu + Q'_\nu\alpha_\mu + Q'\alpha_{\mu\nu}) \\ + \frac{Q'\alpha_\mu\alpha_\nu}{F^2}\left(2\frac{F_\alpha^2}{F} - F_{\alpha\alpha}\right), \end{aligned} \tag{2.14}$$

where

$$F = \frac{\phi_\alpha}{2\eta_\phi} = Q'e^{\psi/2}. \tag{2.15}$$

Case I: Let $\mu \neq \nu$. Equation (2.13) becomes

$$\begin{aligned} \alpha_\mu\alpha_\nu\left[\frac{Q'}{F^2}\left(2\frac{F_\alpha^2}{F} - F_{\alpha\alpha}\right) - \frac{Q'}{2F}\phi_\alpha^2\right] \\ = \frac{F_\alpha}{F^2}(Q'_\mu\alpha_\nu + Q'_\nu\alpha_\mu + Q'\alpha_{\mu\nu}). \end{aligned} \tag{2.16}$$

From Eq. (2.11) we get

$$(\alpha Q')_{\mu\nu} = P'_{\mu\nu} = 0 \quad (\text{for } \mu \neq \nu)$$

which gives

$$Q'_\mu\alpha_\nu + Q'_\nu\alpha_\mu + Q'\alpha_{\mu\nu} = 0 \quad (\text{for } \mu \neq \nu).$$

From Eq. (2.16) we get

$$\frac{1}{F}\left(2\frac{F_\alpha^2}{F} - F_{\alpha\alpha}\right) - \frac{1}{2}\phi_\alpha^2 = 0. \tag{2.17}$$

Case II: Let $\mu = \nu$. Using equations (2.13) and (2.17) it can easily be shown that the field equation for $\mu = \nu$ turns out to be

$$\frac{F_\alpha}{F}(A - B\alpha) = B + \frac{F^2m^2\phi^2}{4Q'}. \tag{2.18}$$

Now, since A, B, m are constants, F and ϕ are functions of α and Q' is not a function of α. So Eq. (2.18) can hold only if $m = 0$. So Eqs. (2.18) and (2.17) give

$$\begin{aligned} &F = C/(A - B\alpha), \\ &\phi = 2\sqrt{2}\log(A - B\alpha), \end{aligned} \tag{2.19a}$$

where

$$\alpha = \frac{P'}{Q'} = \frac{Ax^1 + A'x^4 - D}{\frac{1}{2}B(x^{1^2} + x^{2^2} + x^{3^2} - x^{4^2}) + B_1x^1 + B_2x^2 + B_3x^3 + B_4x^4 - D'} \quad (2.19b)$$

and A, B, C are all constants.

III. APPLICATION TO THE CASE OF GRAVITATIONAL THEORIES WITH ASYMPTOTIC FLAT INSTANTONS

In a recent paper in connection with "Gravitational theories with asymptotic flat instantons" Auria and Regge[3] have sought to obtain a conformally flat solution of

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = \frac{3}{8}(\partial_\mu\phi\partial_\nu\phi - \frac{1}{2}g_{\mu\nu}\partial^\alpha\phi\partial_\alpha\phi) - \frac{3}{8}g_{\mu\nu}M(\phi), \quad (3.1)$$

where

$$\phi_{;\mu}^{;\mu} = \frac{1}{2}\delta M/\partial\phi,$$

where the metric is in the form $(+\ +\ +\ +)$. Auria and Regge[3] assumed

$$g_{\mu\nu} = e^{\psi}(dx^{1^2} + dx^{2^2} + dx^{3^2} + dx^{4^2}), \quad (3.2)$$

where

$$\psi = \psi(x^{1^2} + x^{2^2} + x^{3^3} + x^{4^2}). \quad (3.3)$$

However, proceeding as in the paper by Ray[1] and the present paper, it is easy to see that for solutions of Eq. (3.1) of the form Eq. (3.2) unless $M(\phi) = 0$, ψ can be expressed as a function of either $x^{1^2} + x^{2^2} + x^{3^2} + x^{4^2}$ or $K_1x^1 + K_2x^2 + K_3x^3 + K_4x^4$ where K_1, K_2, K_3, K_4 etc. are constant. However, if ψ is a function of $K_1x^1 + K_2x^2 + K_3x^3 + K_4x^4$, it is obvious that the solution cannot be asymptotically flat. Hence ψ must be a function of $x^{1^2} + x^{2^2} + x^{3^2} + x^{4^2}$. So Eq. (3.3) is not an assumption but can be obtained automatically from other conditions.

IV. CONCLUSION

In summary, we find that Ray's conclusion that for massive and massless scalar meson fields in general relativity with a conformally flat metric $e^{\psi}(dx^{1^2} + dx^{2^2} + dx^{3^2} - dx^{4^2})$ that one must have ψ as a function of $K_0(x^{1^2} + x^{2^2} + x^{3^2} - x^{4^2}) + K_1x^1 + K_2x^2 + K_3x^3 + K_4x^4$, is true except that another solution given by Eq. (2.19) is also possible for massless mesons. It should be noted that the case of massless mesons was also studied by Gursay[4] and he also overlooked the above possibility. However, the calculations presented by Gursay[4] are too sketchy for us to find where the oversight occurred. Furthermore we find that the above result has significant application in the study of the "Gravitational theories with asymptotically flat instantons" by Auria and Regge,[3] i.e., Eq. (3.3) is not an assumption but it naturally follows from Eqs. (3.1) and (3.2).

[1] D. Ray, J. Math. Phys. **18**, 1899 (1977).

[2] In the work of Ray,[1] the expression for P was given by

$$P = \frac{1}{2}A(x^{1^2} + x^{2^2} + x^{3^2} - x^{4^2}) + A_1x^1 + A_2x^2 + A_3x^3 + A_4x^4.$$

But this expression for P can be reduced to the form given by Eq. (1.7a) by suitable choice of coordinates.

[3] R. D'Auria and T. Regge, Nucl. Phys. B **195**, 2, 308 (1982).

[4] M. Gürses, Phys. Rev. D **15**, 2731 (1977).

Killing vectors in algebraically special space-times

G. F. Torres del Castillo
Departamento de Física Matemática, Instituto de Ciencias de la Universidad Autónoma de Puebla, Puebla, México and Departamento de Física, Centro de Investigación y de Estudios Avanzados del IPN, Apartado Postal 14-740, 07000, México D. F., Mexico

(Received 18 October 1983; accepted for publication 23 December 1983)

The form of the isometric, homothetic, and conformal Killing vectors for algebraically special metrics which admit a shear-free congruence of null geodesics is obtained by considering their complexification, using the existence of a congruence of null strings. The Killing equations are partially integrated and the reasons which permit this reduction are exhibited. In the case where the congruence of null strings has a vanishing expansion, the Killing equations are reduced to a single master equation.

PACS numbers: 04.20.Me, 04.20.Cv

I. INTRODUCTION

The mathematical difficulties which appear when dealing with the equations of general relativity, in the case of the algebraically special metrics which admit a shear-free congruence of null geodesics, become manageable. Among the different approaches used in the study of this class of metrics, a very powerful one is that based on the concept of null strings, which arises when one considers their complexification. In fact, this concept and the spinorial formalism allows one to analyze and to integrate the equations for this wide class of metrics with relative ease. As an example, the reduction of the Einstein vacuum field equations to a single differential equation for one function and the reduction, in the case of vacuum, of the Killing equations, including the case of homothetic Killing vectors, to a single equation can be obtained by this approach (see Refs. 1,2 and 3,4, respectively).

In this paper the equations for the isometric, homothetic, and conformal Killing vectors of any algebraically special metric which admits a shear-free congruence of null geodesics along the repeated principal direction of the Weyl tensor are partially integrated under the only further restriction that the conformal curvature does not vanish. In the case in which the congruence of null strings has a vanishing expansion and the energy-momentum tensor of the matter has a constant trace and is suitably aligned to the congruence, the Killing equations are reduced to a single master equation. In Sec. II, a brief description of the coordinate systems and the induced bases to be used is given. In Sec. III, the partial integration of the Killing equations is presented, while in Sec. IV, the integrability conditions on those equations are considered. Finally, in Sec. V, a single master equation is obtained which determines the possible Killing vectors for the particular case mentioned above. The formalism and notation used in this paper are those of Ref. 5. All the spinorial indices are manipulated according to the convention $\psi_A = \epsilon_{AB}\psi^B$, $\psi^B = \psi_A \epsilon^{AB}$, and similarly for dotted indices.

II. PRELIMINARIES

The coordinate systems to be used are adapted to the foliation that every space-time of the class under consideration possesses. If $l^A l^{\dot{B}} \partial_{A\dot{B}}$ is a geodesic, shear-free field of null directions,[6] then the system of differential equations[7]

$$l_A g^{A\dot{B}} = 0, \tag{2.1}$$

defines a congruence of two-dimensional totally geodesic null surfaces (null strings). When, additionally, $l^A l^{\dot{B}} \partial_{A\dot{B}}$ is a repeated principal null direction of the Weyl tensor, there exist coordinates $q^{\dot{A}}, p^{\dot{A}}$ such that the metric has the form[2]

$$g = 2\phi^{-2}\, dq^{\dot{A}} \underset{s}{\otimes} (dp_{\dot{A}} + Q_{\dot{A}\dot{B}}\, dq^{\dot{B}}), \tag{2.2}$$

where ϕ satisfies

$$l^B \nabla_{A\dot{C}} l_B = l_A l^B \partial_{B\dot{C}} \ln \phi, \tag{2.3}$$

and $Q_{\dot{A}\dot{B}}$ is a symmetric object. Thus, the set of vector fields

$$\begin{aligned} \partial_{1\dot{A}} &= \sqrt{2}\,\frac{\partial}{\partial p^{\dot{A}}} \equiv \sqrt{2}\,\partial_{\dot{A}}, \\ \partial_{2\dot{A}} &= \sqrt{2}\phi^2 \left(\frac{\partial}{\partial q^{\dot{A}}} - Q_{\dot{A}}{}^{\dot{B}} \frac{\partial}{\partial p^{\dot{B}}}\right) \equiv \sqrt{2}\phi^2 D_{\dot{A}}, \end{aligned} \tag{2.4}$$

constitutes a null tetrad.[8] The connection one-forms for this tetrad and the components of the curvature can be found in Ref. 2. In terms of the coordinates $q^{\dot{A}}, p^{\dot{A}}$, the null strings are defined by $dq^{\dot{A}} = 0$.

In place of the coordinates $q^{\dot{A}}$ one can use any other pair of independent functions $q'^{\dot{A}}$ which are constant on the null strings. Therefore, $q'^{\dot{A}} = q'^{\dot{A}}(q^{\dot{B}})$ and, in order to preserve the form of the metric (2.2), the coordinates $p^{\dot{A}}$ must be replaced by

$$p'^{\dot{A}} = -\rho^{-1} T^{-1}{}_{\dot{B}}{}^{\dot{A}} p^{\dot{B}} + \sigma^{\dot{A}}, \tag{2.5}$$

where $(T^{-1\dot{A}}{}_{\dot{B}})$ is the inverse of

$$(T^{\dot{A}}{}_{\dot{B}}) \equiv \left(\frac{\partial q'^{\dot{A}}}{\partial q^{\dot{B}}}\right), \tag{2.6}$$

$\sigma^{\dot{A}}$ and $\rho \equiv \phi^2/\phi'^2$ are functions of $q^{\dot{A}}$ only.

III. REDUCTION OF KILLING'S EQUATIONS

Let $K = -\frac{1}{2} K^{A\dot{B}} \partial_{A\dot{B}}$ be a vector field. The covariant derivatives of K can be decomposed into their irreducible components as

$$\nabla_R{}^{\dot{A}} K_S{}^{\dot{B}} = E_{RS}{}^{\dot{A}\dot{B}} + l_{RS}\epsilon^{\dot{A}\dot{B}} + \epsilon_{RS} l^{\dot{A}\dot{B}} - 2\chi \epsilon_{RS} \epsilon^{\dot{A}\dot{B}}, \tag{3.1a}$$

where

 0022-2488/84/061980-05$02.50

$$E_{RS}{}^{\dot{A}\dot{B}} = \nabla_{(R}{}^{(\dot{A}}K_{S)}{}^{\dot{B})}, \tag{3.1b}$$

$$l_{RS} = \tfrac{1}{2}\nabla_{(R}{}^{\dot{A}}K_{S)\dot{A}}, \quad l^{\dot{A}\dot{B}} = \tfrac{1}{2}\nabla^{R(\dot{A}}K_R{}^{\dot{B})}, \tag{3.1c}$$

$$\chi = -\tfrac{1}{8}\nabla^{R\dot{A}}K_{R\dot{A}}. \tag{3.1d}$$

The quantities $E_{RS\dot{A}\dot{B}}$ are the spinorial components of the traceless part of $K_{(\mu;\nu)}$ and $\chi = \frac{1}{4}K^{\mu}{}_{;\mu}$.

The vector field K is a conformal Killing vector if $K_{(\mu;\nu)}$ is proportional to $g_{\mu\nu}$; therefore K is a conformal Killing vector when $E_{RS\dot{A}\dot{B}} = 0$. The vector field K is an homothetic or an isometric Killing vector when, additionally, χ is a nonzero constant or zero, respectively. The set of equations, $E_{RS\dot{A}\dot{B}} = 0$, will be referred to as Killing's equations. If these equations are fulfilled, then K will be called, generically, a Killing vector.

Using the Ricci and Bianchi identities one finds that

$$\tfrac{1}{2}\nabla_{(A}{}^{\dot{R}}\nabla_B{}^{\dot{S}}E_{CD)\dot{R}\dot{S}} = C_{(AB}{}^{\dot{R}\dot{S}}E_{CD)\dot{R}\dot{S}} + K^{\dot{R}\dot{S}}\nabla_{\dot{R}\dot{S}}C_{ABCD} + 4l_{S(A}C^S{}_{BCD)} - 4\chi C_{ABCD}, \tag{3.2}$$

where $C_{AB\dot{R}\dot{S}}$ and C_{ABCD} denote the spinorial components of the traceless part of the Ricci tensor and of the self-dual part of the conformal curvature tensor, respectively. Referring now all the quantities to the null tetrad (2.4), taking $A = B = C = D = 1$ in Eq. (3.2) one obtains that $\partial^{\dot{R}}\partial^{\dot{S}}(\phi^2 E_{11\dot{R}\dot{S}}) = 0$, which implies that $E_{11\dot{R}\dot{S}}$ is of the form

$$E_{11\dot{R}\dot{S}} = \phi^{-2}\partial_{(\dot{R}}U_{\dot{S})}. \tag{3.3}$$

Thus, the first triple of Killing's equations, $E_{11\dot{R}\dot{S}} = 0$, has the solution $U_{\dot{S}} = -2(\alpha p_{\dot{S}} + \delta_{\dot{S}})$, where α and $\delta_{\dot{S}}$ are functions of $q^{\dot{A}}$ only.

Assuming that the equations $E_{11\dot{R}\dot{S}} = 0$ have been integrated, taking $A = B = C = 1$, $D = 2$ in Eq. (3.2) it follows that, $\partial^{\dot{R}}\partial^{\dot{S}}E_{12\dot{R}\dot{S}} = 3\sqrt{2}\phi^{-2}C_{1122}\partial_{\dot{S}}(\phi^2 K_1{}^{\dot{S}})$. Hence, in order to fulfill the second triple of Killing's equations, $E_{12\dot{R}\dot{S}} = 0$, it is necessary to have

$$C_{1122}\partial_{\dot{S}}(\phi^2 K_1{}^{\dot{S}}) = 0. \tag{3.4}$$

If the condition (3.4) is satisfied, then $E_{12\dot{R}\dot{S}}$ can be written in the form

$$E_{12\dot{R}\dot{S}} = \partial_{(\dot{R}}S_{\dot{S})}, \tag{3.5}$$

and the equations $E_{12\dot{R}\dot{S}} = 0$ imply that $S_{\dot{A}} = -\xi p_{\dot{A}} + \epsilon_{\dot{A}}$, where ξ and $\epsilon_{\dot{A}}$ are functions of $q^{\dot{A}}$ only.

Assuming now that $E_{11\dot{R}\dot{S}} = E_{12\dot{R}\dot{S}} = 0$, taking $A = B = 1$, $C = D = 2$ in Eq. (3.2) one finds that

$$\tfrac{1}{3}\phi^2\partial^{\dot{R}}\partial^{\dot{S}}(\phi^{-2}E_{22\dot{R}\dot{S}}) = 2\sqrt{2}\phi^{-2}C_{1222}\partial_{\dot{S}}(\phi^2 K_1{}^{\dot{S}}) + 2K^{\dot{R}\dot{S}}\partial_{\dot{R}\dot{S}}C_{1122} - 8\chi C_{1122}.$$

Hence, the last triple of Killing's equations require that

$$\sqrt{2}\phi^{-2}C_{1222}\partial_{\dot{S}}(\phi^2 K_1{}^{\dot{S}}) + K^{\dot{R}\dot{S}}\partial_{\dot{R}\dot{S}}C_{1122} - 4\chi C_{1122} = 0. \tag{3.6}$$

If the condition (3.6) is fulfilled, then $E_{22\dot{R}\dot{S}}$ can be written in the form

$$E_{22\dot{R}\dot{S}} = \phi^2\partial_{(\dot{R}}R_{\dot{S})}, \tag{3.7}$$

and the equations $E_{22\dot{R}\dot{S}} = 0$ have the solution $R_{\dot{A}} = -2(\nu p_{\dot{A}} + \Delta_{\dot{A}})$, where ν and $\Delta_{\dot{A}}$ are functions of $q^{\dot{A}}$ only.

Equation (3.2) gives two additional conditions. Assuming $E_{AB\dot{R}\dot{S}} = 0$, one of these is given by

$$0 = \nabla_{(1}{}^{\dot{R}}\nabla_{2}{}^{\dot{S}}E_{22)\dot{R}\dot{S}} = \sqrt{2}\phi^{-2}C_{2222}\partial_{\dot{S}}(\phi^2 K_1{}^{\dot{S}}) + \text{terms which vanish if } C_{1122} = C_{1222} = 0. \tag{3.8}$$

From Eqs. (3.4), (3.6), and (3.8) it follows that if $\partial_{\dot{S}}(\phi^2 K_1{}^{\dot{S}})$ is different from zero, then $C_{ABCD} = 0$. In the forthcoming, the discussion will be restricted to the case in which $\partial_{\dot{S}}(\phi^2 K_1{}^{\dot{S}}) = 0$.

A straightforward computation yields $E_{11}{}^{\dot{R}\dot{S}} = \sqrt{2}\phi^{-2}\partial^{\dot{R}}(\phi^2 K_1{}^{\dot{S}})$. Comparing with (3.3) one has, $U_{\dot{A}} = \sqrt{2}\phi^2 K_{1\dot{A}}$. Hence, the first triple of Killing's equations implies that $K_{1\dot{A}} = -\sqrt{2}\phi^{-2}(\alpha p_{\dot{A}} + \delta_{\dot{A}})$. Then, one finds, $\partial_{\dot{S}}(\phi^2 K_1{}^{\dot{S}}) = 2\sqrt{2}\alpha$; thus, if $C_{ABCD} \neq 0$,

$$K_{1\dot{A}} = -\sqrt{2}\phi^{-2}\delta_{\dot{A}}. \tag{3.9}$$

Similarly, by a direct computation one has $E_{12}{}^{\dot{R}\dot{S}} = (1/\sqrt{2})[\partial^{(\dot{R}}K_2{}^{\dot{S})} + D^{(\dot{R}}\phi^2 K_1{}^{\dot{S})} + \phi^2 K_1{}^{\dot{C}}\partial^{(\dot{R}}Q^{\dot{S})}{}_{\dot{C}}]$. Then, substituting Eq. (3.9) one obtains,

$$E_{12}{}^{\dot{R}\dot{S}} = \partial^{(\dot{R}}\left[(1/\sqrt{2})K_2{}^{\dot{S})} + \delta_{\dot{C}}Q^{\dot{S})\dot{C}} - \delta_{\dot{C}}{}^{,\dot{S})}p^{\dot{C}}\right],$$

where ${}_{,\dot{S}} \equiv \partial/\partial q^{\dot{S}}$. Therefore, the solution of the second triple of Killing's equations yields

$$K_{2\dot{A}} = \sqrt{2}(\delta^{\dot{C}}Q_{\dot{C}\dot{A}} + M_{\dot{A}}), \tag{3.10}$$

where, following Ref. 4,

$$M_{\dot{A}} \equiv \delta_{\dot{C},\dot{A}}\,p^{\dot{C}} - \xi p_{\dot{A}} + \epsilon_{\dot{A}}. \tag{3.11}$$

Hence, the Killing vector K has the form

$$K = M^{\dot{A}}\partial_{\dot{A}} + \delta^{\dot{A}}\frac{\partial}{\partial q^{\dot{A}}}, \tag{3.12}$$

and

$$\chi = -\tfrac{1}{2}\xi - K[\ln\phi], \tag{3.13}$$

where $K[f]$ denotes the directional derivative of the function f along K.

In order to complete the set of Killing's equations, one finds that $E_{22}{}^{\dot{R}\dot{S}} = \sqrt{2}\phi^2[D^{(\dot{R}}K_2{}^{\dot{S})} - \phi^2 K_1{}^{(\dot{R}}D_{\dot{C}}Q^{\dot{S})\dot{C}} + K_2{}^{\dot{C}}\partial_{\dot{C}}Q^{\dot{R}\dot{S}}]$. Then, using (3.9) and (3.10) one gets

$$\tfrac{1}{2}\phi^{-2}E_{22}{}^{\dot{R}\dot{S}} = K[Q^{\dot{R}\dot{S}}] + \xi Q^{\dot{R}\dot{S}} + 2\delta^{\dot{C},(\dot{R}}Q^{\dot{S})}{}_{\dot{C}} + M^{(\dot{R},\dot{S})}. \tag{3.14}$$

This expression can be written in the form given in (3.7) only after the condition (3.6) has been satisfied. Under the present assumptions, condition (3.6) reduces to

$$K[C_{1122}] = -2\chi C_{1122}, \tag{3.15a}$$

equivalently, in view of Eq. (3.13),

$$K[\phi^{-2}C_{1122}] = \xi\phi^{-2}C_{1122}. \tag{3.15b}$$

Thus the integration of the third triple of Killing's equations requires additional information concerning the form of the metric. However, in the rest of this section and in the next, some general relations will be established using the form of the Killing vector given by (3.12).

The Lie derivative along K of the members of the tetrad (2.4) are given by

$$\mathscr{L}_K\partial_{1\dot{A}} = [K,\partial_{1\dot{A}}] = (\xi\delta_{\dot{A}}^{\dot{C}} - \delta_{\dot{A}}{}^{,\dot{C}})\partial_{1\dot{C}} \tag{3.16a}$$

and

$$\mathscr{L}_K\partial_{2\dot{A}} = [K,\partial_{2\dot{A}}] = (2K[\ln\phi]\,\delta_{\dot{A}}^{\dot{C}} - \delta^{\dot{C}}{}_{,\dot{A}})\partial_{2\dot{C}} - \tfrac{1}{2}(E_{22\dot{A}}{}^{\dot{C}} + \phi^2 M^{\dot{R}}{}_{,\dot{R}}\delta_{\dot{A}}^{\dot{C}})\partial_{1\dot{C}}, \tag{3.16b}$$

where Eqs. (3.12) and (3.14) have been used. Equation (3.16a) means that the flow generated by K preserves the foliation of the space-time given by the null strings. This fact depends on the assumption that C_{ABCD} does not vanish.

To close this section, the transformation laws of the objects introduced in the integration of Killing's equations will be given. Since $\delta^{\dot{A}} = K[q^{\dot{A}}]$ and $M^{\dot{A}} = K[p^{\dot{A}}]$ [see Eq. (3.12)], from (2.5) and (2.6) one finds that if $q'^{\dot{A}}, p'^{\dot{A}}$ is another set of coordinates, then the components of K with respect to the basis induced by the primed coordinates are $\delta'^{\dot{A}} = K[q'^{\dot{A}}] = \delta^{\dot{C}} \partial q'^{\dot{A}} / \partial q^{\dot{C}}$, that is,

$$\delta'^{\dot{A}} = T^{\dot{A}}{}_{\dot{C}} \delta^{\dot{C}}, \tag{3.17}$$

and similarly

$$M'^{\dot{A}} = \rho^{-1} T_{\dot{C}}^{-1\dot{A}} M^{\dot{C}} + \delta^{\dot{C}} \frac{\partial p'^{\dot{A}}}{\partial q^{\dot{C}}}. \tag{3.18}$$

Expressing $M'_{\dot{A}}$ in an analogous form to that given for $M_{\dot{A}}$ in (3.11), using (3.17), (3.18), and (2.5) one gets

$$\xi' = \xi + \delta^{\dot{A}} (\ln \rho)_{,\dot{A}} \tag{3.19}$$

and

$$\begin{aligned} \epsilon'_{\dot{A}} = T^{-1\dot{B}}{}_{\dot{A}} [\rho^{-1} \epsilon_{\dot{B}} + \delta^{\dot{C}} (T^{\dot{D}}{}_{\dot{B}} \sigma_{\dot{D}})_{,\dot{C}} \\ + \delta^{\dot{C}}{}_{,\dot{B}} T^{\dot{D}}{}_{\dot{C}} \sigma_{\dot{D}}] + \xi' \sigma_{\dot{A}}. \end{aligned} \tag{3.20}$$

These transformation properties can be used to simplify the equations, for example, by eliminating ξ or $\epsilon_{\dot{A}}$, however, such possible simplification will not be elaborated here (for some applications of these transformations see Refs. 3 and 4).

IV. INTEGRABILITY CONDITIONS

It is shown in Ref. 9 that the integrability conditions for the Killing equations may be written as

$$\begin{aligned} L_{RST\dot{A}} \equiv & \nabla_{R\dot{A}} l_{ST} + 2K_{U\dot{A}} C^{U}{}_{RST} + 2K_{(R}{}^{\dot{U}} C_{ST)\dot{U}\dot{A}} \\ & + \tfrac{4}{3} \epsilon_{R(S} C_{T)U\dot{U}\dot{A}} K^{U\dot{U}} - (R/6) \epsilon_{R(S} K_{T)\dot{A}} \\ & + 2\epsilon_{R(S} \nabla_{T)\dot{A}} \chi = 0, \end{aligned} \tag{4.1}$$

$$M_{RSTU} \equiv K_{P\dot{A}} \nabla^{P\dot{A}} C_{RSTU} - 4\chi C_{RSTU} + 4 l_{V(R} C^{V}{}_{STU)} = 0,$$

and similar equations for the objects with dotted indices. These conditions restrict the form of the functions $\delta_{\dot{A}}$, ξ, and $\epsilon_{\dot{A}}$, which appear in the Killing vector [Eq. (3.12)]. Evidently, a direct computation of the conditions (4.1) would be very lengthy. However, this computation can be abbreviated by using the Ricci identities which give

$$L_{RST\dot{A}} = \nabla_{(R}{}^{\dot{B}} E_{ST)\dot{B}\dot{A}} - \tfrac{1}{3} \epsilon_{R(S} \nabla^{B\dot{C}} E_{T)B\dot{C}\dot{A}}, \tag{4.2a}$$

$$M_{RSTU} = \tfrac{1}{2} \nabla_{(R}{}^{\dot{A}} \nabla_{S}{}^{\dot{B}} E_{TU)} - C_{(RS}{}^{\dot{A}\dot{B}} E_{TU)\dot{A}\dot{B}} \tag{4.2b}$$

[cf., Eq. (3.2)].

The main simplification that is obtained by determining the integrability conditions by means of (4.2) comes from the fact that, with K given by (3.12), $E_{11\dot{A}\dot{B}}$ and $E_{12\dot{A}\dot{B}}$ are zero. Hence, one finds that

$$L_{111\dot{A}} = L_{112\dot{A}} = L_{211\dot{A}} = 0,$$

$$L_{122\dot{A}} = \sqrt{2} E_{22\dot{A}\dot{B}} \partial^{\dot{B}} \ln \phi,$$

$$L_{212\dot{A}} = (1/\sqrt{2}) \phi^{4} \partial^{\dot{B}} (\phi^{-4} E_{22\dot{A}\dot{B}}),$$

$$L_{222\dot{A}} = \sqrt{2} \phi^{2} \{ \phi^{3} D^{\dot{B}} (\phi^{-3} E_{22\dot{A}\dot{B}}) - E_{22}{}^{\dot{B}\dot{C}} \partial_{\dot{B}} Q_{\dot{C}\dot{A}} \},$$

$$M_{1111} = M_{1112} = 0,$$

$$\begin{aligned} M_{1122} &= \tfrac{1}{6} \phi^{2} \partial^{\dot{A}} \partial^{\dot{B}} (\phi^{-2} E_{22\dot{A}\dot{B}}), \\ M_{1222} &= \tfrac{1}{2} \phi^{2} \{ \phi^{2} D^{\dot{A}} [\phi^{2} \partial^{\dot{B}} (\phi^{-4} E_{22\dot{A}\dot{B}})] \\ &\quad + 2 [\phi^{3} D^{\dot{B}} (\phi^{-3} E_{22\dot{A}\dot{B}}) \\ &\quad - E_{22}{}^{\dot{B}\dot{C}} \partial_{\dot{B}} Q_{\dot{C}\dot{A}}] \partial^{\dot{A}} \ln \phi \} - C_{12}{}^{\dot{A}\dot{B}} E_{22\dot{A}\dot{B}}, \\ M_{2222} &= (1/\sqrt{2}) \phi^{2} [\phi^{3} D^{\dot{A}} (\phi^{-3} L_{222\dot{A}}) + L_{222\dot{A}} \partial_{\dot{C}} Q^{\dot{C}\dot{A}}] \\ &\quad - \phi^{7} (D_{\dot{C}} Q^{\dot{C}\dot{A}}) \partial^{\dot{B}} (\phi^{-3} E_{22\dot{A}\dot{B}}) - C_{22}{}^{\dot{A}\dot{B}} E_{22\dot{A}\dot{B}}. \end{aligned} \tag{4.3}$$

Substituting the expression for $E_{22\dot{A}\dot{B}}$ given by (3.14) and using the commutation relations (3.16) one obtains

$$\begin{aligned} (1/\sqrt{2}) L_{212\dot{A}} = & - E_{22\dot{A}\dot{B}} \partial^{\dot{B}} \ln \phi + \phi^{2} \{ K [\partial^{\dot{B}} Q_{\dot{A}\dot{B}}] \\ & + \delta^{\dot{B}}{}_{,\dot{B}\dot{A}} + \tfrac{3}{2} \xi_{,\dot{A}} + \delta_{\dot{C},\dot{A}} \partial_{\dot{B}} Q^{\dot{B}\dot{C}} \}, \end{aligned} \tag{4.4a}$$

$$\begin{aligned} (1/\sqrt{2}) L_{222\dot{A}} = & - E_{22\dot{A}\dot{B}} \phi D^{\dot{B}} \phi + \phi^{4} \{ 2K [D^{\dot{B}} Q_{\dot{A}\dot{B}}] \\ & + 2(\delta^{\dot{C}}{}_{,\dot{C}} + \xi) D^{\dot{B}} Q_{\dot{A}\dot{B}} + 2\delta_{\dot{C},\dot{A}} D_{\dot{B}} Q^{\dot{B}\dot{C}} \\ & - (\xi_{,\dot{C}} p^{\dot{C}} - \epsilon^{\dot{C}}{}_{,\dot{C}}) \partial^{\dot{B}} Q_{\dot{A}\dot{B}} \\ & + \xi^{,\dot{B}} Q_{\dot{A}\dot{B}} + \xi_{,\dot{A}\dot{C}} p^{\dot{C}} - \epsilon^{\dot{C}}{}_{,\dot{A}\dot{C}} \}, \end{aligned} \tag{4.4b}$$

while the condition $M_{1122} = 0$ amounts to Eq. (3.15b).

V. APPLICATIONS

The results obtained in the previous sections can be applied directly to find the Killing vectors of any given real metric which admits a shear-free congruence of null geodesics defined by a repeated Debever–Penrose vector or any complex metric which admits a congruence of null strings defined by a repeated Debever–Penrose spinor. Following the procedure outlined in Ref. 2, the given metric can be brought to the form (2.2); which allows one to identify the coordinates $q^{\dot{A}}, p^{\dot{A}}$ and the structural functions ϕ and $Q_{\dot{A}\dot{B}}$. Then the use of the integrability conditions and Eq. (3.14) leads to the explicit form of the allowed Killing vectors.

Conversely, one can obtain the general form of the metric which admits a given Killing vector. In view of Eq. (3.17), the Killing vectors admitted by this class of metrics can be classified according to whether $\delta_{\dot{A}}$ vanishes or not. In the case where $\delta_{\dot{A}}$ is nonzero, by an appropriate change of coordinates, it can be made equal to a constant, e.g., $\delta_{\dot{A}} = \delta^{\dot{1}}_{\dot{A}}$ and simultaneously, ξ and $\epsilon_{\dot{A}}$ can be made equal to zero [see Eqs. (3.17), (3.19)–(3.20)]. Then the last triple of Killing's equations [Eq. (3.14)] amounts to $Q_{\dot{A}\dot{B},\dot{2}} = 0$ and from (3.13) it follows that $\chi = -(\ln \phi)_{,\dot{2}}$.

When $\delta_{\dot{A}}$ is zero, the function ξ is an invariant. In the subcase where ξ does not vanish, $\epsilon_{\dot{A}}$ can be eliminated and from (3.14) and (3.13) it follows that $Q_{\dot{A}\dot{B}} = P_{\dot{A}\dot{B}} - f (\ln \xi)_{,(\dot{A}} p_{\dot{B})}$, where each $P_{\dot{A}\dot{B}}$ is a homogeneous function of $p^{\dot{R}}$ of order 1 and f is a function such that $p^{\dot{A}} \partial_{\dot{A}} f = 1$, e.g., $f = \frac{1}{2} \ln[(p^{\dot{1}})^2 + (p^{\dot{2}})^2]$, and $\chi = \xi (p^{\dot{A}} \partial_{\dot{A}} \ln \phi - \frac{1}{2})$. Finally, in the subcase where ξ is zero, $\epsilon_{\dot{A}}$ can be made equal to a constant, e.g., $\epsilon_{\dot{A}} = \delta^{\dot{1}}_{\dot{A}}$. Then from (3.14) and (3.13) one obtains $\partial_{\dot{2}} Q_{\dot{A}\dot{B}} = 0$ and $\chi = -\partial_{\dot{2}} \ln \phi$. In general, these metrics will be complex.

In the remaining part of this section it is shown that in the case where the congruence of null strings has a vanishing expansion, i.e., $l^{A} \nabla_{B\dot{C}} l_{A} = 0$, the Killing equations can be reduced to a single master equation, assuming that the scalar curvature is a constant.

As a consequence of the condition $l^A\nabla_{B\dot{C}}l_A = 0$, ϕ can be chosen as equal to one [see Eq. (2.3)] and, with respect to the basis (2.4), $C_{11\dot{A}\dot{B}} = 0$. Therefore, in order to satisfy the Einstein field equations, the trace of the energy-momentum tensor of the matter must be a constant and, denoting by $T_{AB\dot{C}\dot{D}}$ the spinorial components of the traceless part of the energy-momentum tensor referred to the basis (2.4), the components $T_{11\dot{A}\dot{B}}$ must vanish. Then the remaining components are of the form[2]

$$T_{12\dot{A}\dot{B}} = (1/8\pi)\partial_{\dot{A}}\partial_{\dot{B}}T_0,$$
$$T_{22\dot{A}\dot{B}} = -(1/8\pi)\partial_{(\dot{A}}T_{\dot{B})}, \tag{5.1}$$

where T_0 and $T_{\dot{A}}$ are some functions made out of the matter field, and $Q_{\dot{A}\dot{B}}$ is given by

$$Q_{\dot{A}\dot{B}} = -\partial_{\dot{A}}\partial_{\dot{B}}\Theta - \partial_{(\dot{A}}G_{\dot{B})} - \tfrac{2}{3}L_{(\dot{A}}\,p_{\dot{B})} - (R/12)p_{\dot{A}}\,p_{\dot{B}}, \tag{5.2}$$

where R denotes the scalar curvature (assumed constant), $L_{\dot{A}} = L_{\dot{A}}(q^{\dot{R}})$, $G_{\dot{A}}$ is such that

$$\partial_{\dot{A}}G^{\dot{A}} = 4T_0, \tag{5.3}$$

and Θ is a function that must satisfy a single second-order partial differential equation with quadratic nonlinearities.[2]

In the case under consideration, $C_{1122} = R/12 =$ constant, therefore condition (3.15a) amounts to

$$R\chi = 0. \tag{5.4}$$

Notice that with $\phi = 1$, from (3.13) it follows that $\xi = -2\chi$. After imposing (5.4) one can write $E_{22\dot{R}\dot{S}}$ in the form given in (3.7). Substituting (5.2) into (3.14) and using (3.11), (3.13), and (3.16a) one finds that $R_{\dot{A}}$ can be taken as

$$\begin{aligned}
-\tfrac{1}{2}R_{\dot{A}} = {} & \partial_{\dot{A}}\{K[\Theta] + (3\xi + 2\delta^{\dot{R}}{}_{,\dot{R}})\Theta\} + K[G_{\dot{A}}] \\
& + (2\xi + \delta^{\dot{R}}{}_{,\dot{R}})G_{\dot{A}} + \delta^{\dot{C}}{}_{,\dot{A}}G_{\dot{C}} \\
& + (\tfrac{2}{3}L_{\dot{R}}\delta^{\dot{R}}{}_{,\dot{C}} + \tfrac{2}{3}\delta^{\dot{R}}L_{\dot{C},\dot{R}} \\
& + \xi_{,\dot{C}} + (R/6)\epsilon_{\dot{C}} + \tfrac{1}{2}\delta^{\dot{R}}{}_{,\dot{R}\dot{C}})p^{\dot{C}}p_{\dot{A}} - \tfrac{1}{2}\delta_{\dot{C},\dot{D}\dot{A}}\,p^{\dot{C}}p^{\dot{D}} \\
& + (\tfrac{2}{3}L_{\dot{C}}\epsilon_{\dot{A}} - \epsilon_{\dot{C},\dot{A}})p^{\dot{C}} + \tfrac{2}{3}L_{\dot{C}}\epsilon^{\dot{C}}p_{\dot{A}}.
\end{aligned} \tag{5.5}$$

Then the last triple of Killing's equations gives $R_{\dot{A}} = -2(\nu p_{\dot{A}} + \Delta_{\dot{A}})$, where ν and $\Delta_{\dot{A}}$ are functions of $q^{\dot{R}}$ only. Since $\partial^{\dot{A}}\partial_{\dot{A}} = 0$ and using (3.16a) on evaluating $\partial^{\dot{A}}R_{\dot{A}}$, one obtains

$$\begin{aligned}
& 4K[T_0] + 4(\xi + \delta^{\dot{R}}{}_{,\dot{R}})T_0 + (2L_{\dot{R}}\delta^{\dot{R}}{}_{,\dot{A}} + 2\delta^{\dot{R}}L_{\dot{A},\dot{R}} \\
& \quad + 3\xi_{,\dot{A}} + (R/2)\epsilon_{\dot{A}} + 2\delta^{\dot{R}}{}_{,\dot{R}\dot{A}})p^{\dot{A}} \\
& \quad - 2L^{\dot{C}}\epsilon_{\dot{C}} - \epsilon_{\dot{C}}{}^{,\dot{C}} = 2\nu.
\end{aligned} \tag{5.6}$$

Hence, in order for K to be a Killing vector, the matter field must satisfy the constraint

$$\partial^{\dot{C}}\partial^{\dot{B}}K[T_0] + (\xi + \delta^{\dot{R}}{}_{,\dot{R}})\partial^{\dot{C}}\partial^{\dot{B}}T_0 = 0. \tag{5.7}$$

Writing Eq. (5.6) in the form

$$\begin{aligned}
& \partial^{\dot{A}}\{K[G_{\dot{A}}] + (2\xi + \delta^{\dot{R}}{}_{,\dot{R}})G_{\dot{A}} + \delta^{\dot{C}}{}_{,\dot{A}}G_{\dot{C}} + \tfrac{1}{3}(2L_{\dot{R}}\delta^{\dot{R}}{}_{,\dot{C}} \\
& \quad + 2\delta^{\dot{R}}L_{\dot{C},\dot{R}} + 3\xi_{,\dot{C}} + (R/2)\epsilon_{\dot{C}} + 2\delta^{\dot{R}}{}_{,\dot{R}\dot{C}})p^{\dot{C}}p_{\dot{A}} \\
& \quad - (L^{\dot{C}}\epsilon_{\dot{C}} + \tfrac{1}{2}\epsilon_{\dot{C}}{}^{,\dot{C}} + \nu)p_{\dot{A}}\} = 0,
\end{aligned}$$

it follows that there exists a function Λ such that the expression between braces is equal to $-\partial_{\dot{A}}\Lambda$. Thus, substituting into (5.5) one gets

$$\begin{aligned}
& \partial_{\dot{A}}\{K[\Theta] + (3\xi + 2\delta^{\dot{R}}{}_{,\dot{R}})\Theta - \Delta_{\dot{B}}\,p^{\dot{B}} \\
& \quad - (\tfrac{1}{2}\epsilon_{\dot{B},\dot{C}} - \tfrac{1}{3}L_{\dot{B}}\epsilon_{\dot{C}})p^{\dot{B}}p^{\dot{C}} - \tfrac{1}{6}\delta_{\dot{B},\dot{C}\dot{D}}\,p^{\dot{B}}p^{\dot{C}}p^{\dot{D}} - \Lambda\} = 0
\end{aligned}$$

from which it follows that the directional derivative of the key function Θ along any Killing vector K must obey the master equation

$$\begin{aligned}
K[\Theta] = {} & -(3\xi + 2\delta^{\dot{R}}{}_{,\dot{R}})\Theta + \Lambda + \tfrac{1}{6}\delta_{\dot{A},\dot{B}\dot{C}}\,p^{\dot{A}}p^{\dot{B}}p^{\dot{C}} \\
& + (\tfrac{1}{2}\epsilon_{\dot{A},\dot{B}} - \tfrac{1}{3}L_{\dot{A}}\epsilon_{\dot{B}})p^{\dot{A}}p^{\dot{B}} + \Delta_{\dot{A}}\,p^{\dot{A}} + \eta,
\end{aligned} \tag{5.8}$$

where η is a function of $q^{\dot{R}}$ only.

Consideration of the integrability condition leads to some additional relations. From (5.2) one has

$$\partial_{\dot{B}}Q^{\dot{A}\dot{B}} = -2\partial^{\dot{A}}T_0 - L^{\dot{A}} - (R/4)p^{\dot{A}}. \tag{5.9}$$

Substituting this expression into (4.4a) and using (3.16a) one finds that $L_{212\dot{A}}$ vanishes if and only if Eqs. (5.4) and (5.7) are satisfied. Regarding the conditions $L_{222\dot{A}} = 0$, from Eqs. (3.11b), (3.32), and (3.34b) of Ref. 2 one gets

$$D_{\dot{B}}Q^{\dot{A}\dot{B}} = T^{\dot{A}} + N^{\dot{A}} + \tfrac{1}{2}(L_{\dot{B}}{}^{,\dot{B}} + \partial^{\dot{B}}T_{\dot{B}} + 2D^{\dot{B}}\partial_{\dot{B}}T_0)p^{\dot{A}}, \tag{5.10}$$

where $N_{\dot{A}} = N_{\dot{A}}(q^{\dot{R}})$. Then, using (5.9) and (5.2) in (4.4b) one obtains

$$\begin{aligned}
-\frac{1}{\sqrt{2}}L_{222\dot{A}} = {} & 2K[T_{\dot{A}}] + 2(\delta^{\dot{B}}{}_{,\dot{B}} + \xi)T_{\dot{A}} - 2\delta_{\dot{B},\dot{A}}T^{\dot{B}} \\
& + 2(\xi p^{\dot{B}} - \epsilon^{\dot{B}})_{,\dot{B}}\partial_{\dot{A}}T_0 \\
& - \xi_{,\dot{A}\dot{B}}\,p^{\dot{B}} + \tfrac{2}{3}L_{(\dot{A}}\xi_{,\dot{B})}\,p^{\dot{B}} \\
& + \xi^{,\dot{B}}(\partial_{\dot{A}}\partial_{\dot{B}}\Theta + \partial_{(\dot{A}}G_{\dot{B})}) \\
& + [\delta^{\dot{C}}(L_{\dot{B}}{}^{,\dot{B}} + \partial^{\dot{B}}T_{\dot{B}} + 2D^{\dot{B}}\partial_{\dot{B}}T_0) \\
& - (R/4)\epsilon^{\dot{C}}]_{,\dot{C}}\,p_{\dot{A}} \\
& + 2(\delta^{\dot{B}}N_{\dot{A}})_{,\dot{B}} + 2\xi N_{\dot{A}} - 2\delta_{\dot{B},\dot{A}}N^{\dot{B}} + \epsilon^{\dot{B}}{}_{,\dot{A}\dot{B}} \\
& - \epsilon^{\dot{B}}{}_{,\dot{B}}L_{\dot{A}} + (L_{\dot{B}}{}^{,\dot{B}} + \partial^{\dot{B}}T_{\dot{B}} \\
& + 2D^{\dot{B}}\partial_{\dot{B}}T_0)\epsilon_{\dot{A}}.
\end{aligned} \tag{5.11}$$

Thus, the conditions $L_{222\dot{A}} = 0$ require

$$\begin{aligned}
& \partial_{(\dot{A}}K[T_{\dot{B})}] + (\delta^{\dot{C}}{}_{,\dot{C}} + \xi)\partial_{(\dot{A}}T_{\dot{B})} - \delta_{\dot{C},(\dot{A}}\partial_{\dot{B})}T^{\dot{C}} \\
& \quad + (\xi p^{\dot{C}} - \epsilon^{\dot{C}})_{,\dot{C}}\partial_{\dot{A}}\partial_{\dot{B}}T_0 - \tfrac{1}{2}\{\xi_{,\dot{A}\dot{B}} \\
& \quad - \xi^{,\dot{C}}(\partial_{\dot{A}}\partial_{\dot{B}}\partial_{\dot{C}}\Theta + \partial_{\dot{C}}\partial_{(\dot{A}}G_{\dot{B})}) - \tfrac{2}{3}L_{(\dot{A}}\xi_{,\dot{B})}\} = 0,
\end{aligned} \tag{5.12}$$

which implies the existence of a and $b_{\dot{A}}$, functions of $q^{\dot{R}}$ only such that

$$\begin{aligned}
& 2K[T_{\dot{A}}] + 2(\delta^{\dot{B}}{}_{,\dot{B}} + \xi)T_{\dot{A}} - 2\delta_{\dot{B},\dot{A}}T^{\dot{B}} \\
& \quad + 2(\xi p^{\dot{B}} - \epsilon^{\dot{B}})_{,\dot{B}}\partial_{\dot{A}}T_0 - \xi_{,\dot{A}\dot{B}}\,p^{\dot{B}} + \tfrac{2}{3}L_{(\dot{A}}\xi_{,\dot{B})}\,p^{\dot{B}} \\
& \quad + \xi^{,\dot{B}}(\partial_{\dot{A}}\partial_{\dot{B}}\Theta + \partial_{(\dot{A}}G_{\dot{B})}) = ap_{\dot{A}} + b_{\dot{A}}.
\end{aligned} \tag{5.13}$$

Comparing (5.11) and (5.13) one concludes that

$$\begin{aligned}
b_{\dot{A}} = {} & -2(\delta^{\dot{B}}N_{\dot{A}})_{,\dot{B}} - 2\xi N_{\dot{A}} + 2\delta_{\dot{B},\dot{A}}N^{\dot{B}} - \epsilon^{\dot{B}}{}_{,\dot{A}\dot{B}} \\
& + \epsilon^{\dot{B}}{}_{,\dot{B}}L_{\dot{A}} - (L_{\dot{B}}{}^{,\dot{B}} + \partial^{\dot{B}}T_{\dot{B}} + 2D^{\dot{B}}\partial_{\dot{B}}T_0)\epsilon_{\dot{A}}
\end{aligned} \tag{5.14}$$

and

$$\begin{aligned}
a & = -[\delta^{\dot{C}}(L_{\dot{B}}{}^{,\dot{B}} + \partial^{\dot{B}}T_{\dot{B}} + 2D^{\dot{B}}\partial_{\dot{B}}T_0) - (R/4)\epsilon^{\dot{C}}]_{,\dot{C}} \\
& = K[\partial_{\dot{B}}T^{\dot{B}}] + \delta^{\dot{C}}{}_{,\dot{C}}\partial_{\dot{B}}T^{\dot{B}},
\end{aligned} \tag{5.15}$$

where Eq. (5.6) has been used [this last equality also follows from (5.13)].

It is perhaps worthwhile to point out that for two of the three possible kinds of Killing vectors found at the beginning of this section, there exists a symplectic structure such that K is a Hamiltonian vector field and the null strings are Lagrangian submanifolds. Indeed, when $\delta_{\dot{A}} \neq 0$ there exist coordinates such that $K = \delta_{\dot{C}}{}^{\dot{A}} p^{\dot{C}} \partial_{\dot{A}} + \delta_{\dot{A}} \partial / \partial q^{\dot{A}}$. Defining $\omega \equiv dp_{\dot{A}} \wedge dq^{\dot{A}}$, one sees that the inner product of K and ω gives $K \lrcorner \omega = d(\delta_{\dot{C}}\, p^{\dot{C}})$. In the case where $\delta_{\dot{A}} = 0 = \dot{\xi}$, there exist coordinates such that $K = \epsilon_0{}^{\dot{A}} \partial_{\dot{A}}$, with $\epsilon_0{}^{\dot{A}}$ constant. Then $K \lrcorner (dp_{\dot{A}} \wedge dq^{\dot{A}}) = d(\epsilon_{0\dot{A}} q^{\dot{A}})$. Clearly, in both cases, $dp_{\dot{A}} \wedge dq^{\dot{A}}$ is a symplectic two-form which vanishes on the tangent space of the null strings.

ACKNOWLEDGMENTS

The author wishes to express his gratitude to Professor J. D. Finley, III, for very valuable discussions and to Professor J. F. Plebański for helpful suggestions and comments.

[1]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). For a spinorial derivation see J. D. Finley, III, and J. F. Plebański, J. Math. Phys. **17**, 2207 (1976).

[2]G. F. Torres del Castillo, J. Math. Phys. **24**, 590 (1983).

[3]J. D. Finley, III, and J. F. Plebański, J. Math. Phys. **19**, 760 (1978).

[4]S. A. Sonnleitner and J. D. Finley, III, J. Math. Phys. **23**, 116 (1982).

[5]J. F. Plebański, J. Math. Phys. **16**, 2395 (1975).

[6]The vector fields $\partial_{A\dot{B}}$ form a null tetrad; their scalar products are given by $\partial_{A\dot{B}} \cdot \partial_{C\dot{D}} = -2\epsilon_{AC}\epsilon_{\dot{B}\dot{D}}$.

[7]The one-forms $g^{A\dot{B}}$ are defined by $g^{A\dot{B}}(\partial_{C\dot{D}}) = -2\delta^A_C \delta^{\dot{B}}_{\dot{D}}$.

[8]The null tetrad (2.4) does not satisfy the hermiticity condition, $\overline{\partial_{A\dot{B}}} = \partial_{B\dot{A}}$. Therefore, with respect to this basis, the dotted components are not the complex conjugates of the corresponding undotted ones.

[9]S. Hacyan and J. F. Plebański, J.Math. Phys. **17**, 2203 (1976).

Coupled electromagnetic and gravitational linearized perturbations of the Kerr–Newman metric

V. Bellezza and V. Ferrari
Istituto di Fisica "G. Marconi," Università di Roma, Rome, Italy

(Received 7 February 1983; accepted for publication 27 January 1984)

A new gauge and tetrad invariant equation governing the coupled electromagnetic-gravitational perturbations for the Kerr–Newman metric and involving a closed 2-form is obtained. Operating on this equation with the complex vectorial formalism as developed by Crossman, we have found three equations for the gauge and tetrad invariant quantities $\chi_{1B}(\equiv 3\psi_{2A}\phi_{0B} - 2\phi_{1A}\psi_{1B})$, $\chi_{-1B}(\equiv 3\psi_{2A}\phi_{2B} - 2\phi_{1A}\psi_{3B})$, and $\chi_{0B}(\equiv 3\psi_{2A}\phi_{1B} - 2\phi_{1A}\psi_{2B} + 2\phi_{1A}^{5/2}\bar{\phi}_{1A}^{-1/2}\bar{\phi}_{1B})$.

PACS numbers: 04.40. + c, 97.60.Lf

I. INTRODUCTION

The main problem in analyzing small perturbations of a Kerr–Newman solution arises from the coupling between the electromagnetic and the gravitational fields. This problem does not arise for the Kerr solution: Teukolsky,[1] using the Newman–Penrose formalism, has found a "master" equation governing scalar, electromagnetic, and gravitational perturbations in the given background. This equation is invariant under infinitesimal tetrad rotations (tetrad invariant) and under infinitesimal coordinate transformations (gauge invariant). Teukolsky's equation is satisfied independently by four gauge and tetrad invariant perturbed quantities: the Weyl scalars ψ_{0B} and ψ_{4B} and the Maxwell scalars ϕ_{0B} and ϕ_{2B} (the subscript capital B denotes first-order perturbed quantities). Moreover it is a separable equation in the Boyer–Lindquist coordinates, i.e., the radial and the angular behavior of the perturbation is described by two different equations obtained from the master equation by means of a separation of variables.

In the following we will call an equation satisfied by only one perturbed quantity a "decoupled equation." In this sense, for example, Teukolsky's equation is a decoupled equation for $\phi_{0B}, \phi_{1B}, \psi_{0B}, \psi_{4B}$.

In the Kerr–Newman solution, the electromagnetic perturbations are coupled to the gravitational perturbations via the energy-momentum tensor of the electromagnetic field in Einstein's field equations and via the term $(-g)^{1/2}$ which appears in Maxwell's equations, as written in the usual tensorial form $[(-g)^{1/2}. F^{\mu\nu}]_{,\nu} = 0$.

Until now attempts to find decoupled and separable equations for gauge and tetrad invariant perturbed quantities have been unsuccessful. Many efforts have been made in this direction following different approaches and formalisms: Chandrasekhar[2,3] using the Newman–Penrose formalism has obtained a set of coupled equations for the gauge invariant quantities

$$\Phi_0 = \psi_{0B}, \quad \Phi_1 = \psi_{1B}\,\rho\sqrt{2},$$
$$K = k_B/(\rho)^2\sqrt{2}, \quad S = \sigma_B\bar{\rho}/(\rho)^2,$$

where $\rho = r - ia\cos\theta$ and k_B and σ_B are two perturbed spin coefficients (from now on a bar means complex conjugation).

Lee[4] has derived a set of equations that couple the purely gravitational quantities ψ_{0B} and ψ_{4B} to the variables

$$\chi_{1B} = 3\psi_{2A}\phi_{0B} - 2\phi_{1A}\psi_{1B}$$
$$\chi_{-1B} = 3\psi_{2A}\phi_{2B} - 2\phi_{1A}\psi_{3B}.$$

$\psi_{0B}, \psi_{4B}, \chi_{1B}, \chi_{-1B}$ are gauge and tetrad invariant. In a recent paper Lee[5] has shown that in a limiting case where both the charge and the angular momentum of the black hole are small ($a \ll M, e \ll M$) it is possible to obtain decoupled and separable equations for χ_{1B} and χ_{-1B}. Crossman,[6] using the gauge and tetrad freedom, has derived decoupled equations for $\phi_{0B}, \phi_{1B}, \phi_{2B}$. The equations for ϕ_{0B} and ϕ_{2B} are separable.

A different approach has been adopted by Fackerell.[7] Using the Cahen–Debever–Defrise[8] complex vectorial formalism as developed by Crossman[9] he has found a gauge and tetrad invariant equation for electromagnetic and gravitational perturbations. This equation has the form $dA = 0$, where A is a 2-form. Operating on this equation with the generalized operators of exterior differentiation introduced by Crossman, Fackerell, and Crossman[10] have found a set of decoupled equations for the gravitational quantities ψ_{1B}, ψ_{2B}, and ψ_{3B}, using the gauge and tetrad freedom. The equations for ψ_{1B} and ψ_{3B} are separable. However, the gauge which allows the decoupling of the equations for the electromagnetic quantities $\phi_{0B}, \phi_{2B}, \phi_{1B}$ is different from that which allows the decoupling of the equations for the gravitational perturbations ψ_{1B}, ψ_{2B}, and ψ_{3B}.

In this paper, using the complex vectorial formalism, we derive a new equation governing the coupled electromagnetic-gravitational perturbations. It is both tetrad and gauge independent and it has the form $dA = 0$, where A is a 2-form.[11] Following Fackerell we call this equation a "conservation" equation in analogy with the conservation of charge in electrodynamics that is stated locally by the law $d\,{}^*J = 0$, where J is the charge current form.[12] Jordan, Ehlers, and Sachs[13] have obtained a conservation equation valid for all uncharged type D geometries

$$d(\psi_{2A}^{2/3} Z_A^2) = 0.$$

They have shown that there are conserved quantities associated to this equation (for example in the case of the Schwarzschild metric this equation leads to the definition of the covariant mass). However, the problem of finding conserved quantities associated to our conservation equation will not be treated in this paper, whose aim is to investigate the possibility of finding gauge and tetrad invariant, decoupled and separable equations using the complex vectorial formalism and the Crossman generalized operators.

Operating on our conservation equation with the Crossman operators, we obtain three gauge and tetrad invariant equations for χ_{1B}, χ_{-1B}, and χ_{0B}, where we have defined $\chi_{0B} = 3\psi_{2A}\phi_{1B} - 2\phi_{1A}\psi_{2B} + 2\phi_{1A}^{5/2}\bar{\phi}_{1A}^{-1/2}\bar{\phi}_{1B}$.

These equations are not separable in the Boyer–Lindquist coordinates for two reasons: The first is that some operators appearing in our coupled equations are not purely radial or purely angular operators. The second is that here the situation is quite different from the Reissner–Nordström perturbations. In that case Chandrasekhar[14] has shown that it is possible to separate a set of coupled equations for Φ_0, Φ_1, K, S, by virtue of known relations existing among the angular functions belonging to the spin-1 and the spin-2 fields. In the case of axial symmetry, on the contrary, the angular functions belonging to different spins are not simply related.

The equations for χ_{1B} and χ_{-1B} in the limiting case of small charge and angular momentum, reduce to the corresponding ones obtained by Lee.

It is interesting to note that the quantities $\chi_{1B}, \chi_{-1B}, \chi_{0B}$ are gauge and tetrad invariant linear combinations of the quantities $(\phi_{0B}, \phi_{1B}, \phi_{2B})$ and $(\psi_{1B}, \psi_{2B}, \psi_{3B})$. Therefore, the set of equations we have derived includes, in a gauge and tetrad invariant formulation, the two sets of equations that Crossman and Fackerell have obtained in two different gauges for the perturbed Weyl scalars and for the perturbed Maxwell scalars. The reason we are interested in equations of the form $dA = 0$ is that the formalism developed by Crossman, applied to these equations, automatically leads to decoupled and separable equations for the components of A, if A is a self-dual 2-form (see Sec. V). The requirement that A be self-dual must be satisfied if we do not want to use the gauge and the tetrad freedom, but it is a difficult condition to satisfy because in the Kerr–Newman geometry, both the Maxwell 2-form and the curvature 2-forms have anti-self-dual components, in contrast with the Kerr case where, for example, the Maxwell 2-form is self-dual. In Sec. II of this paper we shall introduce the complex vectorial formalism in the Debever later version and the Crossman generalized operators. In Sec. III we derive our conservation equation and in Sec. IV we discuss its invariance under tetrad rotation and coordinate transformations. In Sec. V the equations for $\chi_{1B}, \chi_{-1B}, \chi_{0B}$ are derived.

II. COMPLEX VECTORIAL FORMALISM

We consider a null tetrad $\{l^\mu, n^\mu, m^\mu, \bar{m}^\mu\}$ satisfying the usual conditions

$$l_\mu n^\mu = -m_\mu \bar{m}^\mu = -1, \quad l_\mu m^\mu = n_\mu m^\mu = 0,$$

and

$$g^{\mu\nu} = l^\mu n^\nu + n^\mu l^\nu - m^\mu \bar{m}^\nu - \bar{m}^\mu m^\nu.$$

The real vectors l^μ and n^μ are chosen along the repeated principal null directions of the Weyl tensor. Following Debever we write the basis 1-forms

$$\theta^1 = n_\mu\, dx^\mu, \quad \theta^2 = l_\mu\, dx^\mu,$$
$$\theta^3 = -\bar{m}_\mu\, dx^\mu, \quad \theta^4 = \bar{\theta}^3$$

and the self-dual basis for the two-forms

$$Z^1 = \theta^1 \wedge \theta^3, \quad Z^2 = \theta^1 \wedge \theta^2 - \theta^3 \wedge \theta^4,$$
$$Z^3 = \theta^4 \wedge \theta^2,$$

where self-dual means that their Hodge duals identically fulfills

$$*Z^a = iZ^a.$$

Let $\sigma_1, \sigma_2, \sigma_3$ be the spinor connection 1-forms expressed in terms of the spin coefficients

$$\begin{aligned} \sigma_1 &= k\theta^1 + \tau\theta^2 + \sigma\theta^3 + \rho\theta^4, \\ \sigma_2 &= \epsilon\theta^1 + \gamma\theta^2 + \beta\theta^3 + \alpha\theta^4, \\ \sigma_3 &= \pi\theta^1 + \nu\theta^2 + \mu\theta^3 + \lambda\theta^4. \end{aligned} \tag{2.1}$$

Then, the first equations of structure are

$$\begin{aligned} dZ^1 &= -2\sigma_2 \wedge Z^1 - \sigma_3 \wedge Z^2, \\ dZ^2 &= 2\sigma_1 \wedge Z^1 - 2\sigma_3 \wedge Z^3, \\ dZ^3 &= 2\sigma_2 \wedge Z^3 + \sigma_1 \wedge Z^2, \end{aligned} \tag{2.2}$$

and the second equations of structure are

$$\begin{aligned} \Sigma_1 &= d\sigma_1 - 2\sigma_2 \wedge \sigma_1, \\ \Sigma_2 &= d\sigma_2 + \sigma_1 \wedge \sigma_3, \\ \Sigma_3 &= d\sigma_3 + 2\sigma_2 \wedge \sigma_3, \end{aligned} \tag{2.3}$$

where Σ_a are the complex curvature two-forms

$$\Sigma_a = C_{ab} Z^b + \tfrac{1}{6} R \gamma_{ab} Z^b + E_{a\bar{b}} \bar{Z}^b$$

C_{ab} and $E_{a\bar{b}}$ are, respectively, the Weyl and the Ricci tensor projected on the null tetrad, R is the Ricci curvature scalar, and γ_{ab} is the metric of the tridimensional complex space of the 2-forms

$$C_{ab} = \begin{bmatrix} \Psi_0 & \Psi_1 & \Psi_2 \\ \Psi_1 & \Psi_2 & \Psi_3 \\ \Psi_2 & \Psi_3 & \Psi_4 \end{bmatrix}, \quad E_{ab} = \begin{bmatrix} \Phi_{00} & \Phi_{01} & \Phi_{02} \\ \Phi_{10} & \Phi_{11} & \Phi_{12} \\ \Phi_{20} & \Phi_{21} & \Phi_{22} \end{bmatrix},$$

$$\gamma_{ab} = \begin{bmatrix} 0 & 0 & \frac{1}{2} \\ 0 & -\frac{1}{4} & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix}.$$

The Bianchi identities are given by

$$\begin{aligned} d\Sigma_1 &= 2\sigma_2 \wedge \Sigma_1 - 2\sigma_1 \wedge \Sigma_2, \\ d\Sigma_2 &= \sigma_3 \wedge \Sigma_1 - \sigma_1 \wedge \Sigma_3, \\ d\Sigma_3 &= -2\sigma_2 \wedge \Sigma_3 + 2\sigma_3 \wedge \Sigma_2, \end{aligned} \tag{2.4}$$

and the source-free Maxwell's equations are

$$dF = 0, \tag{2.5}$$

where

$$F = \phi_0 Z^1 + \phi_1 Z^2 + \phi_2 Z^3.$$

The electrovacuum field equations are

$$\phi_{ab} = \phi_a \bar{\phi}_b. \tag{2.6}$$

These are the relevant equations we are going to use in the following. We will operate on them with the formalism developed by Crossman.

Crossman introduces a generalized operator of exterior differentiation

$$\begin{aligned} O^{rs}_{pq} G = [&d - (p+1)\sigma_2 \wedge -(r+1)\bar{\sigma}_2 \wedge \\ &+ q(\rho\theta^1 - \mu\theta^2 + \tau\theta^3 - \pi\theta^4) \wedge \\ &+ s(\bar{\rho}\theta^1 - \bar{\mu}\theta^2 - \bar{\pi}\theta^3 + \bar{\tau}\theta^4) \wedge]G, \end{aligned} \tag{2.7}$$

where G is a p-form.

Besides the useful generalization of the directional derivatives are

$$\begin{aligned} D^{rs}_{pq} &= D-(p+1)\epsilon+\rho q-(r+1)\bar{\epsilon}+s\bar{\rho}, \\ \Delta^{rs}_{pq} &= \Delta+(p+1)\gamma-q\mu+(r+1)\bar{\gamma}-s\bar{\mu}, \\ \delta^{rs}_{pq} &= \delta-(p+1)\beta+q\tau+(r+1)\bar{\alpha}-s\bar{\pi}, \\ \bar{\delta}^{rs}_{pq} &= \bar{\delta}+(p+1)\alpha-q\pi-(r+1)\bar{\beta}+s\bar{\tau}. \end{aligned} \tag{2.8}$$

We remember that if f is a 0-form

$$df=Df\theta^1+\Delta f\theta^2+\delta f\theta^3+\bar{\delta}f\theta^4.$$

In the case of type D charged geometry and with our choice of the tetrad it results that

$$\begin{aligned} &k=\sigma=\lambda=\nu=0, \\ &\Psi_0=\Psi_1=\Psi_3=\Psi_4=0, \\ &\phi_0=\phi_2=0. \end{aligned}$$

Crossman introduces generalized wave operators $N_{c,s}$ given by

$$\begin{aligned} N_{c,s}\Psi = {} & [D^{-2\;\;\;\;-1}_{2(s-1)\;-(s+c)}\cdot\Delta^{-1\;\;\;\;\;\;0}_{-(1+2s)\;s-c-1} \\ & -\delta^{-2\;\;\;\;-1}_{2(s-1)\;-(s+c)}\cdot\bar{\delta}^{-1\;\;\;\;\;\;0}_{-(1+2s)\;(s-c-1)} \\ & -(s-1)(2s-1)\Psi_2]\Psi. \end{aligned} \tag{2.9}$$

For charged type D metrics, $N_{c,s}$ is related to $N_{s,s}$ by

$$N_{c,s}\Psi=\phi_1^{(c-s)/2}N_{s,s}[\phi_1^{(s-c)/2}\Psi], \tag{2.10}$$

where $N_{s,s}$ is separable and it is the charged generalization of Teukolsky's spin-weighted operator.

III. EQUATIONS FOR ELECTROVACUUM PERTURBATION

We now consider the first-order perturbation of the null tetrad $\{l^\mu,n^\mu,m^\mu,\bar{m}^\mu\}$ expressed in terms of the unperturbed tetrad (from now on a subscript capital A denotes the unperturbed quantities and B the first-order perturbed ones)

$$\begin{aligned} l^\mu_B &= L_1 l^\mu_A+L_2 n^\mu_A+L_3 m^\mu_A+\bar{L}_3\bar{m}^\mu_A, \\ n^\mu_B &= N_1 l^\mu_A+N_2 n^\mu_A+N_3 m^\mu_A+\bar{N}_3\bar{m}^\mu_A, \\ m^\mu_B &= M_1 l^\mu_A+M_2 n^\mu_A+M_3 m^\mu_A+M_4\bar{m}^\mu_A, \end{aligned} \tag{3.1}$$

where L_1, L_2, N_1, N_2 are real functions and L_3, N_3, M_1, M_2, M_3, M_4 are complex functions of the local coordinates. From these expressions of the perturbed tetrad it is possible to write the perturbed basis for 1-forms

$$\begin{aligned} \theta^1_B &= -L_1\theta^1_A-N_1\theta^2_A-M_1\theta^3_A-\bar{M}_1\theta^4_A, \\ \theta^2_B &= -L_2\theta^1_A-N_2\theta^2_A-M_2\theta^3_A-\bar{M}_2\theta^4_A, \\ \theta^3_B &= -L_3\theta^1_A-N_3\theta^2_A-M_3\theta^3_A-\bar{M}_4\theta^4_A, \end{aligned} \tag{3.2}$$

and 2-forms

$$\begin{aligned} Z^1_B = {} & -(L_1+M_3)Z^1_A-\tfrac{1}{2}(\bar{M}_1+N_3)Z^2_A-\bar{M}_4\bar{Z}^1_A \\ & +\tfrac{1}{2}(\bar{M}_1-N_3)\bar{Z}^2_A+N_1\bar{Z}^3_A, \\ Z^2_B = {} & -(M_2+\bar{L}_3)Z^1_A-\tfrac{1}{2}(L_1+N_2+M_3+\bar{M}_3)Z^2_A \\ & -(\bar{M}_1+N_3)Z^3_A+(L_3-\bar{M}_2)\bar{Z}^1_A \\ & -\tfrac{1}{2}(L_1+N_2-M_3-\bar{M}_3)\bar{Z}^2_A+(\bar{N}_3-M_1)\bar{Z}^3_A, \end{aligned} \tag{3.3}$$

$$\begin{aligned} Z^3_A = {} & -\tfrac{1}{2}(M_2+\bar{L}_3)Z^2_A-(N_2+\bar{M}_3)Z^3_A+L_2\bar{Z}^1_A \\ & +\tfrac{1}{2}(M_2-\bar{L}_3)\bar{Z}^2_A-M_4\bar{Z}^3_A. \end{aligned}$$

We have then the basis to write the first and the second perturbed structure equations and the perturbed Bianchi identities.

In order to derive our conservation equation let us start considering the following perturbed 2-form:

$$P=3\Psi_{2A}F_B+2\phi^2_{1A}\bar{F}_B-2\phi_{1A}(\Sigma_{2B}+\tfrac{1}{2}\Psi_{2A}Z^2_B), \tag{3.4}$$

where F_B is the perturbed Maxwell's 2-form and Σ_{2B} is the perturbed second curvature 2-form (2.3)

$$\begin{aligned} F_B &= \phi_{0B}Z^1_A+\phi_{1B}Z^2_A+\phi_{2B}Z^3_A+\phi_{1A}Z^2_B, \\ \Sigma_{2B} &= \Psi_{1B}Z^1_A+\Psi_{2B}Z^2_A+\Psi_{3B}Z^3_A+\Psi_{2A}Z^2_B \\ &\quad +\phi_{10B}\bar{Z}^1_A+\phi_{11B}\bar{Z}^2_A+\phi_{12B}\bar{Z}^3_A+\phi_{11A}\bar{Z}^2_B. \end{aligned}$$

The components of P are

$$\begin{aligned} P = {} & \chi_{1B}Z^1_A+\chi_{0B}Z^2_A+\chi_{-1B}Z^3_A-2\phi_{11A}\phi_{1B}\bar{Z}^2_A \\ & -2\phi^{5/2}_{1A}\bar{\phi}^{-1/2}_{1A}\bar{\phi}_{1B}Z^2_A. \end{aligned} \tag{3.5}$$

being

$$\begin{aligned} \chi_{1B} &= 3\Psi_{2A}\phi_{0B}-2\phi_{1A}\Psi_{1B}, \\ \chi_{0B} &= 3\Psi_{2A}\phi_{1B}-2\phi_{1A}\Psi_{2B}+2\phi^{5/2}_{1A}\bar{\phi}^{-1/2}_{1A}\bar{\phi}_{1B}, \\ \chi_{-1B} &= 3\Psi_{2A}\phi_{2B}-2\phi_{1A}\Psi_{3B}. \end{aligned} \tag{3.6}$$

Now our purpose is to compute the external differential of P, but to do this some further equations are needed. We first define two 1-forms

$$\begin{aligned} \Gamma &= \rho_A\theta^1_A-\mu_A\theta^2_A+\tau_A\theta^3_A-\pi_A\theta^4_A, \\ \tilde{\Gamma} &= \rho_A\theta^1_A-\mu_A\theta^2_A-\tau_A\theta^3_A+\pi_A\theta^4_A. \end{aligned} \tag{3.7}$$

From Eqs. (2.3) and (2.5) in the stationary state it results that

$$\begin{aligned} d\Psi_{2A} &= 3\Psi_{2A}\Gamma+2\phi_{11A}\tilde{\Gamma}, \\ d\phi_{1A} &= 2\phi_{1A}\Gamma. \end{aligned} \tag{3.8}$$

We will use in the following the perturbed Maxwell's equations (2.5) written with the generalized operator

$$O^{-1\,0}_{-1\,0}F_B=0 \tag{3.9}$$

and the Fackerell conservation equation

$$O^{-1\;\;\;0}_{-1\;-1}[(\Sigma_{2B}+\tfrac{1}{2}\Psi_{2A}Z^2_B)+\phi_{1A}\bar{F}_B-\phi^{1/2}_{11A}F_B]=0. \tag{3.10}$$

Besides we note that

$$O^{-1\,0}_{-1\,n}\Psi=(d+n\Gamma\wedge)\Psi. \tag{3.11}$$

Let us compute the external differential of the first part of (3.4): from (3.9) and (3.8) it results that

$$\begin{aligned} & O^{-1\,0}_{-1\,0}[3\Psi_{2A}F_B+2\phi^2_{1A}\bar{F}_B] \\ & \quad =9\Psi_{2A}\Gamma\wedge F_B+6\phi_{11A}\tilde{\Gamma}\wedge F_B+8\phi^2_{1A}\Gamma\wedge\bar{F}_B; \end{aligned}$$

then from (3.10) we obtain

$$\begin{aligned} & O^{-1\,0}_{-1\,3}[3\Psi_{2A}F_B+2\phi^2_{1A}\bar{F}_B] \\ & \quad =6\phi_{11A}\tilde{\Gamma}\wedge F_B+2\phi^2_{1A}\Gamma\wedge\bar{F}_B. \end{aligned} \tag{3.12}$$

For the second part of P, using (3.9) and (3.10) we find

$$\begin{aligned} & O^{-1\;\;\;0}_{-1\;-3}\{2\phi_{1A}[\Sigma_{2B}+\tfrac{1}{2}\Psi_{2A}Z^2_B]\} \\ & \quad =-2\phi^2_{1A}\Gamma\wedge\bar{F}_B+2\phi^{3/2}_{1A}\bar{\phi}^{1/2}_{1A}\bar{\Gamma}\wedge F_B. \end{aligned} \tag{3.13}$$

Adding (3.12) and (3.13) we have

$$O^{-1\ \ 0}_{-1\ -3}(P) = 4\phi^{2}_{1A}\Gamma \wedge F_B + 6\phi_{11A}\tilde{\Gamma} \wedge F_B - 2\phi^{3/2}_{1A}\bar{\phi}^{1/2}_{1A}\bar{\Gamma} \wedge F_B. \tag{3.14}$$

For the Kerr–Newman metric it results that

$$\frac{\rho_A}{\bar{\rho}_A} = \frac{\mu_A}{\bar{\mu}_A} = \frac{\tau_A}{\bar{\pi}_A} = \frac{\pi_A}{\bar{\tau}_A}. \tag{3.15}$$

Then

$$\tilde{\Gamma} = (\phi_{1A}/\bar{\phi}_{1A})^{1/2}\bar{\Gamma}, \tag{3.16}$$

and from (3.14) and (3.16)

$$O^{-1\ \ 0}_{-1\ -3}(P) = 4\phi^{3/2}_{1A}(\phi^{1/2}_{1A}\Gamma \wedge \bar{F}_B + \bar{\phi}^{1/2}_{1A}\bar{\Gamma} \wedge F_B). \tag{3.17}$$

But from (3.8)

$$\phi^{1/2}_{1A}\Gamma \wedge \bar{F}_B + \bar{\phi}^{1/2}_{1A}\bar{\Gamma} \wedge F_B = O^{-1\,0}_{-1\,0}[\phi^{1/2}_{1A}\bar{F}_B + \bar{\phi}^{+1/2}_{1A}F_B]$$

and

$$\phi^{3/2}_{1A}O^{-1\,0}_{-1\,0}[\phi^{1/2}_{1A}\bar{F}_B + \bar{\phi}^{1/2}_{1A}F_B] = O^{-1\ \ 0}_{-1\ -3}\phi^{3/2}_{1A}[\phi^{1/2}_{1A}\bar{F}_B + \bar{\phi}^{1/2}_{1A}F_B].$$

Finally we conclude that

$$O^{-1\ \ 0}_{-1\ -3}[P - 4\phi^{3/2}_{1A}(\phi^{1/2}_{1A}\bar{F}_B + \bar{\phi}^{-1/2}_{1A}F_B)] = 0. \tag{3.18}$$

We note that in the case of Kerr $\phi_{1A} = 0$ and the equation (3.18) becomes

$$O^{-1\ \ 0}_{-1\ -3}P = 0. \tag{3.19}$$

But in this case $P = 3\Psi_{2A}F_B$ and from (3.8) and (3.19) we obtain

$$O^{-1\ \ 0}_{-1\ -3}3\Psi_{2A}F_B = 3\Psi_{2A}O^{-1\,0}_{-1\,0}F_B = 0$$

which are the Maxwell equations.

In order to write Eq. (3.18) in a more convenient way we note that taking $\phi^{3/2}_{1A}$ off the operator $O^{-1\ \ 0}_{-1\ -3}$ and writing P in components we obtain

$$\begin{aligned} O^{-1\ \ 0}_{-1\ -3}\{\phi^{3/2}_{1A}[\phi^{-3/2}_{1A}P - 4(\bar{\phi}^{1/2}_{1A}F_B + \phi^{1/2}_{1A}\bar{F}_B)]\} &= O^{-1\,0}_{-1\,0}[\phi^{-3/2}_{1A}(\chi_{1B}Z^{1}_{A} + \chi_{0B}Z^{2}_{A} + \chi_{-1B}Z^{3}_{A}) \\ &\quad - 2\frac{\phi_{1B}}{\phi^{1/2}_{1A}}\bar{\phi}_{1A}\bar{Z}^{2}_{A} - 2\frac{\bar{\phi}_{1B}}{\bar{\phi}^{1/2}_{1A}}\phi_{1A}Z^{2}_{A} \\ &\quad - 4(\bar{\phi}^{1/2}_{1A}F_B + \phi^{1/2}_{1A}\bar{F}_B)] = 0. \end{aligned}$$

Being

$$(\phi^{1/2}_{1})_B = \tfrac{1}{2}(\phi_{1B}/\phi^{1/2}_{1A}) \quad \text{and} \quad \phi_{1A}Z^{2}_{A} = F_A$$

if we define a 2-form χ_B whose components are χ_{1B}, χ_{0B}, and χ_{-1B}, Eq. (3.18) becomes

$$O^{-1\,0}_{-1\,0}[\phi^{-3/2}_{1A}\chi_B - 4(\bar{\phi}^{1/2}_{1}F + \phi^{1/2}_{1}\bar{F}_B)] = 0, \tag{3.20}$$

which is the conservation equation we were looking for.

IV. PROPERTIES OF THE CONSERVATION EQUATION

We now subject the perturbed tetrad to infinitesimal rotations and to infinitesimal translations. A general rotation of the tetrad can be considered as made up of three classes of rotations

$$\begin{aligned} &(1)\ l_B \to l_B,\ m_B \to m_B + a l_A,\ n_B \to n_B + \bar{a}m_A + a\bar{m}_A, \\ &(2)\ n_B \to n_B,\ m_B \to m_B + b n_A,\ l_B \to l_B + \bar{b}m_A + b\bar{m}_A, \\ &(3)\ l_B \to A^{-1}l_A,\ n_B \to A n_A,\ m_B \to e^{i\sigma}m_A, \end{aligned} \tag{4.1}$$

where a and b are two infinitesimal complex functions and A and σ are two infinitesimal real functions. From the definition of the Maxwell and Weyl scalars we have the following rules of transformation:

$$\begin{aligned} \Psi'_{0B} &= \Psi_{0B}, \\ \Psi'_{4B} &= \Psi_{4B}, \\ \Psi'_{1B} &= \Psi_{1B} + 3b\Psi_{2A}, \\ \Psi'_{2B} &= \Psi_{2B}, \\ \Psi'_{3B} &= \Psi_{3B} + 3\bar{a}\Psi_{2A}, \\ \phi'_{0B} &= \phi_{0B} + 2b\phi_{1A}, \\ \phi'_{2B} &= \phi_{2B} + 2\bar{a}\phi_{1A}, \\ \phi'_{1B} &= \phi_{1B}, \end{aligned} \tag{4.2}$$

where the prime indicates the quantities computed in the rotated frame.

From (4.2) it is easy to check that both χ_B and $(\bar{\phi}^{1/2}_{1}F + \phi^{1/2}_{1}\bar{F})_B$ are invariant under infinitesimal tetrad rotations. Thus Eq. (3.20) is tetrad invariant. We can still perform an infinitesimal coordinate transformation at every point in space-time

$$x'^{\mu} = x^{\mu} + \xi^{\mu},$$

where

$$\xi^{\mu} = X l^{\mu}_{A} + Y n^{\mu}_{A} + Z m^{\mu}_{A} + \bar{Z}\bar{m}^{\mu}_{A}$$

and X, Y are real functions and Z is a complex function. Under this coordinate tranformation each perturbed quantity Q_B is changed in Q'_B in this way

$$Q'_B = Q_B - L_{\xi}Q_A, \tag{4.3}$$

where $L_{\xi}Q_A$ is the Lee derivative in the ξ-direction of the unperturbed quantity Q_A. From (4.3) it follows that if a quantity is constant or zero or if it is a product of Kronecker functions in the stationary state, it will be invariant under coordinate translations, namely it will be gauge invariant.

Thus Ψ_{0B}, Ψ_{1B}, Ψ_{3B}, Ψ_{4B}, ϕ_{0B}, and ϕ_{2B} are gauge invariant while

$$\begin{aligned} \Psi'_{2B} &= \Psi_{2B} - 3\Psi_{2A}(\rho_A X - \mu_A Y + \tau_A Z - \pi_A\bar{Z}) \\ &\quad - 2\phi_{11A}(\rho_A X - \mu_A Y - \tau_A Z + \pi_A\bar{Z}) \\ \phi'_{1B} &= \phi_{1B} - 2\phi_{1A}(\rho_A X - \mu_A Y + \tau_A Z - \pi_A\bar{Z}). \end{aligned} \tag{4.4}$$

We immediately see from (3.6) that χ_{1B} and χ_{-1B} are gauge invariant.

Besides, being from (3.15)

$$\bar{\rho}_A X - \bar{\mu}_A Y - \bar{\pi}_A Z + \bar{\tau}_A\bar{Z} = (\bar{\phi}_{1A}/\phi_{1A})^{1/2}(\rho_A X - \mu_A Y - \tau_A Z + \pi_A\bar{Z})$$

it follows that χ_{0B} is also gauge invariant. Thus the 2-form χ_B appearing in Eq. (3.20) is gauge invariant.

Let us consider now the second 2-form in (3.20) namely $(\bar{\phi}^{1/2}_{1}F + \phi^{1/2}_{1}\bar{F})_B$. Under coordinate translations, F_B transforms as

$$F'_B = F_B + O^{-1\,0}_{-1\,0}[\phi_{1A}(Y\theta^{1}_{A} - X\theta^{2}_{A} - \bar{Z}\theta^{3}_{A} + Z\theta^{4}_{A})],$$

from which

$$
\begin{aligned}
&(\phi_1^{1/2}\overline{F}+\bar\phi_1^{1/2}F)'_B \\
&\quad=(\phi_1^{1/2}\overline{F}+\bar\phi_1^{1/2}F)_B+\{\phi_{1A}^{1/2}\bar\phi_{1A}O^{\,-1\,2}_{\,-1\,0} \\
&\qquad\times(Y\theta^1_A-X\theta^2_A+\overline{Z}\theta^3_A-Z\theta^4_A) \\
&\qquad+\phi_{1A}\bar\phi_{1A}^{1/2}O^{\,-1\,0}_{\,-1\,2}(Y\theta^1_A-X\theta^2_A-\overline{Z}\theta^3_A+Z\theta^4_A) \\
&\qquad-\phi_{1A}^{1/2}\bar\phi_{1A}[(\rho_AX-\mu_AY-\tau_AZ+\pi_A\overline{Z})Z^2_A \\
&\qquad+(\rho_AX-\mu_AY+\tau_AZ-\pi_A\overline{Z})\overline{Z}^2_A]\}.
\end{aligned}\tag{4.5}
$$

It is evident that the second 2-form appearing in Eq. (3.20) is not gauge invariant and the part gauge dependent of this 2-form is isolated in (4.5) within curly brackets. But if we apply the operator of external differential to this term we find, with straightforward calculations, that the result is zero. Thus we can conclude that the equation

$$O^{\,-1\,0}_{\,-1\,0}[\phi_{1A}^{-3/2}\chi_B-4(\bar\phi_1^{1/2}F+\phi_1^{1/2}\overline{F})_B]=0$$

is both tetrad and gauge invariant.

V. EQUATIONS FOR $\chi_{1B},\chi_{0B},\chi_{-1B}$

In order to obtain the equations for χ_{1B},χ_{0B}, and χ_{-1B} it is convenient to write Eq. (3.20) in this form

$$
\begin{aligned}
&O^{\,-1\,0}_{\,-1\,0}\phi_{1A}^{-3/2}[\chi_{1B}Z^1_A+\chi_{0B}Z^2_A+\chi_{-1B}Z^3_A] \\
&\quad=4O^{\,-1\,0}_{\,-1\,0}[G_{1B}Z^1_A+G_{2B}Z^2_A+G_{3B}Z^3_A+\text{c.c.}]\,,
\end{aligned}\tag{5.1}
$$

where c.c. indicates complex conjugation and G_{1B},G_{2B},G_{3B} and their conjugates are the components of the perturbed 2-form $(\phi_1^{1/2}\overline{F}+\bar\phi_1^{1/2}F)_B$,

$$
\begin{aligned}
G_{1B}&=\{\bar\phi_{1A}\phi_{1A}^{1/2}(\overline{L}_3-M_2)+\bar\phi_{1A}^{1/2}[\phi_{0B}-\phi_{1A}(M_2+\overline{L}_3)]\},\\
G_{2B}&=\{\bar\phi_{1A}^{1/2}[\phi_{1B}-\tfrac12\phi_{1A}(L_1+N_2+M_3+\overline{M}_3)]-\tfrac12\phi_{1A}^{1/2}\bar\phi_{1A}(L_1+N_2-M_3-\overline{M}_3)+\tfrac12(\phi_{1A}/\bar\phi_{1A}^{1/2})\bar\phi_{1B}\},\\
G_{3B}&=\{\bar\phi_{1A}\phi_{1A}^{1/2}(N_3-\overline{M}_1)+\bar\phi_{1A}^{1/2}[\phi_{2B}-\phi_{1A}(\overline{M}_1+N_3)]\}.
\end{aligned}
$$

The 2-form appearing on the left-hand side of equation (5.1) is a self-dual form, while the 2-form on the right-hand side is not.

Operating on Eq. (5.1) with $*O^{\,-1\,0}_{\,-1\,-2}*$ and only considering the self-dual part of the resulting equations we obtain the set of equations for χ_{1B},χ_{0B}, and χ_{-1B}

$$
\begin{aligned}
N_{1,1}[\phi_{1A}^{-3/2}\chi_{1B}]&=4\{N_{1,1}G_{1B}+\delta^{\,-2\,-1}_{\,0\,-2}\delta^{\,-3\,-1}_{\,-1\,0}\overline{G}_{1B}\\
&\quad+D^{\,-2\,-1}_{\,0\,-2}D^{\,-3\,-1}_{\,-1\,0}\overline{G}_{3B}-(D^{\,-2\,-1}_{\,0\,-2}\delta^{\,-1\,-2}_{\,-1\,0}+\delta^{\,-2\,-1}_{\,0\,-2}D^{\,-1\,-2}_{\,-1\,0})\overline{G}_{2B}\},\\
[N_{0,0}+\Psi_{2A}][\phi_{1A}^{-2}\chi_{0B}]&=4\{[N_{0,0}+\Psi_{2A}]\phi_{1A}^{-1/2}G_{2B}\\
&\quad-\tfrac12[\Delta^{\,-2\,-1}_{\,-2\,0}\delta^{\,-3\,-1}_{\,-1\,1}+\delta^{\,-2\,-1}_{\,-2\,0}\Delta^{\,-3\,-1}_{\,-1\,1}]\phi_{1A}^{-1/2}\overline{G}_{1B}-\tfrac12[D^{\,-2\,-1}_{\,-2\,0}\bar\delta^{\,-3\,-1}_{\,-1\,1}\\
&\quad+\bar\delta^{\,-2\,-1}_{\,-2\,0}D^{\,-3\,-1}_{\,-1\,1}]\phi_{1A}^{-1/2}\overline{G}_{3B}\\
&\quad+\tfrac12[D^{\,-2\,-1}_{\,-2\,1}\Delta^{\,-1\,-2}_{\,-1\,1}+\Delta^{\,-2\,-1}_{\,-2\,0}D^{\,-1\,-2}_{\,-1\,1}+\delta^{\,-2\,-1}_{\,-2\,0}\bar\delta^{\,-1\,-2}_{\,-1\,1}\\
&\quad+\bar\delta^{\,-2\,-1}_{\,-2\,0}\delta^{\,-1\,-2}_{\,-1\,1}]\phi_{1A}^{-1/2}\overline{G}_{2B}\},\\
N_{-1,-1}[\phi_{1A}^{-5/2}\chi_{-1B}]&=4\{N_{-1,-1}\phi_{1A}^{-1}G_{3B}+\Delta^{\,-2\,-1}_{\,0\,0}\Delta^{\,-3\,-1}_{\,-1\,-2}\phi_{1A}^{-1}\overline{G}_{1B}\\
&\quad+\bar\delta^{\,-2\,-1}_{\,0\,0}\bar\delta^{\,-3\,-1}_{\,-1\,-2}\phi_{1A}^{-1}\overline{G}_{3B}-[\Delta^{\,-2\,-1}_{\,0\,0}\bar\delta^{\,-1\,-2}_{\,-1\,2}+\bar\delta^{\,-2\,-1}_{\,0\,0}\Delta^{\,-1\,-2}_{\,-1\,2}]\phi_{1A}^{-1}\overline{G}_{2B}.
\end{aligned}\tag{5.2}
$$

As we said in Sec. II, the operator $N_{s,s}$ is separable but the other operators appearing in the previous equations are not separable in Boyer–Lindquist coordinates in the Kerr–Newman metric. It is clear from Eq. (5.2) that if $\overline{G}_{1B},\overline{G}_{2B}$, and $\overline{G}_{3B}$, namely the anti-self-dual part of the two-form $(\bar\phi_1^{1/2}F+\phi_1^{1/2}\overline{F})_B$, were zero, the three equations would be decoupled and separable. In fact if A is a self-dual closed two-form, it is possible to show that the self-dual components of the equations

$$*O^{\,-1\,0}_{\,-1\,-2}*O^{\,-1\,0}_{\,-1\,0}A=0$$

are three decoupled equations for the three components of A:

$$N_{1,1}[A_1]=0,$$

$$(N_{1,0}+\psi_{2A})[A_2]=0,$$

$$N_{-1,1}[\psi_{2A}A_3]=0.$$

Therefore, we are in the same situation as if we had used the equation $dF=0$ to obtain decoupled and separable equations. In fact Eq. (3.9) written in components is

$$
\begin{aligned}
&O^{\,-1\,0}_{\,-1\,0}\{[\phi_{0B}-(M_2+\overline{L}_3)\phi_{1A}]Z^1_A\\
&\quad+[\phi_{1B}-\tfrac12\phi_{1A}(L_1+N_2+M_3+\overline{M}_3)]Z^2_A\\
&\quad+[\phi_{2B}-\phi_{1A}(\overline{M}_1+N_3)]Z^3_A\}\\
&=O^{\,-1\,0}_{\,-1\,0}\{\overline{G}'_{1B}\overline{Z}^1_A+\overline{G}'_{2B}\overline{Z}^1_A+\overline{G}'_{3B}\overline{Z}^3_A\},
\end{aligned}\tag{5.3}
$$

where the right-hand side 2-form has only anti-self-dual components

$$\overline{G}'_{1B}=\phi_{1A}(L_3-\overline{M}_2),$$

$$\overline{G}'_{2B}=-\phi_{1A}(L_1+N_2-M_3-\overline{M}_3),$$

$$\overline{G}'_{3B}=\phi_{1A}(\overline{N}_3-M_1).$$

Operating on (5.3) with $*O^{\,-1\,0}_{\,-1\,-2}*$ we obtain a set of equations that are exactly the same as Eqs. (5.2) when we replace the set of quantities $(\chi_{1B},\chi_{0B},\chi_{-1B})$ with $\{[\phi_{0B}-(M_2+\overline{L}_3)\phi_{1A}]$, $[\phi_{1B}-\tfrac12\phi_{1A}(L_1+N_2+M_3+\overline{M}_3)]$, $[\phi_{2B}-\phi_{1A}(\overline{M}_1+N_3)]\}$, the set $(\overline{G}_{1B},\overline{G}_{2B},\overline{G}_{3B})$ with $(\overline{G}'_{1B},\overline{G}'_{2B},\overline{G}'_{3B})$ and we put $G_{1B}=G_{2B}=G_{3B}=0$.

It is however evident that Eq. (3.20) has a different information content from Eq. (3.9), because it involves, via $\chi_{1B},\chi_{0B},\chi_{1B}$, both Weyl's scalars and Maxwell's scalars.

The first and the third equations (5.2) reduce to those obtained by Lee in the case of small charge and angular momentum ($e \ll M$, $a \ll M$), being all the terms on the right-hand side quadratic or more both in e and a, when written in Boyer–Lindquist coordinates.

VI. CONCLUDING REMARKS

In this paper we have derived a new gauge and tetrad invariant equation for the Kerr–Newman metric, involving first-order perturbed quantities. We think that the problem of decoupling and separability of the perturbation equations could be solved by finding a 2-form A, whose components involve gauge and tetrad invariant perturbed quantities (as $\chi_{1B}, \chi_{0B}, \chi_{-1B}, \psi_{0B}, \psi_{4B}$), that satisfies the conservation equation $dA = 0$ and that is self-dual. Our equation can be used for further investigations of the problem in the following way: It is necessary to find further conservation equations and properly combine them with our equation, Fackerell's equation, and Maxwell's equation, in such a way that the anti-self-dual part of the resulting 2-form cancels. Once the requirement of self-duality is satisfied, Crossman's formalism will automatically lead to decoupled and separable equations for the components of A.

ACKNOWLEDGMENTS

We wish to thank Professor R. Ruffini and Professor M. Bordoni for their encouragement and useful suggestions.

[1]S. A. Teukolsky, Astrophys. J. **185**, 635 (1973).

[2]S. Chandrasekhar, Proc. R. Soc. London, Ser. A **358**, 421–439 (1978).

[3]Chandrasekhar's equations for Φ_0, Φ_1, K, S, are only gauge invariant. In a forthcoming paper we will show that they can be generalized, within the complex vectorial formalism, in a tetrad and gauge-invariant form.

[4]C. H. Lee, J. Math. Phys. **17**, 1226–1235 (1976).

[5]C. H. Lee, Research Institute for Fundamental Physics, Kyoto University, Preprint RIFP-427, 1981.

[6]R. G. Crossman, Lett. Math. Phys. **1**, 105–109 (1976).

[7]E. D. Fackerell, in *Proceedings of the Second Marcel Grossman Meeting on General Relativity*, edited by R. Ruffini (North-Holland, Amsterdam, 1982).

[8]M. Cahen, E. Debever, and L. Defrise, J. Math. Mech. **16**, 761–785 (1967).

[9]R. G. Crossman, "Electrovac perturbations of black holes using complex vectorial formalism," Ph.D. dissertation, Monash University, June 1978 (unpublished).

[10]R. G. Crossman and E. D. Fackerell, *Proceedings of the Einstein Centenary Summer School on Gravitational Radiation and Collapsed Objects*, edited by C. Edwards (Springer-Verlag, Berlin, 1980).

[11]V. Bellezza, Tesi di laureo, University of Rome, 1982 (unpublished).

[12]L. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).

[13]P. Jordan, J. Ehlers, and R. K. Sachs, Akad. Wiss. Lit. Mainz Abh. Math.-Naturwiss. Kl. **1**, 3 (1961).

[14]S. Chandrasekhar, *The Mathematical Theory of Black Holes* (Oxford U. P., New York, 1983).

Isometries and dimensional reduction[a]

Freydoon Mansouri and Louis Witten
Department of Physics, University of Cincinnati, Cincinnati, Ohio 45221

(Received 17 August 1983; accepted for publication 9 December 1983)

We study dimensional reduction in $(d+n)$-dimensional space-times, with or without torsion, which admit n commuting Killing vectors. The field content of theories in $(d+n)$ dimensions reduced to d dimensions are completely worked out. The equations of motion satisfied by these fields are also derived explicitly.

PACS numbers: 04.50. + h, 02.40. + m, 11.30.Ly, 11.15. − q

I. INTRODUCTION

Dimensional reduction provides an attractive framework for a possible geometrical unification of gravity with gauge theories of internal symmetry. To this end, a number of suggestions, based on the generalization of Kaluza–Klein theory,[1] already exists in the literature.[1–4] From the contents of these works, it is clear that dimensional reduction is not a unique and universal procedure. Moreover, the extent to which one can take any such scheme seriously depends crucially on the definiteness and clarity of the initial assumptions and their relation to the final results.

The main objective of this article is to report a formulation of dimensional reduction for $(d+n)$-dimensional space-times, with or without torsion, which admit n commuting Killing vectors. This is the generalization to arbitrary n of our previous work[5] for $n=1$ and $n=2$. In our approach to dimensional reduction, a number of interesting features arise which are worth noting. One of these is the appearance of an SO(n) symmetry which was absent in the initial theory. It is the anlog of the global symmetries which arise in supergravity theories. Another interesting feature is that the cosmological constant does not appear in the dimensionally reduced set of equations which we obtain, if it were zero in the initial $(d+n)$-dimensional theory.

II. BASIC ELEMENTS OF DIMENSIONAL REDUCTION

In some applications, such as in Einstein–Cartan theory and supergravity, the torsion tensor is, in general, not zero, so that we must deal with nonsymmetric connections. To accommodate such possibilities, we will introduce vielbein fields and covariant derivatives in a way which is familiar from the gauge theory approach to gravity and supergravity.[6] This language makes our formalism accessible to a larger number of physicists. There is no sacrifice of rigor, however. The mathematically minded reader will note that our starting point is to introduce a connection in a $(d+n)$-dimensional manifold via the corresponding bundle of linear or orthonormal frames.

We consider a $(d+n)$-dimensional manifold M, which is characterized by the structure group GL$(d+n,R)$ or one of its subgroups. We specify a connection in M by the covariant derivative

$$\hat{D}_a = \partial_a + \hat{H}_a^{\ AB}\hat{X}_{AB},\qquad a=0,\ldots,d+n-1,\quad A,B=0,\ldots,d+n-1, \tag{1}$$

where $\hat{H}_a^{\ AB}$ are the gauge fields (connection coefficients) and $\hat{X}_{AB}$ are the representations (in general reducible) of the generators of the group GL$(d+n,R)$.

Here and in what follows the "caret" on top of any quantity indicates that it corresponds not to a dimensionally reduced manifold but to M itself. When no ambiguity arises, the caret will be omitted. For definiteness, we assume that M has one time and $d+n-1$ spacial dimensions. We also introduce a set of $(d+n)$-bein matrices, $\hat{K}_a^{\ A}$, with their corresponding inverses, $\hat{K}^b_{\ B}$, satisfying the orthonormality conditions

$$\hat{K}_a^{\ A}\hat{K}^a_{\ B} = \delta^A_B,\quad \hat{K}_a^{\ A}\hat{K}^b_{\ A} = \delta^b_a. \tag{2}$$

One can then express the metric on M in the form

$$\hat{g}_{ab} = \hat{K}_a^{\ A}\hat{K}_b^{\ B}\eta_{AB}, \tag{3}$$

where η_{AB} are the components the Minkowskian metric with signature $(+,-,\ldots,-)$. For many purposes, it is more convenient to express the covariant derivative (1) in a new basis with components $\hat{X}_A$ given by

$$\hat{X}_A = \hat{K}^a_{\ A}\hat{D}_a. \tag{4}$$

One advantage of using the basis $\{\hat{X}_A\}$ is that the components of both the metric tensor and the Levi-Civita tensor will become constants. The commutators of the elements of the set $\{\hat{X}_A\}$ are given by[6,7]

$$[\hat{X}_A,\hat{X}_B] = \hat{K}^a_{\ A}K^b_{\ B}[\hat{R}_{ab}^{\ \ CD}\hat{X}_{CD} - \hat{T}^C_{ab}\hat{X}_C] \equiv \hat{R}_{AB}^{\ \ CD}\hat{X}_{CD} - \hat{T}^C_{AB}\hat{X}_C, \tag{5}$$

where $\hat{R}_{AB}^{\ \ CD}$ and $\hat{T}^C_{AB}$ are, respectively, the components of the curvature and torsion tensors in the basis $\{\hat{X}_A\}$.

Up to this point we have introduced a number of structures on M to describe its intrinsic differential geometry. To proceed with dimensional reduction, we now assume that there are n independent everywhere spacelike Killing vectors ${}^k\xi$ on M. In other words, we assume that metric tensor of M has an n-parameter group of isometries. In this paper we further assume that these isometries are abelian, i.e., that the corresponding Killing vectors commute. The nonabelian isometries will be dealt with elsewhere.[8] Depending on the basis we work in, we represent the components of the Killing vectors by $\{{}^i\xi^A\}$ or $\{{}^i\xi^a\}$, where

$${}^j\xi^A = \hat{K}_a^{\ A}\,{}^j\xi^a. \tag{6}$$

The presence of n Killing vectors is the signal that n of the $d+n$ coordinates which parametrize the manifold M are redundant and that it might be possible to utilize the action of the symmetry to obtain a d-dimensional formulation of an initially $(d+n)$-dimensional theory. This would then consti-

[a] Supported in part by Department of Energy under Contract No. DE-AS-2-76ER02978.

tute a dimensional reduction from a $(d+n)$-dimensional manifold M to a d-dimensional manifold S. A method of carrying out this procedure was given by Geroch[4] in his four- to three- and four- to two-dimensional reductions. His formalism can easily be restated to apply to more general dimensional reductions: One starts by dividing M into orbits under the desired isometry. Two points p and q belong to the same orbit if there exists a curve from p to q the tangent to which at any point is a linear combination of Killing vectors ${}^j\xi$. Thus, the orbits are spacelike n-surfaces. The elements of the manifold S are then taken to be the collection of all such orbits in M. As pointed out by Geroch, it is not possible to regard the manifold S as a hypersurface in M except when the n-manifold spanned by the Killing vectors ${}^j\xi$ is orthogonal to a d-manifold in M. It is more natural to regard S as a quotient space of M by the action of the isometry group. In fact, let ψ be a mapping from M to S which assigns to each point p of M the orbit $\psi(p)$ passing through p. Then, one can establish a one-to-one correspondence[4] between the tensor fields on S and tensor fields on M, which satisfy the following two conditions:

$$
\begin{aligned}
{}^j\xi_A \hat{T}^{A\cdots B}_{C\cdots D} &= \cdots = {}^j\xi_B \hat{T}^{A\cdots B}_{C\cdots D} = {}^j\xi^C \hat{T}^{A\cdots B}_{C\cdots D} \\
&= \cdots = {}^j\xi^D T^{A\cdots B}_{C\cdots D} = 0, \quad j = 1,\ldots,n
\end{aligned} \tag{7}
$$

$$
\underset{{}^j\xi}{\mathscr{L}} \hat{\mathbf{T}} = 0, \quad j = 1,\ldots,n, \tag{8}
$$

where $\mathscr{L}$ is the symbol for "Lie derivative," and the condition (8) states that the Lie derivative of the tensor $\hat{\mathbf{T}}$ along the direction specified by the Killing vector ${}^j\xi$ vanishes. The requirements (7) and (8) ensure that the intrinsic geometry of S is patterned after M and yet is self-supporting in the sense that, once established, it need not make any reference to M as long as the corresponding isometries are exact. Moreover, since various tensor operations commute with the map ψ from M to S, to set up theories which make sense in S, it is not necessary to restrict the range of indices to those natural to S, so that we can retain manifest covariance with respect to M. From this point of view, the creation of manifold S is a secondary notion which helps visualize unified theories in $(d+n)$ dimensions, which satisfy the conditions (7) and (8).

We now proceed with dimensional reduction. Consider the $n \times n$ matrix of scalar fields with elements

$$
\lambda_{jk} = {}^j\xi^A\, {}^k\xi_A = {}^j\xi^a\, {}^k\xi_a, \quad j,k = 1,\ldots,n. \tag{9}
$$

It is easy to verify that the $n(n+1)/2$ fields λ_{jk} satisfy the conditions (7) and (8) and are therefore scalar fields on S. Similarly, consider the quantity

$$
K^A_a = \hat{K}^A_a - {}^j\xi^A (\lambda^{-1})_{jk}\, {}^k\xi_a. \tag{10}
$$

From (6) it follows that ${}^i\xi_A K^A_a = {}^i\xi^a K^A_a = 0$, for all components K^A_a. Since K^A_a also may be chosen to satisfy the condition (8), it is a mixed tensor (vielbein) on S. From λ_{jk} and K^A_a one can then construct the following tensors on S:

$$
h^b_a = K^A_a K^b_A = \delta^b_a - {}^j\xi_b (\lambda^{-1})_{jk}\, {}^k\xi_a, \tag{11}
$$

$$
h^B_A = K^B_b K^b_A = \delta^B_A - {}^j\xi^B (\lambda^{-1})_{jk}\, {}^k\xi_A, \tag{12}
$$

$$
h_{ab} = \eta_{AB} K^A_a K^B_b = \hat{g}_{ab} - {}^j\xi_a (\lambda^{-1})_{jk}\, {}^k\xi_b, \tag{13}
$$

$$
h_{AB} = h_{ab} K^a_A K^b_B = \hat{\eta}_{AB} - {}^j\xi_A (\lambda^{-1})_{jk}\, {}^k\xi_B. \tag{14}
$$

These expressions define the components of the metric and the Kronecker delta on S.

We now turn to the requirements demanded by the structure of isometries. The requirement of commutativity implies that for $j,k = 1,\ldots,n$

$$
{}^j\xi^A \hat{X}_A\, {}^k\xi^B - {}^k\xi^A \hat{X}_A\, {}^j\xi^B = {}^k\xi^C\, {}^j\xi^D \hat{T}^B_{CD}, \tag{15}
$$

where $\hat{T}^B_{CD}$ is the torsion tensor. Moreover, since by the definition of an isometry the Lie derivative of the metric tensor in the direction of the Killing vectors vanishes, we must have

$$
\hat{X}_A\, {}^k\xi_B + \hat{X}_B\, {}^k\xi_A = {}^k\xi^C [\eta_{AD} \hat{T}^D_{BC} + \eta_{BD} \hat{T}^D_{AC}]. \tag{16}
$$

For vanishing right-hand side, this reduces to the familiar Killing equation. Finally, following Geroch, we introduce the covariant derivative on S in terms of the covariant derivative on M in the following way:

$$
X_A T^{B\cdots C}_{D\cdots E} = h^L_A h^M_D \cdots h^N_E h^B_I \cdots h^C_J \hat{X}_L T^{I\cdots J}_{M\cdots N}. \tag{17}
$$

Since our connection on M is, in general, nonsymmetric, the connection on S will also be nonsymmetric, in general. Moreover, since the nonsymmetric part of the connection is related to the torsion tensor, this tensor must be compatible with the covariant derivative (17) on S. One way to ensure this is to require that the torsion tensor of M be a tensor on S, i.e.,

$$
{}^k\xi^A \hat{T}^C_{AB} = {}^k\xi^B \hat{T}^C_{AB} = 0, \quad \underset{\xi}{\mathscr{L}} T = 0. \tag{18}
$$

This is in line with some of the field theoretic applications, e.g., to supergravity, which we have in mind. With conditions (18), it is straightforward to check that expression (17) satisfies all the requirements of a covariant derivative on S.

In terms of $\{X_A\}$, the components of curvature and torsion of S are given by

$$
[X_A, X_B] = R^{CD}_{AB} X_{CD} - T^C_{AB} X_C. \tag{19}
$$

Using (17), one can then uniquely express these in terms of the components of curvature and torsion of M, the scalars λ_{jk}, and the covariant derivatives of ${}^j\xi$:

$$
\begin{aligned}
R^{LM}_{AB} &= h^C_{[A} h^D_{B]} h^L_{[J} h^M_{I]} [\hat{R}^{JI}_{CD} \\
&\quad - 2(\lambda^{-1})_{kj} (\hat{X}_C\, {}^k\xi^I)(\hat{X}_D\, {}^j\xi^J) \\
&\quad - 2(\lambda^{-1})_{kj} (\hat{X}^I\, {}^k\xi^J)(\hat{X}_C\, {}^j\xi_D)], \\
h^C_{[A} h^D_{B]} &= \tfrac{1}{2}(h^C_A h^D_B - h^C_B h^D_A).
\end{aligned} \tag{20}
$$

III. SPECTRA AND EQUATIONS OF MOTION

The ultimate objective of any dimensional reduction is to formulate $(d+n)$-dimensional theories in such a way that in the limit of exact symmetry, they would have the appearance of a d-dimensional theory. So, we must go one step further and express the covariant derivatives of the Killing vectors in terms of fields which are intrinsically on S. Thus, consider the antisymmetric field strength tensors

$$
\omega^{A\cdots B}_j = [n!(d-2)!]^{-1} \epsilon_{l\ldots m} \epsilon^{A\cdots BC\cdots DEF}\, {}^l\xi_C \cdots {}^m\xi_D \hat{X}_E\, {}^j\xi_F, \tag{21}
$$

$$
A,B = 0,\ldots,d+n-1, \quad j,l,m = 1,\ldots,n,
$$

where $\epsilon^{A\cdots F}$ is the $(d+n)$-dimensional Levi-Civita tensor (constant in the basis we work in), and $\epsilon_{l\ldots m}$ is the Euclidean n-dimensional Levi-Civita tensor. Clearly, there will be n

such $(d-2)$-forms which satisfy the conditions (7) and (8) and which are, therefore, tensors on S. Starting from the definition (21), one can show that

$$\begin{aligned}\hat{X}_I{}^j\xi_J &= (-1)^{d+n-1}[2\lambda\,(n!)]^{-1}\epsilon_{l\cdots m}\epsilon_{A\cdots BG\cdots HIJ}\,\omega_j^{A\cdots B}\,{}^l\xi^G\cdots{}^m\xi^H \\ &\quad -(\lambda^{-1})_{lm}\,{}^l\xi_{[I}\hat{X}_{J]}\lambda_{mj}. \end{aligned} \tag{22}$$

$$\lambda = \det(\lambda_{ij}); \tag{23}$$

the $(\lambda^{-1})_{lm}$ are elements of the matrix inverse to (λ_{ij}):

$$(\lambda^{-1})_{lm}\lambda_{mk} = \delta_{lk} \tag{24}$$

and

$${}^l\xi_{[I}\hat{X}_{J]}\lambda_{mj} = \tfrac{1}{2}({}^l\xi_I\hat{X}_J - {}^l\xi_J\hat{X}_I)\lambda_{mj}. \tag{25}$$

In the dimensionallly reduced theory the $(d-2)$ forms given by (21) may play the role of field strength tensors associated with gauge fields coupled to gravity. It is therefore of interest to know their curls and divergences. We define the generalized curl by the expression

$$\begin{aligned}\hat{X}^{[H}\omega_j^{L\cdots M]} &= (-1)^{d+n-1}[(n+1)!(d-1)!]^{-1} \\ &\quad\times\epsilon^{HL\cdots MI\cdots J}\epsilon_{GA\cdots BI\cdots J}\hat{X}^G\omega_j^{A\cdots B}.\end{aligned} \tag{26}$$

Then, after a lengthy computation, one can show that

$$\begin{aligned}\hat{X}^{[H}\omega_j^{L\cdots M]} &= (-1)^{d+n-1}[n!(d-1)!]^{-1}\epsilon_{l\cdots qm}\epsilon^{HL\cdots MA\cdots BDE}\,{}^l\xi_A\cdots{}^q\xi_B \\ &\quad\times[2\,{}^m\xi_D({}^j\xi_F\hat{R}^F_E + T^F_{EG}\hat{X}_F\,{}^j\xi^G)].\end{aligned}$$

Similarly,

$$\begin{aligned}X^{[H}\omega_j^{L\cdots M'M]} &= 2(-)^{d+n-1}[n!(d-2)!]^{-1} \\ &\quad\times\epsilon_{l\cdots qm}\epsilon^{HL\cdots MA\cdots BDE}\,{}^l\xi_A\cdots{}^q\xi_B\,{}^m\xi_D\,{}^i\xi_F\hat{R}^F_E \\ &\quad+\omega_j^{[L\cdots M}\eta^{N]B}T^A_{AB} \\ &\quad+(-)^d(d-2)\omega_j^{A[L\cdots M'}T^M_{AB}\eta^{N]B}.\end{aligned} \tag{27}$$

From this, it follows that in the special case of vanishing $\hat{R}^F_E$ and of vanishing torsion our $(d-2)$-forms are curl-free. Also,

$$\begin{aligned}X_A\omega_j^{AB\cdots CD} &= (\lambda^{-1})_{mk}[\omega_m^{AB\cdots CD}X_A\lambda_{kj} + \tfrac{1}{2}\omega_j^{AB\cdots CD}X_A\lambda_{mk}] \\ &\quad+\omega_j^{AB\cdots CD}T^F_{FA} + \tfrac{1}{2}(-1)^d\omega_j^{AF[B\cdots C}T^{D]}_{AF}.\end{aligned} \tag{28}$$

One would expect that in the dimensionally reduced theory each of the scalar fields λ_{ij} satisfy a Klein–Gordon equation with a source. We find that this is indeed the case. In fact, a straightforward application of the formulas we have derived above shows that

$$\begin{aligned}X^AX_A\lambda_{jk} &= (-1)^{d+n-1}(d-2)!\lambda^{-1}\omega_{jC\cdots D}\omega_k^{C\cdots D} - 2\,{}^j\xi^D\hat{R}^F_D\,{}^k\xi_F \\ &\quad+(\lambda^{-1})_{mq}(X^B\lambda_{qj})(X_B\lambda_{mk}) - \tfrac{1}{2}(\lambda^{-1})_{pq}(X^A\lambda_{pq})(X_A\lambda_{jk}).\end{aligned} \tag{29}$$

Our final task is to obtain expressions for the Ricci tensor and the scalar curvature of the manifold S. By definition,

$$R^M_B = h^A_L R^{LM}_{AB}, \tag{30}$$

where R^{LM}_{AB} is given by Eq. (20). Making use of Eq. (22), one obtains the following expression for the Ricci tensor of S:

$$\begin{aligned}R^M_B &= h^D_B h^M_I\hat{R}^I_D + \tfrac{1}{2}(\lambda^{-1})_{kj}X^MX_B\lambda_{kj} \\ &\quad+\tfrac{1}{4}(X^M\lambda_{kj})(X_B\lambda^{-1}_{kj}) \\ &\quad+\tfrac{1}{2}(-1)^{d+n-1}[d-2)!]\lambda^{-1}(\lambda^{-1})_{kj} \\ &\quad\times[h^M_B\omega_{jA\cdots D}\omega_k^{A\cdots D} - (d-2)\omega_{jBA\cdots D}\omega_k^{MA\cdots D}] \\ &\quad+\tfrac{1}{2}\eta^{MA}(\lambda^{-1})_{kj}T^E_{AB}X_E\lambda_{kj}.\end{aligned} \tag{31}$$

Similarly,

$$\begin{aligned}R &= h^B_M R^M_B \\ &= \hat{R} - {}^j\xi^D(\lambda^{-1})_{jk}\,{}^k\xi_I\hat{R}^I_D + \tfrac{1}{2}(\lambda^{-1})_{jk}X^AX_A\lambda_{jk} \\ &\quad+(-1)^{d+n-1}[(d-2)!]\lambda^{-1}\lambda^{-1}_{pq}\omega_p^{A\cdots E}\omega_q^{A\cdots E} \\ &\quad+\tfrac{1}{4}(X^A\lambda_{pq})(X_A\lambda^{-1}_{pq}).\end{aligned} \tag{32}$$

IV. DISCUSSION

We now turn to a discussion of the results and their possible significance. Interpreted in the most straightforward manner, Eqs. (27), (28), (29), and (31) are the main results of our dimensional reduction. To see this, suppose that in $(d+n)$-dimensional the source-free Einstein–Cartan equations $\hat{R}^I_D = 0$ are satisfied. One can also postulate a suitable set of equations for the torsion tensor. Then, the above four d-dimensional equations determine the dynamics of the scalar fields λ_{jk}, $(d-2)$-forms, $\omega_k^{A\cdots E}$, and the gravitational field h_{AB}. Its nongravitational sector is highly nonlinear, and the interactions of scalar and gauge fields are not of the minimal type. The field content of the reduced theory is fixed by the intrinsic geometry and the symmetries of the initial theory. There are n field strengths or $(d-2)$-forms and $\frac{1}{2}n(n+1)$ scalar fields. There is a natural global SO(N) symmetry associated with the matter field sector of the reduced theory, which was not present in the initial $(d+n)$-dimensional theory. It provides an explicit example of how such symmetries arise in supergravity theories.[3,9] There are also n-abelian local gauge symmetries.

Our results are to be compared and contrasted with other recent approaches to dimensional reduction[10]: (a) Our Killing vectors represent the isometries not of some submanifold of M but of the entire $(d+n)$-dimensional manifold. As a result, their scalar products determine the scalar fields and their twists the gauge fields of the reduced theory. (b) As long as the isometries are exact, there will be no dependence on the extra dimensions and no need for harmonic expansions in those variables. (c) The dependence on the extra n-coordinates can be associated with the breakdown of the initial symmetry of the metric tensor g. This could be triggered, e.g., by some sort of spontaneous symmetry breaking. It could also happen because of the anomalies which arise when the theory is quantized. (d) There is no cosmological constant in the reduced set of equations if there were none present in the $(d+n$-dimensional theory. (e) Since we have an explicit set of equations for the dimensionally reduced theory, then such questions as the ground state solution, its stability, etc., can be unambiguously determined for a given set of boundary conditions.

ACKNOWLEDGMENTS

We would like to thank Paul Esposito and Peter Suranyi for discussions and the Aspen Center for Physics for its hospitality.

[1]For a modern version and references to earlier works see, e.g., L. N. Chang, K. Macrae, and F. Mansouri, Phys. Lett. B **57**, 59 (1975); Phys. Rev. D **13**, 235 (1976).
[2]See, e.g., J. Rayski, Acta Phys. Pol. **27**, 89 (1965); J. Scherk and J. H. Schwarz, Phys. Lett. B **57**, 463 (1975); Y. M. Cho and P. G. O. Freund, Phys. Rev. D **12**, 1711 (1975); G. Domokos and S. Kovesi-Domokos, Nuovo Cimento A **44**, 318 (1978); J. F. Luciani, Nucl. Phys. B **135**, 111 (1978); J. Scherk and J. H. Schwarz, Nucl. Phys. B **153**, 61 (1979); E. Witten, Nucl. Phys. B. **186**, 412 (1981); A. Salam and J. Strathdee, Ann. Phys. **141**, 316 (1982); R. Percacci and S. Ranjbar-Daemi, ICTP preprint, 1982; F. Ardalan and F. Mansouri, Phys. Rev. D. **9**, 3341 (1974).
[3]E. Cremmer and B. Julia, Phys. Lett. B **80**, 48 (1978).
[4]R. Geroch, J. Math. Phys. **12**, 918 (1971); **13**, 394 (1972); F. P. Esposito and L. Witten, Proceedings of the Second Latin American Symposium on General Relativity and Gravitation, Caracas, 1976, edited by C. Aragone, Simon Bolivar University, Department of Physics, Caracus.
[5]F. Mansouri and L. Witten, Phys. Lett. B **127**, 341 (1983).
[6]See, e.g., F. Mansouri, *Proceedings of Integrative Conference on Group Theory and Mathematical Physics*, edited by W. Beiglbuck, A. Bohm, and E. Takasugi (Springer-Verlag, Berlin, 1978); F. Mansouri and C. Schaer, Phys. Lett. B **101**, 51 (1981).
[7]F. Mansouri, *Proceedings of the Fifth Workshop on Current Problems in High Energy, Florence, Italy*, edited by R. Casalbuoni *et al.* (John Hopkins U.P., Baltimore, 1979).
[8]F. Mansouri and L. Witten, Proceedings of the 10th International Conference on General Relativity and Gravitation, edited by B. Bertotti, F. De Filice,and A. Paseolini, Padua, Italy, 4–9 July, 1983, p. 1193; Cincinnati Preprint UCTP-102-83, 1983, and in preparation.
[9]See also B. Julia, Preprint LPTENS-83/22, Ecole Normale 1983, and the references cited therein.
[10]See, e.g., A.Salam and J. Strathdee, Ann. Phys. **141**, 316 (1982); G.Chapline and R. Slansky, Nucl. Phys. B **209**, 461 (1982); S. Weinberg, Phys. Lett. B **125**, 265 (1983).

Stochastic convergence of the saturation coverage of one-dimensional arrays of β-bell particles

B. Mellein
Universidad de los Andes, Apartado Aéreo 4976, Bogotá, Colombia

A. Maltz and E. E. Mola[a)]
INIFTA, Sección Química Teórica, Sucursal 4, Casilla de Correo 16, 1900 La Plata, Argentina

(Received 24 March 1983; accepted for publication 23 September 1983)

This paper considers the first two moments of $R(\beta,N) = A(\beta,N)/N$, $A(\beta,N)$ denoting the random number of unoccupied sites of a one-dimensional array of N compartments saturated by randomly placed β-bell particles ($\beta \geqslant 2$). It is shown that, as $N \to \infty$, the mean of $R(\beta,N)$ approaches a (nonzero) limit $L(\beta)$ while its variance tends to zero thus yielding the stochastic convergence of $R(\beta,N)$ to $L(\beta)$. $L(\beta)$ is explicitly determined and its behavior for large β is also studied.

PACS numbers: 05.50. + q, 02.50. + s, 82.65.Jv, 68.20. + t

1. INTRODUCTION

In this paper we will be concerned with the following model.[1,2] A linear array of N identical compartments is ceaselessly bombarded by particles which occupy β ($\beta \geqslant 2$) contiguous lattice sites. These so-called β-bell particles are assumed to contact in a spatially random manner and to stick only if striking β (adjacent) vacant compartments. Thus, as time elapses, the number of unoccupied compartments diminishes until finally being left only fragments of $1,2,\ldots,\beta - 1$ vacant compartments (see Fig. 1). As soon as this saturation situation arises, the total number $A(\beta,N)$ of unoccupied lattice sites may be determined.

Due to the nature of the model, $A(\beta,N)$ is a random variable. Its mean $a(\beta,N)$ has been seen[2] to satisfy the recursion relation

$$a(\beta,N) = \frac{N-\beta}{N-\beta+1}\, a(\beta,N-1) + \frac{2}{N-\beta+1}\, a(\beta,N-\beta), \quad N \geqslant \beta + 1, \tag{1}$$

with initial conditions $a(\beta,1) = 1$, $a(\beta,2) = 2,\ldots,a(\beta,\beta-1) = \beta - 1$, $a(\beta,\beta) = 0$ and in the case $\beta = 2$ it has been found[1] that

$$\lim_{N\to\infty} a(2,N)/N = \exp(-2).$$

Making use of (1) we will show in Sec. 2 that $L(\beta)$, the average fraction of vacant compartments of an infinite one-dimensional lattice space saturated by β-bell particles, is

$$L(\beta) = \lim_{N\to\infty} a(\beta,N)/N = 1 - \beta e^{-P(1,\beta)} \int_0^1 e^{P(z,\beta)}\, dz, \tag{2}$$

where

$$P(z,\beta) = 2\int_0^z \frac{1-x^{\beta-1}}{1-x}\, dx = 2\left(z + \frac{z^2}{2} + \cdots + \frac{z^{\beta-1}}{\beta-1}\right), \quad z > 0. \tag{3}$$

In Sec. 3 we will discuss the physically interesting case of large polymers, i.e., the case of large β. More precisely, starting out from (2) we will establish that L^*, the average fraction of occupied sites at saturation, is

$$\lim_{\beta\to\infty} L(\beta) = 1 - L^*, \tag{4}$$

where

$$L^* = \int_0^\infty \exp\left\{-2\int_0^t \frac{1-e^{-v}}{v}\, dv\right\} dt. \tag{5}$$

It is worthwhile to mention that Rényi[3] has shown that

$$\lim_{x\to\infty} F(x)/x = L^*, \tag{6}$$

where $F(x)$ is the mean number of unit intervals placed randomly (without intersecting each other) in the interval $(0,x)$. A connection between Rényi's continuous and the present discrete model has been observed recently by Maltz and Mola.[4] They have gotten the relation

$$\lim_{\beta\to\infty} L(\beta) = 1 - \lim_{x\to\infty} F(x)/x \tag{7}$$

by theoretical considerations.

In Sec. 4, we shall be concerned with $v(\beta,N)$, the mean of $A^2(\beta,N)$. Transforming a difference equation for $v(\beta,N)$ into a first-order linear differential equation for the generating function of the $v(\beta,N)$, $N = \beta + 1, \beta + 2,\ldots$, we find from the analytic expression of its solution, by way of a Tauberian argument, that

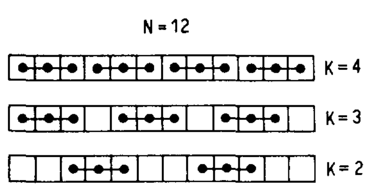


FIG. 1. Three of all possible saturated arrangements when trimers are placed on a linear array composed of 12 compartments.

[a)] To whom the correspondence related to this article should be sent.

 0022-2488/84/061995-05$02.50

$$\lim_{N\to\infty} v(\beta,N)/N^2 = L^2(\beta). \tag{8}$$

This together with (2) shows that, as N becomes large, the variance of $A(\beta,N)/N$ approaches zero, symbolically

$$\lim_{N\to\infty} \mathrm{Var}(A(\beta,N)/N) = 0.$$

The stochastic convergence of $A(\beta,N)/N$ to $L(\beta)$, i.e.,

$$\lim_{N\to\infty} P(|A(\beta,N)/N - L(\beta)| > \epsilon) = 0, \quad \epsilon > 0, \tag{9}$$

is an easy consequence of Eqs. (2), (8) and Chebychev's inequality: for any fixed $\epsilon > 0$

$$P(|A(\beta,N)/N - L(\beta)| > \epsilon) \leqslant \epsilon^{-2} E([A(\beta,N)/N - L(\beta)]^2)$$
$$= \epsilon^{-2}[v(\beta,N)/N^2 - 2L(\beta)a(\beta,N)/N + L^2(\beta)],$$

and (9) follows from Eqs. (2) and (8).

2. CALCULATION OF L(β)

Starting with the recursion relation (1) and with a corresponding one for $N - a(\beta,N)$, the average number of occupied sites of a β-bell particle saturated $1\times N$ array, and utilizing a Tauberian argument (as we will do in Sec. 4) we have shown[5] [for the definition of $P(z,\beta)$, see Eq. (3)]

$$L(\beta) = \lim_{N\to\infty} \frac{a(\beta,N)}{N} = e^{-P(1,\beta)} \int_0^1 2\left((1-x)\sum_{k=1}^{\beta-1} kx^k\right) e^{P(x,\beta)}\,dx \tag{10}$$

and Eq. (2), respectively. Clearly, as a consequence, the expressions on the right-hand sides of Eqs. (2) and (10) are equal. Adopting a different method we will reproduce the result given by Eq. (2).

As β remains fixed in what follows we shall use the notation $a_N \equiv a(\beta,N)$.

Setting

$$s_N = a_N - a_{N-1}, \quad N\geqslant 2, \quad s_1 = 1,$$

and thus obtaining

$$a_N = s_1 + s_2 + \cdots + s_N, \quad N\geqslant 1,$$

we first observe that Eq. (1) may be written as

$$s_N = \frac{2(N-\beta)}{(N-\beta+1)}\frac{a(\beta,N-\beta)}{(N-\beta)} - \frac{(N-1)}{(N-\beta+1)}\frac{a(\beta,N-1)}{(N-1)}, \quad N\geqslant \beta+1, \tag{11}$$

as well as

$$s_N = \frac{1}{N-\beta+1}[s_1 + s_2 + \cdots + s_{N-\beta} - (s_{N-\beta+1} + s_{N-\beta+2} + \cdots + s_{N-1})], \quad N\geqslant \beta+1. \tag{12}$$

Performing the limit $(N\to\infty)$ on both sides of Eq. (11) we find from Eq. (2) that

$$L(\beta) = \lim_{N\to\infty} s_N. \tag{13}$$

From Eq. (12) we deduce

$$s_N - s_{N-1} = \frac{2}{N-\beta+1}(s_{N-\beta} - s_{N-1}), \quad N\geqslant \beta+1.$$

Setting

$$r_N = s_N - s_{N-1}, \quad N\geqslant 2, \quad r_1 = 1$$

or equivalently

$$s_N = r_1 + r_2 + \cdots + r_N, \quad N\geqslant 1, \tag{14}$$

which in turn implies

$$r_N = -\frac{2}{N-\beta+1}(r_{N-1} + r_{N-2} + \cdots + r_{N-\beta+1}), \quad N\geqslant \beta+1$$

or

$$r_N = \frac{N-\beta-2}{N-\beta+1} r_{N-1} + \frac{2}{N-\beta+1} r_{N-\beta}, \quad N\geqslant \beta+2. \tag{15}$$

Now multiplying both sides of Eq. (15) by $(N-\beta+1)z^{N-\beta}$ and summing over $N = 2\beta$ to ∞ leads to

$$\sum_{N=2\beta}^{\infty}(N-\beta+1) r_N z^{N-\beta} = z\sum_{N=2\beta}^{\infty}(N-\beta) r_{N-1} z^{N-\beta-1} - 2\sum_{N=2\beta}^{\infty} r_{N-1} z^{N-\beta} + 2z^{\beta-1}\sum_{N=2\beta}^{\infty} r_{N-\beta} z^{N-2\beta+1}$$

or

$$H'(z) - r_\beta - 2r_{\beta+1} z = z[H'(z) - r_\beta - 2r_{\beta+1} z(1-\delta_{\beta 2})] - 2[H(z) - zr_\beta - z^2 r_{\beta+1}(1-\delta_{\beta 2})] + 2z^{\beta-1}H(z), \tag{16}$$

where we set

$$H(z) = \sum_{N=\beta}^{\infty} r_N z^{N-\beta+1}$$

and $\delta_{\beta 2} = 0$ if $\beta \neq 2$, $\delta_{\beta 2} = 1$ if $\beta = 2$.

Taking into account that $r_\beta = -\beta$ and $r_{\beta+1} = \beta$ (see

TABLE I. The initial values of a_i, s_i, r_i, v_i, and p_i.

i	a_i	s_i	r_i	v_i	p_i
0					0
1	1	1	1	1	0
2	2	1	0	4	1
3	3	1	0	9	4
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
$\beta-1$	$\beta-1$	1	0	$(\beta-1)^2$	
β	0	$1-\beta$	$-\beta$	0	
$\beta+1$	1	1	β	1	
$\beta+2$	2	1	0		
.	.	.	.		
.	.	.	.		
.	.	.	.		
$2\beta-1$	$\beta-1$	1	0		

Table I), then Eq. (16) reduces to the first order linear differential equation

$$H'(z) + 2\frac{1-z^{\beta-1}}{1-z}H(z) = -\beta$$

with initial condition $H(0) = 0$ whose solution is [for the definition of $P(z,\beta)$ see Eq. (3)]

$$H(z) = -\beta e^{-P(z,\beta)}\int_0^z e^{P(x,\beta)}\,dx. \tag{17}$$

From the definition of the generating function H, from Table I, Eqs. (13) and (15) we learn that

$$L(\beta) = 1 + H(1),$$

which in combination with Eq. (17) yields Eq. (2).

3. THE CASE OF LARGE PARTICLE SIZE

As mentioned in the Introduction Rényi's continuous model[3] has been found[4] to be—in some sense—a limit of the present lattice model. As a consequence, Eq. (7) has been established. Owing to Eqs. (6) and (7), Eq. (4) is proved true. Nevertheless it is interesting to see how Eq. (4) comes out proceeding from Eq. (2):

[MT

$$\begin{aligned}
1 - L(\beta) &= \beta\int_0^1 \exp\{P(x,\beta) - P(1,\beta)\}\,dx \\
&= \beta\int_0^1 \exp\left\{-2\int_x^1 \frac{1-z^{\beta-1}}{1-z}dz\right\}dx \quad \text{due to Eq. (3)} \\
&= \int_0^\beta \exp\left\{-2\int_{y/\beta}^1 \frac{1-z^{\beta-1}}{1-z}dz\right\}dy \quad \text{by substituting } y=\beta x \\
&= \int_0^\beta \exp\left\{-2\int_0^{\beta-y}\left[1-\left(1-\frac{v}{\beta}\right)^{\beta-1}\right]\frac{1}{v}dv\right\}dy \quad \text{by substituting } z = 1 - v/\beta \\
&= \int_0^\beta \exp\left\{-2\int_0^t\left[1-\left(1-\frac{v}{\beta}\right)^{\beta-1}\right]\frac{1}{v}dv\right\}dt \quad \text{by substituting } t = \beta - y
\end{aligned}$$

and sending β to infinity one gets the expression L^* stated in Eq. (5). The values of $L(\beta)$, $\beta = 2,3,\ldots$ range from $e^{-2} = 0.135\ldots$, to $0.251\ldots = 1 - L^*$.

4. THE ASYMPTOTIC BEHAVIOR OF $v(\beta,N)$

In this section we shall be concerned with the limit relation stated by Eq. (8). We shall break up its proof into a number of steps.

Throughout this section we will write a_N and v_N instead of $a(\beta,N)$ and $v(\beta,N)$, respectively.

A. Preliminaries

We shall need the following two lemmas. Their easy proofs will be omitted.

Lemma 1: For a real-valued function, continuous on $[0,1]$, the following holds:

$$\int_0^s \frac{f(t)}{(1-t)^2}dt \sim \frac{f(1)}{1-s} \quad \text{as } s \uparrow 1.$$

Lemma 2: Let $(c_n)_{n=1}^\infty$ be a sequence of real numbers such that $c_n \geqslant 0$ for any integer number n, and $\lim_{n\to\infty} c_n/n = c < \infty$. Then

$$\lim_{n\to\infty} n^{-3}\sum_{k=1}^n c_k c_{n-k} = \frac{c^2}{6}.$$

The proof of the succeeding result may be found, e.g., in Feller.[6]

Tauberian theorem for power series: Let $c_n \geqslant 0$ and suppose that $C(s) = \sum_{n=0}^\infty c_n s^n$ converges for $0 \leqslant s < 1$. If L varies slowly at infinity [L varies slowly at infinity if, for every fixed $x > 0$, $\lim_{t\to\infty} L(tx)/L(t) = 1$] and $0 \leqslant r < \infty$ then each of the two relations,

$$C(s) \sim (1-s)^{-r}L\left(\frac{1}{1-s}\right) \quad \text{as } s \uparrow 1$$

and

$$\sum_{k=0}^{n-1} c_k \sim \frac{n^r L(n)}{\Gamma(r+1)} \quad \text{as } n \to \infty,$$

implies the other.

B. A recursion relation for $v_N \equiv v(\beta,N)$

We will make use of the following well-known result on conditional expectations.[6] The mean of a random variable X can be expressed as the mean of a conditional expectation given a random variable Y, symbolically

$$E(X) = E[E(X\,|\,Y)].$$

Setting $X = A^2(\beta,N)$ and defining the random variable $Y \equiv Y(\beta,N)$ by

$Y = k$, if the *first* β-bell particle getting stuck to the $1\times N$ array occupies sites $k, k+1,\ldots, k+\beta-1$ (counted from the left), $k = 1, 2,\ldots, N-\beta+1$.

$Y = 0$, otherwise

we obtain for $N \geqslant \beta$

$$\begin{aligned}
v_N &= E[A^2(\beta,N)] = E[E(A^2(\beta,N)\,|\,Y)] \\
&= \sum_{k=0}^{N-\beta+1} E(A^2(\beta,N)\,|\,Y=k)P(Y=k) \\
&= \frac{1}{N-\beta+1}\sum_{k=1}^{N-\beta+1} E(A^2(\beta,N)\,|\,Y=k),
\end{aligned}$$

where the last equality is due to the assumption that the β-bell particles strike in a spatially random manner. Now, $Y \equiv Y(\beta,N) = k$, $k = 2,3,\ldots,N-\beta$, means that our initial

$1\times N$ array is subdivided into a $1\times(k-1)$ array and a $1\times(N-k-\beta+1)$ array. Thus

$$v_N=\frac{1}{N-\beta+1}\Big(\sum_{k=2}^{N-\beta}E([A(\beta,k-1)+A(\beta,N-k-\beta+1)]^2)+2E[A^2(\beta,N-\beta)]\Big),$$

where $A(\beta,k-1)$ and $A(\beta,N-k-\beta+1)$ are independent since placements of any kind being realized on the right-hand array, say, do not affect the odds of the occupation process of the left-hand array at all. Clearly, given $Y=k$ (with $k\geqslant\beta+1$), the (overall) probability that, for example, the sites $i, i+1, ..., i+\beta-1, i\in\{1,...,k-\beta\}$, of the left-hand $1\times(k-1)$-array will be occupied by some β-bell particle is altered by every placement of a particle onto the right-hand $1\times(N-k-\beta+1)$-array; however, when for the first time a position on the left-hand array is to be occupied, the (conditional) probability that the sites $i, i+1,... i+\beta-1$ will be chosen is equal to $1/(k-\beta)$ independent of what has happened on the right-hand array. Symbolically,

$$P_k(X_L=i|B_R)\equiv P_k(X_L=i)=\frac{1}{k-\beta},\quad i\in\{1,2,...,k-\beta\},$$

where B_R is any event which refers exclusively to the right-hand array (e.g., $A(\beta,N-k-\beta+1)=0$), X_L is the random variable giving the left-hand position of the first β-bell particle placed on the left-hand array and where the subscript k of P_k indicates that we are dealing with a conditional probability, given $Y=k$. Now observe that the upper argument may be iterated, i.e., equally well applicated successively to sub-subarrays of the $1\times(k-1)$-subarray, to yield the desired result

$$P_k(A(\beta,k-1)=r, A(\beta,N-k-\beta+1)=s)=P_k(A(\beta,k-1)=r)\times P_k(A(\beta,N-k-\beta+1)=s)$$

with $r,s\geqslant 0$. From this fact we get

$$v_N=\frac{1}{N-\beta+1}\Big[\sum_{k=2}^{N-\beta}(v_{k-1}+2a_{k-1}\times a_{N-k-\beta+1}+v_{N-k-\beta+1})+2v_{N-\beta}\Big]$$

or

$$v_N=\frac{2}{N-\beta+1}\Big(\sum_{k=1}^{N-\beta}v_k+\sum_{k=1}^{N-\beta-1}a_k a_{N-k-\beta}\Big),\quad N\geqslant\beta+1 \tag{18}$$

with initial conditions listed in Table I.

C. The generating function of $a_N\equiv a(\beta,N)$

Let

$$G(s)=\sum_{N=\beta+1}^{\infty}a_N s^{N-\beta+1},\quad A(s)=\sum_{N=1}^{\infty}a_N s^N, \tag{19}$$

and

$$Q(s,\beta)=\sum_{n=1}^{\beta-1}ns^n. \tag{20}$$

To get an analytic representation of G we multiply both sides of Eq. (1) by $(N-\beta+1)s^{N-\beta}$, sum over $N=\beta+1$ to ∞ and obtain the first-order linear differential equation

$$G'(s)-\frac{2s^{\beta-1}}{1-s}G(s)=\frac{2}{1-s}Q(s,\beta),$$

whose unique solution satisfying the initial condition $G(0)=0$ is

$$G(s)=2(1-s)^{-2}e^{-P(s,\beta)}\times\int_0^s(1-t)Q(t,\beta)e^{P(t,\beta)}\,dt. \tag{21}$$

Now consult Table I to confirm that

$$A(s)=Q(s,\beta)+s^{\beta-1}G(s). \tag{22}$$

D. The generating function of $v_N\equiv v(\beta,N)$

Let p_N be the coefficient of s^N in the Taylor series representation of A^2, where A is the generating function of the a_N introduced in Eq. (19). Then, clearly,

$$p_0=p_1=0,\quad p_N=\sum_{k=1}^{N-1}a_k a_{N-k},\quad N\geqslant 2, \tag{23}$$

and Eq. (18) may be rewritten as

$$v_N=\frac{2}{N-\beta+1}\Big(\sum_{k=1}^{N-\beta}v_k+p_{N-\beta}\Big),\quad N\geqslant\beta+1, \tag{24}$$

or equally well as

$$(N-\beta+1)v_N-(N-\beta)v_{N-1}-2v_{N-\beta}=2(p_{N-\beta}-p_{N-\beta-1}),\quad N\geqslant\beta+1 \tag{25}$$

with initial values to be found in Table I. Multiplying both sides of Eq. (25) by $s^{N-\beta}$ and summing over $\beta+1$ to ∞ leads to the first-order linear differential equation

$$V'(s)-\frac{2s^{\beta-1}}{1-s}V(s)=R(s) \tag{26}$$

for

$$V(s)=\sum_{N=\beta+1}^{\infty}v_N s^{N-\beta+1},$$

where we put

$$R(s)=2A^2(s)+\frac{2}{1-s}\sum_{n=1}^{\beta-1}n^2s^n. \tag{27}$$

The unique solution of Eq. (26) obeying the initial condition $V(0)=0$ is

$$V(s)=(1-s)^{-2}e^{-P(s,\beta)}\int_0^s R(t)(1-t)^2e^{P(t,\beta)}\,dt. \tag{28}$$

E. The behavior of $V(s)$ as $s\uparrow 1$

Substituting R in Eq. (28) and recalling Eqs. (20)–(22) and (27) we obtain

$$V(s)=(1-s)^{-2}e^{-P(s,\beta)}\times\int_0^s\{2[(1-t)t^{\beta-1}G(t)]^2e^{P(t,\beta)}+F(t)\}\,dt, \tag{29}$$

where $F(t)=o(1/(1-t)^2)$ as $t\uparrow 1$. Putting

$$f(t) = \int_0^t 2(1-x)Q(x,\beta)e^{P(x,\beta)}\,dx, \tag{30}$$

it follows from Eqs. (21) and (29), as $s \uparrow 1$

$$V(s) = 2(1-s)^{-2}e^{-P(s,\beta)}\int_0^s e^{-P(t,\beta)} \times \left[\frac{t^{\beta-1}}{1-t}f(t)\right]^2 dt + o(1/(1-s)^3).$$

An application of Lemma 1 then yields

$$V(s) \sim 2(1-s)^{-3}f^2(1)e^{-2P(1,\beta)}, \quad \text{as } s\uparrow 1. \tag{31}$$

F. An application of the Tauberian theorem

From Eq. (31) and the Tauberian theorem for power series we conclude

$$\sum_{k=1}^{N} v_k \sim \frac{N^3}{3}f^2(1)e^{-2P(1,\beta)} \quad \text{as } N\to\infty. \tag{32}$$

G. The asymptotic behavior of $v_N \equiv v(\beta,N)$

As stated in Eq. (10)

$$a_N \sim Nf(1)e^{-P(1,\beta)} \quad \text{as } N\to\infty. \tag{33}$$

Hence Lemma 2 together with Eqs. (23) and (33) implies

$$p_N \sim \frac{N^3}{6}f^2(1)e^{-2P(1,\beta)} \quad \text{as } N\to\infty,$$

which in turn, in combination with Eqs. (24) and (32), gives

$$v_N \sim N^2f^2(1)e^{-2P(1,\beta)} \quad \text{as } N\to\infty. \tag{34}$$

Recalling Eqs. (10), (20), and (30) one sees that Eq. (34) coincides with Eq. (8), the relation we wanted to prove.

ACKNOWLEDGMENTS

B. M. is grateful to COLCIENCIAS (Project 2004-1-37-81) and E. E. M. to Consejo Nacional de Investigaciones Científicas y Técnicas for financial support.

[1]R. B. McQuistan and D. Lichtman, J. Math. Phys. **9**, 1680 (1968).
[2]A. Maltz and E. E. Mola, Surf. Sci. **115**, 599 (1982).
[3]A. Rényi, Commun. Math. Res. Inst. Hung. Acad. Sci. **3**, 109 (1958).
[4]A. Maltz and E. E. Mola, *Stochastic Processes Applied to Physics and other Related Fields*, edited by B. Gómez *et al.* (World Scientific, Singapore, 1983), pp. 171–183.
[5]B. Mellein, A. Maltz, and E. E. Mola, *Stochastic Processes Applied to Physics and other Related Fields*, edited by B. Gómez *et al.* (World Scientific, Singapore, 1983), pp. 199–206.
[6]W. Feller, *An Introduction to Probability Theory and Its Applications 2*, 2nd ed. (Wiley, New York, 1971).

An explicit formula for the Casimir energy of an arbitrary compact smooth surface without boundary

Siddhartha Sen
School of Mathematics, Trinity College, Dublin, Ireland

(Received 13 October 1982; accepted for publication 23 September 1983)

An explicit formula for the finite part of the Casimir energy of a scalar field $\phi(\mathbf{x},t)$ defined on a smooth, compact two-dimensional Riemannian manifold X without boundary is derived. The formula involves the metric tensor on X and its derivatives.

PACS numbers: 11.10. – z, 02.40.Ky

I. INTRODUCTION

In this paper we will obtain an explicit formula for the finite part of the Casimir energy[1] of an arbitrary smooth surface X without boundary. By the Casimir energy of a surface we mean the following: Let the Laplacian on X be $-\nabla^2$ and let the eigenvalues and eigenfunctions of $-\nabla^2$ on X be λ_n^2 and $u_n(x)$, respectively, i.e.,

$$-\nabla^2 u_n(\mathbf{x}) = \lambda_n^2 u_n(\mathbf{x}). \tag{1.1}$$

The Casimir energy E_R is then defined to be the regularized sum:

$$E_R = \tfrac{1}{2}\hbar \sum \lambda_n \tag{1.2}$$

The methods used to calculate (1.2) have involved either knowledge of all the eigenvalues $\lambda_n{}^2$ or knowledge of the Green's function $G(\mathbf{x},\mathbf{y},\omega)$,[3] where

$$[-\nabla^2 - \omega^2] G(\mathbf{x},\mathbf{y},\omega) = \delta^D(\mathbf{x}-\mathbf{y}). \tag{1.3}$$

A Green's function method capable of dealing with arbitrary regions with boundaries has been discussed by Balian and Duplantier.[4] These authors were also able to obtain an explicit formula for the Casimir energy at high temperature involving one global parameter. In this paper an equally explicit formula for (1.2) in the case of a region without boundary will be obtained. The case of a region with a boundary is more complicated, and it is not clear that the methods of this paper can be generalized to deal with boundaries.

Our explicit formula for the Casimir energy is:

$$E_R = \tfrac{1}{2}\hbar \frac{1}{(2\pi)^3}\int_X d^2x \int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} d\tau \times \int_{-\infty}^{\infty} ds\, e^{is} C_{-4}(x,\xi,\tau,-is), \tag{1.4}$$

where $C_{-4}(x,\xi,\tau,-is)$ is an explicit function of the metric tensor g_{ij} of X and its derivatives (see Sec. IV and Appendix C). Thus once the surface X is given and the Laplacian is defined on X then E_R can be computed in terms of g_{ij} and its derivatives. In Sec. II we explain our method. This involves the pseudodifferential calculus. In Sec. III we briefly review the portions of the pseudodifferential calculus that we need. In Sec. IV we obtain our explicit formula, and in Sec. V we briefly summarize our results.

II. THE METHOD

We will proceed in a formal manner. The meaning of our expressions will become clear in terms of the pseudodifferential calculus which we briefly review in the Sec. III.

Let us introduce $P = +(-\nabla^2)^{1/2}$ as a positive elliptic operator on X. It is possible to prove that P has a discrete spectrum $\lambda_n = (\lambda_n^2)^{1/2}$ with corresponding orthonormal eigenfunction $u_n(x)$. It has been proved by Duistermaat and Guilleman[5] that

$$(2\pi)^n \Theta_p(t,x) \sim \sum_{\substack{K \neq n + lm \\ l \in N}} \Gamma(n-K)\omega_k(x)\cdot t^{(k-n)} + \sum_{\substack{l=1 \\ n+l \in Z}}^{\infty} \frac{(-1)^{l+1}}{l!}\omega_{n+lm}(x) t^l \ln t + \sum_{l=0}^{\infty} \nu_1(x) t^l, \quad \text{for } t \downarrow 0, \tag{2.1}$$

where

$$\Theta_p(t,x) \equiv \sum_m e^{-\lambda_m t} |u_m(x)|^2, \tag{2.2}$$

ω_k, ν_k, are smooth functions of x and n represents the dimension of the smooth compact boundary less manifold X. On the other hand,

$$\Theta_p(t,x) = \frac{1}{(2\pi i)} \int_{\mathrm{Re}\, s = c} t^{-s} \zeta_p(s,x)\Gamma(s)\, ds \tag{2.3}$$

with $\zeta_p(s,x) \equiv \Sigma_m [1/(\lambda m)^s] |u_m(x)|^2$ holds for c sufficiently large. From this formula it can be shown quite easily[5] that the terms in (2.1) correspond to the poles of $\zeta_p(s,x)\Gamma(s)$. In particular, the logarithmic terms correspond to terms where the poles of $\zeta_p(s,x)$ and $\Gamma(s)$ coincide to produce a double pole, while the $\nu_k(x)$ terms are related in a one-to-one way to the values of $\zeta_p(s,x)$ at $s = -k$ [after taking off of the pole of $\zeta_p(s,x)$ if necessary]. The fact that the singularity of $\zeta_p(s,x)$ in general consists of simple poles at $s = n - k$, $k = 0,1,2,\ldots$, with residues equal to $(2\pi)^{-n}\omega_k$, $k = 0,1,2,\ldots$, follows from a result of Duistermaat and Guilleman,[5] which we quote:

Theorem 2.1. $\zeta_p(s,x)$ has a meromorphic extension to the complex plane having only simple poles at $s = (n-k)$ with residues equal to $(2\pi)^{-n}\omega_k$, $k = 0,1,2,\cdots$. Moreover, $\zeta_p(s,x)$ is of at most polynomial growth on each half-space $\mathrm{Re}\, s \geqslant c$ excluding neighborhoods of poles.

We can thus define the regularized value of $\zeta_p(s,x)$ at $s = -1$ to be $\nu_1(x)$, i.e.,

$$E_R = \sum_R \frac{1}{2}\hbar\lambda_n \equiv \frac{1}{2}\hbar \int_X \nu_1(x)\, d^2x. \tag{2.4}$$

As we have shown elsewhere[6] E_R represents the finite part of

 0022-2488/84/062000-07$02.50

the Casimir energy. Our aim will thus be to find an asymptotic expansion for $\Theta_p(t,x)$ and determine the coefficient $\nu_1(x)$

III. REVIEW OF PSEUDO-DIFFERENTIAL CALCULUS[7]

In this section we briefly review those portions of the pseudodifferential calculus that we need in order to calculate $\nu_1(x)$. We recall that a linear differential operator is a polynomial expression

$$P(x,D) = \sum_{|\alpha|<m_n} a_\alpha(x) D_x^\alpha, \tag{3.1}$$

where we utilize the multi-index notation, i.e.,

$$D_x^\alpha = \left(\frac{1}{i}\right)^{|\alpha|} \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \left(\frac{\partial}{\partial x_2}\right)^{\alpha_2} \cdots \left(\frac{\partial}{\partial x_N}\right)^{\alpha_N}, \tag{3.2}$$

when $x \in R^N$ and the coordinates of X are $(x_1,x_2,...,x_N)$. We will also use the notation

$$|\alpha| \equiv \sum_{i=1}^{N} \alpha_i, \quad \alpha! \equiv \alpha_1!\alpha_2!\cdots\alpha_N!,$$

when $(\alpha_1,\alpha_2,...,\alpha_N)$ are all integers. The class of functions on which the operators of interest to us act will be defined[7] as follows. Consider the class of function $S(R^n) = \{u \in C^\infty(R^N); \forall(\alpha,n) \ni \text{constants } C_{\alpha,n} \text{ with } |D^\alpha u| \leqslant C_{\alpha,n}(1+x)^{-n}\}$. For s a real number and $u \in S$ define

$$\|u\|_s = \left(\int \hat{u}(\xi)^2 (1+|\xi|^2)^s \, d_\xi^N\right)^{1/2}. \tag{3.3}$$

The space $H_s(R^N)$ defined to be the completion of $S(R^N)$ in the norm $\|u\|_s$ is the space of the functions on which our operators of interest will act. $\hat{u}(\xi)$ is the Fourier transform of $u(x)$, and is defined as

$$u(x) = \int e^{ix}\hat{u}(\xi)\, d_\xi^N. \tag{3.4}$$

Using (3.4) the action of $P(x,D)$ given by (3.1) on (3.4) is clearly

$$P(x,D)u(x) = \int e^{ix} p(x,\xi)\hat{u}(\xi) d_\xi^N, \tag{3.5}$$

where $p(x,\xi) = \Sigma_{|\alpha|<n} a_\alpha(x)\xi^\alpha, p(x,\xi)$, is called the symbol of the linear differential operator $P(x,D)$ and is written as $\sigma(P) = p(x,\xi)$. A pseudo-differential operator ("$\psi_{\rm DO}$") can now be defined as an operator whose symbol $p(x,\xi)$ need not by a polynomial. We have the following definition:

Definition 3.1: $p(x,\xi)$ is said to be a symbol of order m, written $p \in S^m$, if

(1) $p(x,\xi)$ is C^∞ in x and ξ,

(2) $p(x,\xi)$ has compact x support,

(3) for all α,β, there is a constant $C_{\alpha,\beta}$ such that

$$|D_x^\alpha D^\beta p(x,)| \leqslant C_{\alpha\beta}(1+|\xi|)^{m-\beta}.$$

Definition 3.2: $P(x,D) = \int e^{ix} p(x,\xi)\hat{u}(\xi)\, d^N\xi$ is a pseudodifferential operator of order m if $p(x,\xi)$ is a symbol of order m. A useful equivalence relation[8] on the class of symbols can be introduced by defining $q_1 \sim q_2$, $q_1 - q_1 \in S^m$ for all m, positive or negative. According to a theorem of Sobolev[7] $(q_1 - q_2)$ represents on operator which acting on a function u changes into a C^∞ function ("infinitely smoothing"). We now state the following theorem that we will need in the next section.

Theorem 3.1: Let $g_j \in S^{m_j}$, where m_j decreases to $-\infty$ and we suppose that the support of $q_j(x,\xi)$ in x is contained in some fixed compact set for all j. Then there is a *unique*

$$q \sim \sum_{j=0}^{\infty} q_j.$$

The crucial property of pseudodifferential operators that we need for computations will now be stated in the form of a theorem.

Theorem 3.2: Let ψ_K denote the class of pseudodifferential operators restricted to functions with support in a fixed compact set K. For $P \in \psi_K$, let P^* be defined by $(P^*u,\vartheta) = (u,P\vartheta)$ [inner product on L^2, where $(u,\vartheta) = \int u(x)\bar{\vartheta}(x) d^N x$]. Then

(1) $P \in \psi_K \Rightarrow P^* \in \psi_K$ and $\sigma(P^*) \sim \Sigma_\alpha D_\xi^\alpha (iD_x^\alpha) p^*/\alpha!$, where $(p(x,\xi))^* = \overline{p(x,\xi)}$, where $\sigma(P) = p$,

(2) $P,Q \in \psi_K \Rightarrow PQ \in \psi_K$ and $\sigma(PQ) = \sigma(P) \circ \sigma(Q) = \Sigma_\alpha [D_\xi^\alpha p]((iD_x)^\alpha q)$, where $\sigma(P) = p$, $\sigma(Q) = q$.

For our calculations we will be interested in pseudodifferential operators which, besides being closed under composition and adjoint, also contain invertible elements. There are the elliptic pseudodifferential operators and are defined as follows:

Definition 3.3: A symbol $p(x,\xi) \in S^m$ is said to be elliptic if $p(x,\xi)^{-1}$ exists and there is a constant such that $p(x,\xi)^{-1} \leqslant C(1+|\xi|)^{-m}$, for large ξ. A pseudodifferential operator is said to be elliptic if its symbol p is elliptic. The following lemma which we quote is useful.

Lemma 3.1: If $p(x,\xi) \in S^m$ is elliptic and $p^{-1}(x,\xi)$ exists for $|\xi| \geqslant c$, then $p^{-1}(x,\xi)$ can be extended in a C^∞ way to $q(x,\xi) \in S^{-m}$ so that $pq \sim qp \sim 1$.

We also have the following theorem.

Theorem 3.3: If P is an elliptic operator, then there is a *unique* pseudodifferential operator Q such that $PQ \sim I \sim QP$.

As our final bit of information regarding pseudodifferential operators we introduce the following definition:

Definition 3.4: Let $S^m = p(x,\xi,\lambda); p$ is C^∞ in x and ξ has compact x support, analytic in λ for $\lambda \in C - R^+$ and $|D_\xi^\alpha D_x^\alpha p| \leqslant C_{\alpha\beta}(1+|\xi|+|\lambda|^{1/m})^{m-|\alpha|}$, then P is said to be a pseudodifferential operator of order m depending on the parameter λ if $\sigma(P_\lambda) = p \in S_\lambda^m$.

The theorems and properties of pseudodifferential operators with symbol $p \in S^m$ stated by us extend without a murmur to psuedodifferential operators with symbol $p \in S_\lambda^m$. This concludes our review of the calculus of pseudodifferential operators.

IV. CALCULATING $\nu_1(x)$

In this section we apply the formalism of the pseudodifferential calculus reviewed in Sec. III to determine the coefficient $\nu_1(x)$ for a compact, two-dimensional, smooth Reimannian manifold X without boundary. We suppose the Laplacian $-\nabla^2$ on X, in terms of local coordinates (x_1,x_2) takes the form

$$A^2 = -\nabla^2 = -G_{11}(x_1,x_2)\frac{\partial^2}{\partial x_1^2} - G_{12}(x_1,x_2)\frac{\partial^2}{\partial x_1 \partial x_2} - G_{22}(x_1,x_2)\frac{\partial^2}{\partial x_2^2} + G_1(x_1,x_2)\frac{1}{i}\frac{\partial}{\partial x_1} + G_2(x_1,x_2) \times \frac{1}{i}\frac{\partial}{\partial x_2} + G_0(x_1,x_2), \quad (4.1)$$

where $G_{11},G_{12},G_{22},G_{12},G_2,G_0$ are smooth funtions of (x_1,x_2). In Appendix A, G_{11},G_{22}, etc., are explicitly related to the metric tensor g_{ij} of X. The symbol of the operator A^2 is given by

$$\sigma(A^2) = G_{11}\xi^2 + G_{12}\xi\tau + G_{22}\tau^2 + G_1\tau + G_2\xi + G_0, \quad (4.2)$$

where ξ,τ are the Fourier transform variables of $u(x_1,x_2)$ defined by the equation

$$u(x_1,x_2) = \int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} d\tau\, e^{ix_1\xi}e^{ix_2\tau}\hat{u}(\xi,\tau). \quad (4.3)$$

Our procedure for determining the coefficient $\nu_1(x)$ will involve the following steps.

1. We will determine the symbol of the operator A, $\sigma(A)$, which has the property that $\sigma(A)\circ\sigma(A) = \sigma(A^2)$. This will be done by assuming that $\sigma(A)$ can be written as $\Sigma_{k=1}^{-\infty} b_K(x,\xi)$, where $b_j(x,\xi,\tau)$ is an element of $S^j, j\downarrow -\infty$. In view of Theorem 3.1, this representation of $\sigma(A)$ is unique, modulo infinitely smoothing terms.[7]

2. Once $\sigma(A)$ has been obtained, we then proceed to determine the inverse of $\sigma(A-\lambda I)$. Again this involves introducing $\sigma(B,\lambda) = \Sigma C_K(x,\xi,\tau,\lambda)$ such that $\sigma(A-\lambda I)\circ\Sigma C_K(x,\xi,\tau,\lambda) = I$. The operator thus determined is again unique in view of Theorems 3.1 and 3.3.

3. Finally writing $e^{-tA} = (1/2\pi i)\int_c [e^{-t\lambda}/(A-\lambda I)] \times d\lambda$ and replacing $(A-I)^{-1}u(x)$ by $\int\sigma(B,\lambda)e^{ix\cdot\xi}\,d^2\xi$, an asymptotic expansion for e^{-tA} is obtained involving only powers of t. No logarithmic terms appear in this procedure. The coefficient of t^1 is provisionally identified with $\nu_1(x)$. By referring to the work of Wodzicki[8] the absence of logarithmic terms is explained and the determination of $\nu_1(x)$ is completed.

Let us now proceed to determine $\sigma(A)$. By using the pseudodifferential calculus product law, we can write

$$\sigma(A)\circ\sigma(A) = \sum_{K=2,1,0} a_K(x,\xi,\tau) \quad (4.4)$$

with

$$\sigma(A) = \sum_{j=1}^{-\infty} b_j(x,\xi,\tau).$$

This gives the set of equations:

$$b_1(x,\xi,\tau)b_{j+1}(x,\xi,\tau) + \sum_{\substack{k\\ |l|<|j+1| \\ k-|\alpha|+l=2+j}} [D_\xi^\alpha b_k(x,\xi,\tau)] \times [(iD_x)^\alpha b_l(x,\xi,\tau)] = a_{2+j}(x,\xi,\tau), \quad (4.5)$$

where

$$j = 0, -1, -2, \cdots.$$

Thus

$$b_1 \equiv b_1(x,\xi,\tau) = +(G_{11}\xi^2 + G_{12}\xi\tau + G_{22}\tau^2)^{1/2}, \quad (4.6)$$

$$b_0 \equiv b_0(x,\xi,\tau) = \frac{G_1\xi + G_2\tau}{(G_{11}\xi^2 + G_{12}\xi\tau + G_{22}\tau^2)^{1/2}}, \quad (4.7)$$

$$b_{-1} \equiv b_{-1}(x,\xi,\tau) = \frac{G_0}{b_1} - \frac{b_0^2}{b_1} - [D_\xi^1 b_1][(iD_x)^1 b_0], \quad (4.8)$$

$$b_{-2} = b_{-2}(x,\xi,\tau) = (-)\frac{1}{b_1} \times \{2b_0b_{-1} + [D_\xi^1 b_1][(iD_x)^1 b_{-1}] + [D_\xi^1 b_0][(iD_x)^1 b_0] + [D_\xi^2 b_0][(iD_x)^2 b_0]/2!, \quad (4.9)$$

etc. The explicit form of b_{-1},b_{-2} in terms of G_{11},G_{22}, etc., is given in Appendix B.

We next determine the inverse of the operator $(A-\lambda I)$. We write

$$\tilde{b}_1(x,\xi,\tau,\lambda) = b_1(x,\xi,\tau) - \lambda I \quad (4.10)$$

and, for $j\neq 1$, $\tilde{b}_j(x,\xi,\tau,\lambda)$ represents the symbols generated by using (3.7), (3.8), (3.9),... starting with $\tilde{b}_1$. Thus $\tilde{\Sigma}_{K=1}^{\infty}\tilde{b}_K$ represent a pseudodifferential operator depending on a parameter λ as can easily be established.

Our problem is to determine the symbols C_K chosen to be homogeneous in ξ, τ, and λ of degree K so that

$$\sigma(A-\lambda I)\circ\sum_{j=0}^{\infty} C_{-1-j} = 1. \quad (4.11)$$

This implies the following:

$$C_{-1} = 1/\tilde{b}_1, \quad (4.12)$$

$$C_{-2} = (-)(1/\tilde{b})\{\tilde{b}_0C_{-1} + [D_\xi^1\tilde{b}_1][(CD_x)^1C_{-1}], \quad (4.13)$$

$$C_{-3} = (-)(1/\tilde{b}_1)\{\tilde{b}_0C_{-2} + \tilde{b}_1C_{-1} + [D_\xi^1\tilde{b}_0][(iD_x)^1C_{-1}] + [D_\xi^2\tilde{b}_1][(iD_x)^2C_{-1}]/2! + [D_\xi^1\tilde{b}_1][iD_x)^1C_{-2}], \quad (4.14)$$

$$C_{-4} = (-)(1/\tilde{b}_1)\{\tilde{b}_0C_{-3} + \tilde{b}_{-1}C_{-2} + \tilde{b}_{-2}C_{-1} + [D_\xi^1\tilde{b}_1][(iD_x)^1C_{-3}] + [D_\xi^1\tilde{b}_0][(iD_x)^1C_{-2}] + [D_\xi^1\tilde{b}_{-1}] \times [(iD_x)^1C_{-1}] + [D_\xi^2\tilde{b}_0][(iD_x)^2C_{-1}]/2! + [D_\xi^2\tilde{b}_1][(iD_x)^2C_{-2}]/3! + [D_\xi^3\tilde{b}_1][(iD_x)^3C_{-1}]/3!\}, \quad (4.15)$$

etc. Thus C_{-4} is a function of $G_{11},G_{22},\cdots$ and their derivatives. The asymptotic expansion $\mathrm{Tr}(e^{-tA})$ can now be determined as follows. Let us write

$$e^{-tA} = \frac{1}{2\pi i}\int_c e^{-t\lambda}(A-\lambda I)^{-1}\,d\lambda \quad (4.16)$$

and define

$$E(t) = \frac{1}{2\pi i}\int_c e^{-t\lambda}B_\lambda\,d\lambda, \quad (4.17)$$

where

$$\sigma(B_\lambda = \sum_{K=-1}^{\infty} C_K(x,\xi,\tau,\lambda). \quad (4.18)$$

We write

$$\mathrm{Tr}\,(E(t)) = \int L(t,x,x)\frac{d^2x}{(2\pi)^2},$$

where

$$L(t,x,y) = \int e^{i(x-y)\xi}\sigma(E(t))\,d\xi, \quad (4.19)$$

with $\sigma((E(t))$ the symbol of the operator $E(t)$. Thus

$$L(t,x,x) = \sum_{K=0}^{N}\sum_{J}\int \frac{C_{K,j}(x,\xi)t^{j-1}}{(j-1)!}e^{-ta_1(x,\xi,\tau)}d\xi d\tau, \quad (4.20)$$

where we have written

$$C_K(x,\xi,\tau)\equiv C_K(x,\xi) = \sum \frac{C_{K,j}(x,\xi)}{(a_1-\lambda)^j} \quad (4.21)$$

and have used the residue theorem to write

$$\int_c e^{-t\lambda}(a_1-\lambda)^{-j}d\lambda = (2\pi i)\frac{t^{j-1}}{(j-1)!}e^{-ta_1(x,\xi,\tau)}. \quad (4.22)$$

Proceeding in this way, the desired asymptotic expansion for $\mathrm{Tr}(e^{-tA})$ involving powers of t only can be obtained.[7] In particular the coefficient of t^1 is given by

$$\nu_1(x_1,x_2) = \frac{1}{(2\pi)^3}\int_{-\infty}^{\infty}d\xi\int_{-\infty}^{\infty}d\tau\int_{-\infty}^{\infty}ds \times e^{is}C_{-4}(x_1,x_2,\xi,\tau,-is). \quad (4.23)$$

The absence of logarithmic terms in the asymptotic expansion is explained by the work of Wodzicki.[8] Wodzicki's result[8] for an elliptic ψ_{DO} A with positive eigenvalues and order unity is the statement that $\zeta_A(s) = \mathrm{Tr}(A^{-s})$ is a meromorphic function in the whole s place with simple poles at the points $s_j = (2-j), j = 0,1,2,\cdots$. Furthermore, the residues at $s = 0, -1, -2,\cdots$, all vanish, i.e., $\zeta_A(s_j)$ is a nonsingular object. Thus no double poles are present in $\varphi_A(s)\Gamma(s)$ for $s = 0,-1,-2,\cdots$; hence no logarithmic terms appear in the asymptotic expansion for $\Theta(x,t)$. The Casimir energy for the surface X can thus be defined to be

$$E_R = \frac{1}{(2\pi)^3}\frac{1}{2}\hbar\int_x d^2x\int_{-\infty}^{\infty}d\xi\int_{-\infty}^{\infty}d\tau\int_{-\infty}^{\infty}ds \times e^{is}C_{-4}(x,\xi,-is). \quad (4.24)$$

This is our desired explicit formula for the finite part of the Casimir energy.

V. SUMMARY AND DISCUSSIONS

We have been able to obtain the finite part of the Casimir energy of a compact, boundaryless, smooth Reimannian manifold X in terms of the coefficients G_1, G_{22}, G_1, G_2, and G_0 introduced in Sec. IV. In a future publication we will discuss the geometrical nature of Eq. (4.24) and apply the equation to calculate the Casimir energy of simple surfaces.

It would appear from our calculation that for all pseudodifferential operators which can be represented in the form $\Sigma_{K=m}^{-\infty}a_K(x,\xi,\lambda)$, logarithmic terms will not be present in the asymptotic expansion for the corresponding $\Theta(x,t)$. This agrees with Wodzicki's result.[8] It is also clear that the method outlined in this paper can be trivially extended to D-dimensional compact manifolds without boundaries. The problem of obtaining a similar general formula for a manifold with a boundary is much more complex. One of the major problems in trying to obtain such a formula must be the problem of finding suitable boundary conditions for the elliptic pseudodifferential operator equation of interest. Simply if $-\nabla^2 u_n(x) = \lambda_n^2 u_n(x)$ is the Laplacian eigenvalue problem in a region Γ, with boundary $\partial\Gamma$, where the eigenvalues correspond to the Dirichlet boundary condition $u_n(x) = 0, x\in\partial\Gamma$. The corresponding pseudodifferential operator problem $(-\nabla^2)^{1/2}\nu_n = \mu_n\nu_n$ will not give $\mu_n = +\sqrt{\lambda_n^2}$ if the boundary condition $\nu_n(x) = 0$, for $x\in\partial\Gamma$ is used. A trivial example will make this point clear. Suppose our equation of interest is $(-d^2/dx^2)u_n(x) = \lambda_n^2 u_n(x)$, where $0\leqslant x\leqslant L$ and $u_n(x) = 0$, when $x = 0$ and $x = L$. The eigenvalues λ_n^2 are then given by $\lambda_n^2 = n^2\pi^2/L^2$. If we now consider the "square-root" equations,

$$\frac{d}{dx}\nu_n^{\pm}(x) = \pm i\lambda_n\nu_n^{\pm}(x), \quad (5.1)$$

and require $\nu^{\pm}(x) = 0$, $x = 0$, and $x = L$. We find $\nu_n^{\pm}(x) = 0$. To determine the right boundary conditions even in this trivial example, so that $\lambda_n = n\pi/L$ is "natural," involves choosing either $\nu_n^{\pm}(x)$ or $\nu_n^{-}(x)$, which is a nonlocal operation, and requiring, for instance, $|\nu^+(0)| = |\nu^+(L)|$. Finding and dealing with similar nonlocal constraints in the general case is one of the problems that has to be tackled when regions Γ with arbitrary smooth boundaries $\partial\Gamma$ are considered.[9]

ACKNOWLEDGMENTS

I would like to thank E. Onofori for drawing my attention to the work of Wodzicki (Ref. 8) and Duistermaat and Guilleman (Ref. 5). The hospitality and financial support of the CERN Theory Division, where this work was started, is also gratefully acknowledged.

APPENDIX A: RELATION BETWEEN G_{11}, G_{22}, G_1, G_2, G_0, AND THE METRIC TENSOR $g_{ij}(x_1,x_2)$

If we suppose

$$(ds)^2 = \sum_{i,j=1,2} g_{ij}\,dx^i\,dx^j, \quad g_{12} = g_{21}. \quad (\mathrm{A1})$$

Then

$$\nabla^2 \equiv \frac{1}{\sqrt{|g|}}\frac{\partial}{\partial x^i}\left(\sqrt{|g|}g^{iK}\frac{\partial}{\partial x^K}\right), \quad (\mathrm{A2})$$

where

$$g \equiv g_{11}g_{22} - (g_{12})^2 \quad (\mathrm{A3})$$

with

$$g^{11} = g_{22}/g, \quad g^{22} = g_{11}/g, \quad g_{12} = (-)g_{12}/g.$$

Thus

$$G_{11} = g_{22}/g, \quad \text{assuming } g > 0,$$
$$G_{22} = g_{11}/g,$$
$$G_{12} = (-)2g_{12}/g,$$
$$\frac{G_1}{i} = \frac{1}{\sqrt{g}}\left[\frac{\partial}{\partial x^1}\left(\frac{g_{22}}{\sqrt{g}}\right) - \frac{\partial}{\partial x^2}\left(\frac{g_{12}}{\sqrt{g}}\right)\right], \quad \frac{G_2}{i} = \frac{1}{\sqrt{g}}\left[\frac{\partial}{\partial x^2}\left(\frac{g_{11}}{\sqrt{g}}\right) - \frac{\partial}{\partial x^1}\left(\frac{g_{12}}{\sqrt{g}}\right)\right]. \tag{A4}$$

APPENDIX B: EXPLICIT EXPRESSIONS FOR $\tilde{b}_1, \tilde{b}_0, \tilde{b}_{-1}, \tilde{b}_{-2}$ IN TERMS OF $G_{11}, G_{12}, G_{22}, G_1, G_2, G_0$

$$\tilde{b}_1 = [G_{11}\xi^2 + G_{22}\tau^2 + G_{12}\xi\tau]^{1/2} - \lambda, \tag{B1}$$

$$\tilde{b}_0 = \frac{(G_1\xi + G_2\tau)}{[G_{11}\xi^2 + G_{22}\tau^2 + G_{12}\xi\tau]^{1/2} - \lambda}, \tag{B2}$$

$$\begin{aligned}\tilde{b}_{-1} = {} & \frac{G_0}{\tilde{b}_1} - \frac{(G_1\xi + G_2\tau)^2}{(\tilde{b}_1)^3} + \frac{i}{2}\frac{(2G_{11}\xi + G_{12}\tau)(G_{1,1}\xi + G_{2,1}\tau)}{(\tilde{b}_1)^2} \\ & + \frac{i}{2(\tilde{b}_1)^2}(2G_{22}\tau + G_{12}\xi)(G_{1,2}\xi + G_{2,2}\tau) - \frac{i}{4(\tilde{b}_1)^4}(G_1\xi + G_2\tau)(2G_{11}\xi + G_{12}\tau)(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau) \\ & - \frac{i}{4(\tilde{b}_1)^4}(G_1\xi + G_2\tau)(2G_{22}\tau + G_{12}\xi)(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau),\end{aligned} \tag{B3}$$

$$\begin{aligned}\tilde{b}_{-2} = {} & (-)\frac{1}{\tilde{b}_1}\Bigg[2\tilde{b}_0\tilde{b}_{-1} + \frac{1}{2i}\frac{1}{\tilde{b}_1}(2G_{11}\xi + G_{12}\tau)\Bigg(\frac{G_{0,1}}{\tilde{b}_1} - \frac{G_0}{2(\tilde{b}_1)^3}(G_{1,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau) \\ & - 2\tilde{b}_0\frac{(G_{1,1}\xi + G_{2,1}\tau)}{(\tilde{b}_1)^2} + \frac{\tilde{b}_0}{(\tilde{b}_1)^4}(G_1\xi + G_2\tau)(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau) \\ & + \frac{(\tilde{b}_0)^2}{2(\tilde{b}_1)^3}(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau) - \frac{i}{2(\tilde{b}_1)^4}(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau)(2G_{11}\xi + G_{12}\tau)(G_{1,1}\xi + G_{2,1}\tau) \\ & + \frac{i}{2(\tilde{b}_1)^2}(2G_{11,1}\xi + G_{12,1}\tau)(G_{1,1}\xi + G_{2,1}\tau) + \frac{i}{2(\tilde{b}_1)^2}(2G_{11}\xi + G_{12}\tau)(G_{1,1,1}\xi + G_{2,1,1}\tau) \\ & + \frac{i}{2(\tilde{b}_1)^2}(2G_{22,1}\tau + G_{12,1}\xi)\times(G_{1,2}\xi + G_{2,2}\tau) - \frac{i}{2(\tilde{b}_1)^4} \\ & \times(G_{11,1}\xi^2 + G_{12,1}\xi\tau + G_{22,1}\tau^2)(G_{12}\xi + 2G_{22}\tau)(G_{1,2}\xi + {}_{2,2}\tau) \\ & + \frac{i}{2(\tilde{b}_1)^2}(2G_{22}\tau + G_{12}\xi)(G_{1,2,1}\xi + G_{2,2,1}\tau) \\ & + \frac{i}{(2\tilde{b}_1)^6}(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau)^2(G_1\xi + G_2\tau)(2G_{11}\xi + G_{12}\tau) \\ & + \frac{i}{2(\tilde{b}_1)^6}(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau)(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau)(G_1\xi + G_2\tau)(2G_{22}\tau + G_{12}\xi) \\ & - \frac{i}{4(\tilde{b}\,)^4}((G_{1,1}\xi + G_{2,1}\tau)((2G_{11}\xi + G_{12}\tau)(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau)(2G_{22}\tau + G_{12}\xi) \\ & \times(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau)) + (G_1\xi + G_2\tau))\Big((2G_{11,1}\xi + G_{12,1}\tau)(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau) \\ & + (2G_{11}\xi + G_{12}\tau)(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1,1}\xi\tau) \\ & + (2G_{22,1}\tau + G_{12,1}\xi)(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau) + (2G_{22}\tau + G_{12}\xi)(G_{11,2,1}\xi^2 + G_{22,2,1}\tau^2 + G_{12,2,1}\xi\tau)\Big) \\ & + \frac{1}{4i}\frac{1}{\tilde{b}_1}(2G_{22}\tau + G_{12}\xi)\Bigg(\frac{G_{0,2}}{\tilde{b}_1} - \frac{G_0}{2(\tilde{b}_1)^3}(G_{1,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau) - \frac{2\tilde{b}_0}{(\tilde{b}_1)^2}(G_{1,2}\xi + G_{2,2}\tau) \\ & + \frac{\tilde{b}_0}{(\tilde{b}_1)^4}(G_1\xi + G_2\tau)(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau) + \frac{(\tilde{b}_0)^2}{2(\tilde{b}_1)^3}(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau)\Bigg) \\ & - \frac{i}{2(\tilde{b}_1)^4}(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau)((2G_{11}\xi + G_{12}\tau)(G_{1,1}\xi + G_{2,1}\tau) + (2G_{22}\tau + G_{12}\xi)(G_{1,2}\xi + G_{2,2}\tau)) \\ & + \frac{i}{2(\tilde{b}_1)^2}((2G_{11,2}\xi + G_{12,2}\tau)(G_{1,1}\xi + G_{2,1}\tau) + (2G_{11}\xi + G_{12}\tau)(G_{1,1,2}\xi + G_{2,1,2}\tau) \\ & + (2G_{22,2}\tau + G_{12,2}\xi)(G_{1,2}\xi + G_{2,2}\tau) + (2G_{22}\tau + G_{12}\xi)(G_{1,2,2}\xi + G_{2,2,2}\tau)) + \frac{i}{2(\tilde{b}_1)^6}(G_{11,2}\xi^2 + G_{22,2}\tau^2 \\ & + G_{12,2}\xi\tau)((G_1\xi + G_2\tau)(2G_{11}\xi + G_{12}\tau)(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau)\end{aligned}$$

$$
\begin{aligned}
&+(G_1\xi+G_2\tau)(2G_{22}\tau+G_{12}\xi)(G_{11,2}\xi^2+G_{22,2}\tau^2+G_{12,2}\xi\tau))\\
&-\frac{i}{4(\tilde b_1)^4}((G_{1,2}\xi+G_{2,2}\tau)((2G_{11}\xi+G_{12}\tau)(G_{11,1}\xi^2+G_{22,1}\tau^2+G_{12,1}\xi\tau)+G_{12,2}\xi\tau))\\
&+(2G_{22}\tau+G_{12}\xi)(G_{11,2}\xi^2+G_{22,2}\tau^2+G_{12,2}\xi\tau))+(G_1\xi+G_2\tau)((2G_{11,2}\xi+G_{12,2}\tau)(G_{11,1}\xi^2+G_{22,1}\tau^2+G_{12,1}\xi\tau)\\
&+(2G_{11}\xi+G_{12}\tau)(G_{11,1,2}\xi^2+G_{22,1,2}\tau^2+G_{12,1,2}\xi\tau)+(2G_{22,2}\tau+G_{12}\xi_2)\times(G_{11,2}\xi^2+G_{22,2}\tau^2\\
&+G_{12,2}\xi\tau)+(2G_{22}\tau+G_{12}\xi)(G_{11,2,2}\xi^2+G_{22,2,2}\tau^2+G_{12,2,2}\xi\tau))+\frac{1}{i}\left(\frac{G_1}{\tilde b_1}-\frac{(G_1\xi+G_2\tau)(2G_{11}\xi+G_{12}\tau)}{2(\tilde b_1)^3}\right)\\
&\times\left(\frac{(G_{1,1}\xi+G_{2,1}\tau)}{\tilde b_1}-\frac{(G_1\xi+G_2\tau)(G_{11,1}\xi^2+G_{22,1}\tau^2+G_{12,1}\xi\tau)}{2(\tilde b_1)^3}\right)\\
&+\frac{1}{i}\left(\frac{G_2}{\tilde b_1}-\frac{(G_1\xi+G_2\tau)(2G_{22}\tau+G_{12}\xi)}{2(\tilde b_1)^3}\right)\\
&\times\left(\frac{(G_{1,2}\xi+G_{2,2}\tau)}{\tilde b_1}-\frac{(G_1\xi+G_2\tau)(G_{11,2}\xi^2+G_{22,2}\tau^2+G_{12,2}\xi\tau)}{2(\tilde b_1)^3}\right)-\frac{1}{2}\left(\frac{2G_{11}}{\tilde b_1}-\frac{(2G_{11}\xi+G_{12}\tau)^2}{2(\tilde b_1)^3}\right)\\
&\times\left(\frac{-(G_{11,1}\xi^2+G_{22,1}\tau^2+G_{12,1}\xi\tau)(G_{1,1}\xi+G_{2,1}\tau)}{2(\tilde b_1)^3}+\frac{(G_{1,1,1}\xi+G_{2,1,1}\tau)}{\tilde b_1}\right)\\
&+\frac{3}{2}\frac{(G_{11,1}\xi^2+G_{22,1}\tau^2+G_{12,1}\xi\tau)^2}{2(\tilde b_1)^5}(G_1\xi+G_2\tau)\\
&-\frac{(G_{1,1}\xi+G_{2,1}\tau)(G_{11,1}\xi^2+G_{22,1}\tau^2+G_{12,1}\xi\tau)}{2(\tilde b_1)^3}-\frac{(G_1\xi+G_2\tau)(G_{11,1,1}\xi^2+G_{22,1,1}\tau^2+G_{12,1,1}\xi\tau)}{2(\tilde b_1)^3}\\
&-\frac{1}{2}\left(\frac{2G_{22}}{b_1}-\frac{(2G_{22}\tau+G_{12}\xi)^2}{2(\tilde b_1)^3}\right)\left(\frac{-(G_{11,2}\xi^2+G_{22,2}\tau^2+G_{12,2}\xi\tau)}{2(\tilde b_1)^3}(G_{1,2}\xi+G_{2,2}\tau)+\frac{(G_{1,2,2}\xi+G_{2,2,2}\tau)}{\tilde b_1}\right)\\
&+\frac{3}{2}(G_{11,2}\xi^2+G_{22,2}\tau^2+G_{12,2}\xi\tau)^2\frac{(G_1\xi+G_2\tau)}{2(\tilde b_1)^5}-\frac{(G_{1,2}\xi+G_{2,2}\tau)}{2(\tilde b_1)^3}(G_{11,2}\xi^2\\
&+G_{22,2}\tau^2+G_{12,2}\xi\tau)-\frac{(G_1\xi+G_2\tau)}{2(\tilde b_1)^3}\\
&\times(G_{11,2,2}\xi^2+G_{22,2,2}\tau^2+G_{12,2,2}\xi\tau)-\left(\frac{G_{12}}{\tilde b_1}-\frac{(2G_{11}\xi+G_{12}\tau)(2G_{22}\tau+G_{12}\xi)}{2(\tilde b_1)^3}\right)\\
&\times\left(\frac{-(G_{11,1}\xi^2+G_{22,1}\tau^2+G_{12,1}\xi\tau)(G_{1,2}\xi+G_{2,2}\tau)}{2(\tilde b_1)^3}+\frac{(G_{1,2,1}\xi+G_{2,2,1}\tau)}{\tilde b_1}\right.\\
&+\frac{3}{2}\frac{(G_{11,1}\xi^2+G_{22,1}\tau^2+G_{12,1}\xi\tau)}{2(\tilde b_1)^5}\\
&\times(G_{11,2}\xi^2+G_{22,2}\tau^2+G_{12,2}\xi\tau)\times(G_1\xi+G_2\tau)-\frac{(G_{1,1}\xi+G_{2,1}\tau)}{2(\tilde b_1)^3}(G_{11,2}\xi^2+G_{22,2}\tau^2+G_{12,2}\xi\tau)\\
&\left.\times\frac{-(G_1\xi+G_2\tau)}{2(\tilde b_1)^3}(G_{11,2,1}\xi^2+G_{22,2,1}\tau^2+G_{12,2,1}\xi\tau)\right),
\end{aligned}
\tag{B4}
$$

where we use the notation $f_{,1}\equiv\partial f/\partial x_1, f_{,2}\equiv\partial f/\partial x_2$, etc.

APPENDIX C: EXPLICIT EXPRESSIONS FOR C_{-1}, C_{-2}, C_{-3}, and C_{-4} IN TERMS OF $G_{11}, G_{22}, G_1, G_2, G_0, G_{12}$

$$
C_{-1}=\frac{1}{[G_{11}\xi^2+G_{22}\tau^2+G_{12}\xi\tau]^{1/2}-\lambda},
\tag{C1}
$$

$$
\begin{aligned}
C_{-2}=&(-)\frac{\tilde b_0}{(\tilde b_1)^2}+\frac{1}{4i(\tilde b_1)^5}(2G_{11}\xi+G_{12}\tau)(G_{11,1}\xi^2+G_{22,1}\tau^2+G_{12,1}\xi\tau)\\
&+\frac{1}{4i(\tilde b_1)^5}(2G_{22}\tau+G_{12}\xi)(G_{11,2}\xi^2+G_{22,2}\tau^2+G_{12,2}\xi\tau),
\end{aligned}
\tag{C2}
$$

$$
\begin{aligned}
C_{-3}=&(-)\frac{1}{\tilde b_1}\left[\frac{(G_1\xi+G_2\tau)}{\tilde b_1}\left(-\frac{(G_1\xi+G_2\tau)}{(\tilde b_1)^3}\right)+\frac{1}{4i}\frac{1}{(\tilde b_1)^5}(2G_{11}\xi+G_{12}\tau)(G_{11,1}\xi^2+G_{22,1}\tau^2+G_{12,1}\xi\tau)\right.\\
&+\frac{1}{4i}\frac{1}{(\tilde b_1)^5}(2G_{22}\tau+G_{12}\xi)(G_{11,2}\xi^2+G_{22,2}\tau^2+G_{12,2}\xi\tau)+\frac{1}{\tilde b_1}\left(\frac{G_0}{\tilde b_1}-\frac{(G_1\xi+G_2\tau)^2}{(\tilde b_1)^3}\right)\\
&+\frac{i}{2}\left(\frac{2(G_{11}\xi+G_{12}\tau)(G_{1,1}\xi+G_{2,1}\tau)}{(\tilde b_1)^2}\right)+\frac{i}{2(\tilde b_1)^2}(2G_{22}\tau+G_{12}\xi)(G_{1,2}\xi+G_{2,2}\tau)\\
&-\frac{i}{4(\tilde b_1)^4}(G_1\xi\ G_2\tau)(2G_{22}\xi+G_{12}\tau)(G_{11,1}\xi_2+G_{22,1}\tau^2+G_{12,1}\xi\tau)-\frac{i}{4(\tilde b_1)^4}
\end{aligned}
$$

$$
\begin{aligned}
&\times (G_1\xi + G_2\tau)(2G_{22}\tau + G_{12}\xi)(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau) + \left(-\frac{1}{2i}\right)\left(\frac{G_1}{\tilde{b}_1} - \frac{(G_1\xi + G_2\tau)(2G_{11}\xi + G_{12}\tau)}{2(\tilde{b}_1)^3}\right) \\
&\times \left(\frac{G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau}{(\tilde{b}_1)^3}\right) + \left(-\frac{1}{2i}\right)\left(\frac{G_2}{\tilde{b}_1} - \frac{(G_1\xi + G_2\tau)(2G_{22}\tau + G_{12}\tau)}{2(\tilde{b}_1)^3}\right) \\
&\times \left(\frac{G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau}{(\tilde{b}_1)^3}\right) + \left(-\frac{1}{i}\right)\left(\frac{2G_{11}\xi + G_{12}\tau}{2(\tilde{b}_1)^3}\right)\left(\frac{(G_{1,1}\xi + G_{2,1}\tau)}{\tilde{b}_1} - \frac{(G_1\xi + G_2\tau)}{2(\tilde{b}_1)^3}\right. \\
&\left.\times (G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau)\right) + \frac{1}{i}\,\frac{(2G_{11}\xi + G_{12}\tau)}{2(\tilde{b}_1)^6}(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau)(G_1\xi + G_2\tau) \\
&+ \frac{5}{16}\cdot\frac{(2G_{11}\xi + G_{12}\tau)}{(\tilde{b}_1)^8}(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau)((2G_{11}\xi + G_{12}\tau)(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau) \\
&+ (2G_{22}\tau + G_{12}\xi)(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau)) - \frac{1}{8}\cdot\frac{(2G_{11}\xi + G_{12}\tau)}{(\tilde{b}_1)^6}((2G_{11,1}\xi + G_{12,1}\tau)(G_{11,1}\xi^2 + G_{22,1}\tau^2 \\
&+ G_{12,1}\xi\tau) + (2G_{11}\xi + G_{12}\tau)(G_{11,1,1}\xi^2 + G_{22,1,1}\tau^2 + G_{12,1,1}\xi\tau) + (2G_{22,2}\tau + G_{12,1}\xi)(G_{11,2}\xi^2 + G_{22,1}\tau^2 \\
&+ G_{12,2}\xi\tau) + (2G_{22}\tau + G_{12}\xi)(G_{11,2,1}\xi^2 + G_{22,2,1}\tau^2 + G_{12,2,1}\xi\tau)) - \frac{1}{2i}\,\frac{(2G_{22}\tau + G_{12}\xi)}{(\tilde{b}_1)^3}\left(\frac{(G_{1,2}\xi + G_{2,2}\tau)}{\tilde{b}_1}\right. \\
&\left.- \frac{(G_1\xi + G_2\tau)}{2(\tilde{b}_1)^3}(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau)\right) + \frac{1}{2i}\,\frac{(2G_{22}\tau + G_{12}\xi)(G_1\xi + G_2\tau)}{(\tilde{b}_1)^6}(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau) \\
&+ \frac{5}{16}\,\frac{(2G_{22}\tau + G_{12}\xi}{(\tilde{b}_1)^8}(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau)((2G_{11}\xi + G_{12}\tau)(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau) \\
&+ (2G_{22}\tau + G_{12}\xi)(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau)) - \frac{1}{8}\cdot\frac{(2G_{22}\tau + G_{12}\xi)}{(\tilde{b}_1)^6}((2G_{11,2}\xi + G_{12,2}\tau)(G_{11,1}\xi^2 \\
&+ G_{22,1}\tau^2 + G_{12,1}\xi\tau) + (2G_{11}\xi + G_{12}\tau)(G_{11,1,2}\xi^2 \\
&+ G_{22,1,2}\tau^2 + G_{12,1,2}\xi\tau) + (2G_{22,2}\tau + G_{12,2}\xi)(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau) + (2G_{22}\tau + G_{12}\xi)(G_{11,2,2}\xi^2 \\
&+ G_{22,2,2}\tau^2 + G_{12,2,2}\xi\tau)) + \left(-\frac{1}{2}\right)\left(\frac{2G_{11}}{\tilde{b}_1} - \frac{(2G_{11}\xi + G_{12}\tau)^2}{2(\tilde{b}_1)^3}\right)\left(-\frac{1}{2(\tilde{b}_1)^3}(G_{11,1,1}\xi^2 + G_{22,1,1}\tau^2 + G_{12,1,1}\xi\tau)\right. \\
&\left.+ \frac{3}{4}\,\frac{1}{(\tilde{b}_1)^5}(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau)^2\right) + \left(-\frac{1}{2}\right)\left(\frac{2G_{22}}{\tilde{b}_1} - \frac{(2G_{22}\tau + G_{12}\xi)^2}{2(\tilde{b}_1)^3}\right) \\
&\times \left(-\frac{1}{2(\tilde{b}_1)^3}(G_{11,2,2}\xi^2 + G_{22,2,2}\tau^2 + G_{12,2,2}\xi\tau)\right) + \frac{3}{4}\cdot\frac{1}{(\tilde{b}_1)^5}(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau)^2 \\
&- \frac{G_{12}}{\tilde{b}_1} - \frac{(2G_{11}\xi + G_{12}\tau)(2G_{22}\tau + G_{12}\xi)}{2(\tilde{b}_1)^3}\left(-\frac{1}{2(\tilde{b}_1)^3}(G_{11,1,2}\xi^2 + G_{22,1,2}\tau^2 + G_{12,1,2}\xi\tau)\right. \\
&\left.+ \frac{3}{4}\cdot\frac{1}{(\tilde{b}_1)^5}(G_{11,1}\xi^2 + G_{22,1}\tau^2 + G_{12,1}\xi\tau)(G_{11,2}\xi^2 + G_{22,2}\tau^2 + G_{12,2}\xi\tau)\right).
\end{aligned}
\tag{C3}
$$

Finally,

$$
\begin{aligned}
C_{-4} = (-)\frac{1}{\tilde{b}_1}\Bigg\{&\tilde{b}_0 C_{-3} + \tilde{b}_{-1}C_{-2} + \tilde{b}_{-2}C_{-1} + \frac{1}{i}\left(\frac{\partial\tilde{b}_1}{\partial\xi}\right)\left(\frac{\partial C_{-3}}{\partial x_1}\right) + \frac{1}{i}\,\frac{\partial\tilde{b}_1}{\partial\tau}\,\frac{\partial C_{-3}}{\partial x_2} + \frac{1}{i}\left(\frac{\partial\tilde{b}_0}{\partial\xi}\right)\left(\frac{\partial C_{-2}}{\partial x_1}\right) \\
&+ \frac{1}{i}\left(\frac{\partial\tilde{b}_0}{\partial\tau}\right)\frac{\partial C_{-2}}{\partial x_2} + \frac{1}{i}\left(\frac{\partial b_{-1}}{\partial\xi}\right)\left(\frac{\partial C_{-1}}{\partial x_1}\right) + \frac{1}{i}\left(\frac{\partial b_{-1}}{\partial\tau}\right)\left(\frac{\partial C_{-1}}{\partial x_2}\right) - \frac{1}{2}\left(\frac{\partial^2\tilde{b}_1}{\partial\xi^2}\right)\left(\frac{\partial^2 C_{-2}}{\partial x_1^2}\right) \\
&- \frac{1}{2}\left(\frac{\partial^2\tilde{b}_1}{\partial\tau^2}\right)\left(\frac{\partial^2 C_{-2}}{\partial x_2^2}\right) - \left(\frac{\partial}{\partial\xi}\,\frac{\partial\tilde{b}_1}{\partial\tau}\right)\left(\frac{\partial}{\partial x_1}\,\frac{\partial C_{-2}}{\partial x_2}\right) - \frac{1}{2}\left(\frac{\partial^2\tilde{b}_0}{\partial\xi^2}\right)\left(\frac{\partial^2 C_{-1}}{\partial x_1^2}\right) - \frac{1}{2}\left(\frac{\partial^2\tilde{b}_0}{\partial\tau^2}\right)\left(\frac{\partial^2 C_{-1}}{\partial x_2^2}\right) \\
&- \left(\frac{\partial}{\partial\xi}\cdot\frac{\partial\tilde{b}_0}{\partial\tau}\right)\left(\frac{\partial}{\partial x_1}\cdot\frac{\partial C_{-1}}{\partial x_2}\right) + i\,\frac{1}{6}\left(\frac{\partial^3\tilde{b}_1}{\partial\xi^3}\right)\left(\frac{\partial^3 C_{-1}}{\partial x_1^3}\right) + \frac{i}{6}\left(\frac{\partial^3\tilde{b}_1}{\partial\tau^3}\right)\left(\frac{\partial^3 C_{-1}}{\partial x_2^3}\right) + \frac{i}{2} \\
&\times\left(\frac{\partial^2}{\partial\xi^2}\cdot\frac{\partial\tilde{b}_1}{\partial\tau}\right)\left(\frac{\partial^2}{\partial x_1^2}\cdot\frac{\partial C_{-1}}{\partial x_1}\right) + \frac{i}{2}\left(\frac{\partial}{\partial\xi}\cdot\frac{\partial^2\tilde{b}_1}{\partial\tau^2}\right)\left(\frac{\partial}{\partial x_1}\cdot\frac{\partial^2 C_{-1}}{\partial x_2^2}\right)\Bigg\}.
\end{aligned}
\tag{C4}
$$

[1] H. B. Casimir, Proc. Kon. Nederl. Akad. Wetensch B **51**, 793 (1948); T. H. Boyer, Phys. Rev. **180**, 19 (1969); Ann. Phys. (N.Y.) **56**, 474 (1970); B. Davies, J. Math. Phys. **13**, 1324 (1972).

[2] For example, J. S Dowker and R. Banach, J. Phys. A: Math. Gen. **2**, 2255 (1978); M. Fierz, Helv. Phys. Acta **33**, 855 (1960); W. Lukosz, Physica, **56**, 101 (1971).

[3] For example, C. Bender and P. Hays, Phys. Rev. D **14**, 2622 (1976); J. Schwinger, Lett. Math. Phys. **1**, 43 (1975); K. A. Milton, L. L. De Raad, Jr., and J. Schwinger, Ann. Phys. (N.Y.) **115**, 388 (1978); S. Sen, Phys. Rev. D **24**, 869 (1981).

[4] R. Balian and B. Duplantier, Ann. Phys. (N.Y.) **112**, 165 (1978); **104**, 300 (1977).

[5] J. J. Duistermaat and V. W. Guilleman, Invent. Math. **29**, 39 (1975).

[6] S. Sen, "Remarks on the ζ function regularization method," preprint, Trinity College, University of Dublin, 1982.

[7] P. B. Gilkey, *The Index Theorem and the Heat Equation* (Publish or Perish, Kensington, CA, 1974).

[8] M. Wodzicki, Invent. Math. **66**, 115 (1982).

[9] G. I. Èskin, *Boundary Value Problems for Elliptic Pseudodifferential Equations* (American Mathematical Society, Providence, RI, 1981).

Hamiltonian formulation of the KdV equation

Y. Nutku
TÜBİTAK Research Institute for Fundamental Sciences, P.O. Box 74, Gebze, Kocaeli, Turkey and Department of Mathematics, Bosphorus University, Bebek, Istanbul, Turkey

(Received 10 June 1983; accepted for publication 2 September 1983)

We consider the canonical formulation of Whitham's variational principle for the KdV equation. This Lagrangian is degenerate and we have found it necessary to use Dirac's theory of constrained systems in constructing the Hamiltonian. Earlier discussions of the Hamiltonian structure of the KdV equation were based on various different decompositions of the field which is avoided by this new approach.

PACS numbers: 11.10.Ef, 11.10.Lm

I. INTRODUCTION

In a recent paper, "A particle representation for KdV solitons," Bowtell and Stuart[1] discuss the Hamiltonian formulation of the KdV equation. Starting with the variational formulation of the KdV equation in terms of the velocity potential[2] they have obtained what one ordinarily expects to be the Hamiltonian and observed that it has various undesirable properties. This Hamiltonian is in fact not the correct one, because the Lagrangian in the variational principle for the KdV equation is degenerate. It is necessary to use Dirac's theory of constraints[3] in order to cast such degenerate systems into canonical form and we shall do so in this paper.

The construction of a Hamiltonian for the KdV equation was first considered by Gardner[4] using a decomposition of the field into normal modes and this method has since been extensively studied.[5] The Hamiltonian formalism we shall present in this paper is based on a completely different approach which is much more general. We shall not require reference to any type of decomposition but instead proceed directly with Dirac's method to find the KdV Hamiltonian. The new Hamiltonian is formulated in terms of two potentials and their conjugate momenta and it is conserved by virtue of the third[6] among the infinite set of conservation laws[7] satisfied by the KdV equation.

II. HAMILTONIAN

The KdV equation can be obtained from a variational principle

$$\delta I = 0, \quad I = \int \mathscr{L}\, dx\, dt, \tag{1}$$

where[2]

$$\mathscr{L} = \tfrac{1}{2}\phi_t \phi_x + \phi_x^3 + \phi_x \psi_x + \tfrac{1}{2}\psi^2 \tag{2}$$

is the Lagrangian density. Here ϕ is the velocity potential

$$u = \phi_x \tag{3}$$

and ψ is another potential which has been introduced to avoid derivatives higher than the first order in the Lagrangian. Considering independent variations with respect to ϕ and ψ we obtain

$$\phi_{xt} + 6\phi_x\phi_{xx} + \psi_{xx} = 0, \tag{4}$$

$$\psi - \phi_{xx} = 0, \tag{5}$$

which in view of Eq. (3) leads to the KdV equation

$$u_t + 6uu_x + u_{3x} = 0. \tag{6}$$

The Lagrangian in Eq. (2) is degenerate, that is, the momenta

$$\pi_\psi = 0, \tag{7a}$$

$$\pi_\phi = \tfrac{1}{2}\phi_x \tag{7b}$$

cannot be inverted for the velocities ψ_t and ϕ_t. Therefore, following Dirac we define the constraints

$$c_1 = \pi_\psi, \tag{8a}$$

$$c_2 = \pi_\phi - \tfrac{1}{2}\phi_x, \tag{8b}$$

and using the canonical Poisson brackets

$$[\psi(x), \pi_\psi(x')] = \delta(x - x'), \tag{9a}$$

$$[\phi(x), \pi_\phi(x')] = \delta(x - x'), \tag{9b}$$

we find that Poisson brackets of the constraints are given by

$$[c_1(x), c_1(x')] = 0, \tag{10a}$$

$$[c_1(x), c_2(x')] = 0, \tag{10b}$$

$$[c_2(x), c_2(x')] = -\delta_x(x - x'), \tag{10c}$$

where δ stands for the Dirac δ-function. From Eqs. (10) we conclude that the constraints are second class.

The total Hamiltonian is given by

$$H = \int \mathscr{H}\, dx, \tag{11a}$$

$$\mathscr{H} = \mathscr{H}_0 + \mathscr{H}_1, \tag{11b}$$

where $\mathscr{H}_0$ is the free part (which is the only piece considered in Ref. 1),

$$\begin{aligned}\mathscr{H}_0 &= \pi_\psi \psi_t + \pi_\phi \phi_t - \mathscr{L} \\ &= -\phi_x^3 - \phi_x \psi_x - \tfrac{1}{2}\psi^2,\end{aligned} \tag{12}$$

and $\mathscr{H}_1$ is a linear combination of the constraints

$$\mathscr{H}_1 = \lambda c_1 + \sigma c_2, \tag{13}$$

where λ and σ are multipliers. The Poisson bracket of the constraints with H must vanish and this requirement will determine λ and σ provided there are not further constraints in the problem. We find however, that

$$[c_1, H] = \psi - \phi_{xx}, \tag{14}$$

which cannot be set equal to zero by any choice of the multipliers and therefore

$$\chi = \psi - \phi_{xx} \tag{15}$$

must be introduced as a secondary constraint. This constraint will be included in the total Hamiltonian so that in

 0022-2488/84/062007-02$02.50

place of Eq. (13) we now have

$$\mathscr{H}_1 = \lambda c_1 + \sigma c_2 + \mu\chi, \tag{16}$$

where μ is another multiplier. We have now obtained all of the constraints in the problem. In order to see this we calculate the Poisson brackets

$$[c_1,H] = \chi - \mu, \tag{17a}$$

$$[c_2,H] = (\mu_x - \sigma - 3\phi_x^2 - \psi_x)_x, \tag{17b}$$

$$[\chi,H] = \lambda - \sigma_{xx}, \tag{17c}$$

and observe that the right-hand sides of Eqs. (17) can be made to vanish by choosing

$$\mu = \psi - \phi_{xx}, \tag{18a}$$

$$\sigma = -\phi_{3x} - 3\phi_x^2, \tag{18b}$$

$$\lambda = -\phi_{5x} - 6\phi_x\phi_{3x} - 6\phi_{xx}^2. \tag{18c}$$

When we put these expressions back into Eqs. (12) and (16) we find the total Hamiltonian density

$$\mathscr{H} = \tfrac{1}{2}\phi_x^3 + \tfrac{1}{2}\psi^2 + \phi_x\psi_x + \tfrac{1}{2}\phi_{xx}^2 - (\phi_{3x} + 3\phi_x^2)\pi_\phi - (\phi_{5x} + 6\phi_{xx}^2 + 6\phi_x\phi_{3x})\pi_\psi, \tag{19}$$

where we have left out a divergence.

We can verify that the Poisson bracket of the canonical variables ϕ, π_ϕ, ψ, π_ψ with the total Hamiltonian results in

$$\phi_t + 3\phi_x^2 + \phi_{3x} = 0, \tag{20a}$$

$$\phi_{xt} + 6\phi_x\phi_{xx} + \phi_{4x} = 0, \tag{20b}$$

$$\psi_t + 6\phi_{xx}^2 + 6\phi_x\phi_{3x} + \phi_{5x} = 0, \tag{20c}$$

$$\psi - \phi_{xx} = 0, \tag{20d}$$

respectively. Equation (20b) is the KdV equation, and with the definition (20d), Eq. (20c) also reduces to the KdV equation. Furthermore Eq. (20a) is also equivalent to the KdV equation up to an arbitrary function of time which can be added to its right-hand side. This arbitrary function of time can be set equal to zero without loss of generality because in solving for σ from Eq. (17b) we have ignored such an arbitrary function. Therefore the content of Eqs. (20) is the KdV equation.

The total Hamiltonian is a conserved quantity and it will be of interest to express it in terms of the original field u. From Eqs. (3), (7a), (7b), (20c), and (19) we find that up to a surface term

$$H = \int (\tfrac{1}{2}u_x^2 - u^3)\,dx, \tag{21}$$

which is Whitham's conserved quantity. It arises from the third, among the infinitely many, conservation laws satisfied by the KdV equation.

III. CONCLUSION

We have constructed the Hamiltonian for the KdV equation by applying Dirac's theory of constrained systems to Whitham's degenerate Lagrangian. We have found that in this system there are two primary constraints which are second class and there is also a secondary constraint. The total Hamiltonian of Dirac is given by a local expression in terms of the canonical variables. This approach to the Hamiltonian structure of the KdV equation is more general that the one encountered in earlier work which starts by decomposing the field into various modes[4,5] and completes earlier results[1] based on Whitham's variational principle.

ACKNOWLEDGMENT

I thank Professor E. İnönü and Y. Akyıldız for their kind interest in this work.

This work was in part supported by the Turkish Scientific Research Council TÜBİTAK.

[1]G. Bowtell, and A. E. G. Stuart, J. Math. Phys. **24**, 969 (1983).
[2]G. B. Whitham, Proc. R. Soc. London Ser. A **299**, 6 (1967).
[3]P. A. M. Dirac, "Lectures on Quantum Mechanics," Belfer Graduate School Monograph Series No. 2, Yeshiva University, New York, 1964.
[4]C. S. Gardner, J. Math. Phys. **12**, 1548 (1971).
[5]V. E. Zakharov and L. D. Faddeev, Functional Anal. Appl. **5**, 280 (1972); J. W. Miles, Appl. Sci. Res. **37**, 103 (1981).
[6]G. B. Whitham, Proc. R. Soc. London Ser. A **283**, 238 (1965).
[7]R. M. Miura, C. S. Gardner, and M. D. Kruskal, J. Math. Phys. **9**, 1204 (1968).

Geometrical properties and invariances of the generalized sigma model

M. Arik and F. Neyzi
Physics Department, Boğaziçi University, Istanbul, Turkey

(Received 7 September 1983; accepted for publication 13 January 1984)

We consider the most general conformal invariant O(3) nonlinear σ-model in two dimensions, whose Lagrangian contains only first derivatives of the field variable. This is the generalized sigma model (GSM). We point out that its integrability condition, the generalized sine–Gordon (GSG) equation, follows from the fact that the Gaussian curvature of the surface underlying the model is a constant. We show that the GSG can be formulated as an imbedding and an inverse scattering problem. Finally we examine the invariance properties of the GSG equation.

PACS numbers: 11.10.Lm

I. INTRODUCTION

In recent years the two-dimensional nonlinear sigma model has attracted a great deal of renewed interest among physicists. The standard action of this model is bilinear in the derivatives of the field variable. Furthermore, the values of these variables are constrained to the surface of a two-dimensional sphere. One type of generalization of this model has been obtained by replacing the spherical surface by a higher-dimensional homogeneous space. This approach has yielded a number of interesting results both in the classical and in the quantum domain.

More recently a different type of generalization of this model has been considered in which the values of the field functions still describe a spherical surface but the Lagrangian is generalized in such a way that the conformal invariance of the resulting action is preserved.[1] If only first-derivative terms are present, then such a Lagrangian contains an arbitrary function of one variable. Depending on the choice of this arbitrary function, the generalized sigma model (GSM) can describe an infinite family of dynamical systems. The nonlinearity of this generalized model arises not solely from the constraint that the field variable take values on a nonlinear space but also from the fact that the Lagrangian is no longer bilinear in the derivatives of the field variable. The integrability condition for the classical equations of motion of this model is a generalization of the sine–Gordon equation, which in turn, is the integrability condition of the standard nonlinear sigma model[2] in two dimensions. In this paper we exploit the methods that have been developed for the sine–Gordon equation and other totally integrable systems to further investigate the generalized sine–Gordon equation (GSG). In Sec. II we utilize geometrical techniques to show that just like the sine–Gordon equation, the GSG equation can be derived from the condition that its underlying surface has a Gaussian curvature equal to 1. We then consider the imbedding of this surface in a three-dimensional flat space. In Sec. III, we formulate the GSG equation as inverse scattering problem according to the general framework provided by Lax[3] and Ablowitz, Kaup, Newell, and Segur (LAKNS).[4] Thereby we derive the soliton connection and obtain a self-Bäcklund transformation. In Sec.IV we use the group theoretical approach to investigate the invariance properties of the GSG equation under a transformation of the variables. This invariance study once more singles out the special cases of the GSG equation exhibiting extra symmetries, which have been shown to be unified with a variable transformation.[1]

II. SURFACE THEORY AND THE GSM

In this section we give a brief description of the GSM and then proceed to investigate the nature of the surface underlying this model. To this end, we identify the first fundamental form and the Gaussian curvature of the surface. Then we consider the imbedding of the surface in a three-dimensional flat space and obtain the appropriate second fundamental form.

The GSM is described by the Lagrangian density

$$L = \mathbf{n}_u \cdot \mathbf{n}_v f(\theta) + \lambda (\mathbf{n}^2 - 1), \tag{2.1}$$

where

$$u = \tfrac{1}{2}(t + x), \quad v = \tfrac{1}{2}(t - x) \tag{2.2}$$

are the light cone variables. The field variable $\mathbf{n}$ is a three-dimensional vector and the Lagrange multiplier λ ensures that $\mathbf{n}^2$ equals 1. The quantity θ is half the angle between $\mathbf{n}_u$ and $\mathbf{n}_v$. Here subscripts denote partial differentiation. With f, an arbitrary function of θ, the action

$$S = \int L\, du\, dv \tag{2.3}$$

is conformally invariant.

The three-dimensional unit vector $\mathbf{n}$ geometrically describes the surface of a unit sphere. Since $\mathbf{n}$ is a function of u and v, provided that this function is nonsingular, these variables, or equivalently x and t, may be used as the coordinates of this two-dimensional surface. In Ref. 1 it was shown that the equations of motion and the conformal invariance of the model can be utilized to choose coordinates such that

$$\mathbf{n}_u^2 = \mathbf{n}_v^2 = [\, f(\theta) - \tfrac{1}{4} f'(\theta)\sin(4\theta)]^{-1} \equiv e^{2\alpha}, \tag{2.4}$$

where α is a function of θ which is determined by the function f in the Lagrangian. Thus the metric on the unit sphere is given by

$$\begin{aligned} ds^2 = d\mathbf{n}^2 &= (\mathbf{n}_v\, dv + \mathbf{n}_u\, du)^2 \\ &= e^{2\alpha}(\cos^2\theta\, dt^2 + \sin^2\theta\, dx^2). \end{aligned} \tag{2.5}$$

From this equation we immediately recognize the basis one-forms

$$\begin{aligned} \omega^1 &= e^{\alpha}\cos\theta\, dt, \\ \omega^2 &= e^{\alpha}\sin\theta\, dx, \end{aligned} \tag{2.6}$$

 0022-2488/84/062009-04$02.50

such that

$$ds_1^2 = (\omega^1)^2 + (\omega^2)^2. \tag{2.7}$$

The connection one-form

$$\omega^\alpha{}_\beta = -\omega^\beta{}_\alpha \tag{2.8}$$

is determined from the integrability condition

$$d\omega^\alpha + \omega^\alpha{}_\beta \wedge \omega^\beta = 0, \tag{2.9}$$

and is given by

$$\omega^1{}_2 = (\alpha_x \cot\theta - \theta_x)dt - (\alpha_t \tan\theta + \theta_t)dx. \tag{2.10}$$

The Gaussian curvature K is defined by

$$d\omega^\alpha{}_\beta = K\omega^\alpha \wedge \omega^\beta. \tag{2.11}$$

Letting the Gaussian curvature equal to 1 we get

$$[(\alpha' \cot\theta - 1)\theta_x]_x + [(\alpha' \tan\theta + 1)\theta_t]_t + e^{2\alpha}\cos\theta\sin\theta = 0. \tag{2.12}$$

This is the integrability condition for the GSM.

Having identified the intrinsic geometry underlying the GSM, we consider the imbedding of the surface S in a three-dimensional flat space M. We define the second fundamental form of S through

$$-ds_2^2 = \omega^1 \otimes \pi^1 + \omega^2 \otimes \pi^2. \tag{2.13}$$

The equation governing the imbedding problem is

$$d\omega^i{}_k + \omega^i{}_j \wedge \omega^j{}_k = 0, \tag{2.14}$$

where the indices range over three values. This is the Gauss–Codazzi equation for the imbedding problem with the identification

$$\omega^1{}_3 = \pi^1, \quad \omega^2{}_3 = \pi^2. \tag{2.15}$$

Furthermore the definition of the surface S as

$$\omega^3 = 0 \tag{2.16}$$

imposes another condition

$$\omega^1 \wedge \pi^1 + \omega^2 \wedge \pi^2 = 0. \tag{2.17}$$

We see that Eqs. (2.14)–(2.17) are all satisfied when we let

$$\pi^1 = \omega^1, \quad \pi^2 = \omega^2. \tag{2.18}$$

III. THE SOLITON CONNECTION AND THE GSM

Having established the geometrical framework underlying the GSM, we construct an SL(2,R) valued connection one-form with zero curvature. This is the soliton connection. We perform a gauge transformation in order to cast it into LAKNS form. Finally we find the Bäcklund transformation for the GSM. Our approach closely resembles that of Ref. 5.

The Gauss–Codazzi equations for embedding surfaces in a three-dimensional flat space form a realization of Cartan's equations for SL(2,R):

$$d\Theta^i + \tfrac{1}{2}C_j{}^i{}_k\Theta^j \wedge \Theta^k = 0, \tag{3.1}$$

where

$$C_1{}^0{}_2 = 1, \qquad C_0{}^1{}_1 = -C_0{}^2{}_2 = 2 \tag{3.2}$$

are the structure constants of SL(2,R), with identification

$$\begin{aligned}\Theta^0 &= (i/2)\omega^1{}_2,\\ \Theta^1 &= -\tfrac{1}{2}(\omega^2 + i\omega^1),\\ \Theta^2 &= \tfrac{1}{2}(\omega^2 - i\omega^1).\end{aligned} \tag{3.3}$$

For the GSM, these are

$$\begin{aligned}\Theta^0 &= (i/2)[(\alpha'\cot\theta - 1)\theta_x]dt - [(\alpha'\tan\theta + 1)\theta_t]dx,\\ \Theta^1 &= -\tfrac{1}{2}(e^\alpha \sin\theta\, dx + ie^\alpha\cos\theta\, dt),\\ \Theta^2 &= \tfrac{1}{2}(e^\alpha\sin\theta\, dx - ie^\alpha\cos\theta\, dt).\end{aligned} \tag{3.4}$$

Now we construct a connection one-form Γ_1

$$\Gamma_1 = \begin{pmatrix}\Theta^0 & \Theta^1\\ \Theta^2 & -\Theta^0\end{pmatrix}, \tag{3.5}$$

with a vanishing curvature

$$d\Gamma_1 + \Gamma_1 \wedge \Gamma_1 = 0. \tag{3.6}$$

This can be traced back to Gauss–Codazzi equations through Eq. (3.1).

Now we can briefly summarize the LAKNS formalism. The integrability conditions for the systems of linear partial differential equations

$$\begin{aligned}V_{1x} - i\xi V_1 &= qV_2,\\ V_{2x} + i\xi V_2 &= rV_1,\end{aligned} \tag{3.7}$$

where the eigenfunctions V_1 and V_2 evolve in time according to

$$\begin{aligned}V_{1t} &= AV_1 + BV_2,\\ V_{2t} &= CV_1 - AV_2,\end{aligned} \tag{3.8}$$

are

$$\begin{aligned}A_x &= qC - rB,\\ B_x - 2i\xi B &= q_t - 2qA,\\ C_x + 2i\xi C &= r_t + 2rA.\end{aligned} \tag{3.9}$$

Here A, B, and C are functions of x, t, and ξ, where ξ is a constant. If we construct a connection one-form with the identification

$$\begin{aligned}\Theta^0 &= -(A\,dt + i\xi\,dx),\\ \Theta^1 &= -(B\,dt + q\,dx),\\ \Theta^2 &= -(C\,dt + r\,dx),\end{aligned} \tag{3.10}$$

the condition that its curvature vanishes yields Eq. (3.9). We want to identify the set described in Eq. (3.10) with that in Eq. (3.4) to read off the LAKNS potentials for the GSM. However, as they stand, these two sets of equations are incompatible since ξ has to be a constant. To circumvent this problem we perform a gauge transformation

$$\Gamma' = S\Gamma_1 S^{-1} + S\,dS^{-1} \tag{3.11}$$

in order to cast Eq. (3.4) into LAKNS form described by Eq. (3.10). Here S has a determinant equal to unity. We find that for the GSM, S has a very simple form

$$S = \frac{1}{\sqrt{2}}\begin{pmatrix}e^{-i\xi x} & e^{-i\xi x}\\ -e^{i\xi x} & e^{i\xi x}\end{pmatrix}. \tag{3.12}$$

Equating the transformed one-form Γ' with (3.10) we get

$$\begin{aligned}A &= (i/2)e^\alpha\cos\theta,\\ C &= (i/2)e^{2i\xi x}(\alpha'\cot\theta - 1)\theta_x,\\ B &= (i/2)e^{-2i\xi x}(\alpha'\cot\theta - 1)\theta_x,\\ q &= \tfrac{1}{2}e^{-2i\xi x}[e^\alpha\sin\theta - i(\alpha'\tan\theta + 1)\theta_t],\\ r &= -\tfrac{1}{2}e^{2i\xi x}[e^\alpha\sin\theta + i(\alpha'\tan\theta + 1)\theta_t].\end{aligned} \tag{3.13}$$

The eigenvalue ξ usually has a real and an imaginary part. If we restrict its value to be real for simplicity then

$$r = -q^*. \tag{3.14}$$

Furthermore,

$$B = -C^*. \tag{3.15}$$

In the literature, it has been shown that the Gelfand–Levitan–Marchenko integral equation associated with the inverse scattering problem is uniquely solvable if $r = -q^*$ (Ref. 6). For this specific case, there is also a general method for deriving the Bäcklund transformation from the equations for the inverse problem.[7] This is as follows. To conform with the notation in Ref. 7 we let ξ go to $-\xi$ in Eq. (3.7). Then, defining a quantity $U = V_1/V_2$ we get a system of equations

$$\begin{aligned} U_x &= -2i\xi U - rU^2 + q, \\ U_t &= 2AU - CU^2 + B. \end{aligned} \tag{3.16}$$

For the GSM we have

$$U_x = -2i\xi U + q^* U^2 + q, \tag{3.17a}$$

$$U_t = 2AU + CU^2 - C^*, \tag{3.17b}$$

where

$$\begin{aligned} A &= \frac{1}{2}\left(\frac{q_t e^{2i\xi x} + \text{c.c.}}{q e^{2i\xi x} - \text{c.c.}}\right), \\ c &= \frac{i}{2}\left(\frac{1}{q + q^* e^{-4i\xi x}}\right)\left(\frac{iq_t e^{2i\xi x} - \text{c.c.}}{-q e^{2i\xi x} + \text{c.c.}}\right)_x. \end{aligned} \tag{3.18}$$

Using Eq. (3.17a) and its complex conjugate we derive an expression for q in terms of U, U^*, U_x, and U_x^*. Substituting this expression in Eq. (3.17b), we can eliminate q and thereby get a nonlinear partial differential equation for U and U^*. This equation is invariant under the transformation

$$(U,\xi) \mapsto (-U,\xi). \tag{3.19}$$

The existence of this gaugelike invariance makes it possible to find a self-Bäcklund transformation since we know we have a second solution q' such that

$$U_x = -2i\xi U - q'^* U^2 - q', \tag{3.20a}$$

$$U_t = 2AU - CU^2 + C^*. \tag{3.20b}$$

Here A and C are functions of q', q'_t, q'_x, and q'_{xt} and their complex conjugates. Subtracting Eq. (3.20a) from Eq. (3.17a) we obtain the following expression for U:

$$U = \pm i[\,|q + q'|/(q + q')^*]H(x - x_0 + 4\xi t), \tag{3.21}$$

where H is the Heaviside step function. In order to get the spatial part of the Bäcklund transformation for the GSG equation, we add Eqs. (3.20a) and (3.17a) and then substitute for U from Eq. (3.21). Rearranging terms we get

$$\begin{aligned} (q - q')_x = &-2i\xi(q + q') + i(q - q')|q \pm q'| \\ &\times H(x - x_0 + 4\xi t). \end{aligned} \tag{3.22}$$

Similarly we add Eqs. (3.20b) and (3.17b) and substitute for U to get the temporal part of the Bäcklund transformation.

IV. INVARIANCE AND INFINITESIMAL PROPERTIES OF THE GSG EQUATION

In this section we shall investigate the invariance properties of the GSG equation under one parameter Lie group of transformations.[8] We shall find the invariant variable in terms of which we can reduce the partial differential equation with two independent variables and one dependent variable to an ordinary differential equation. Finally we find the explicit solution of this ODE for a special case.

Let us rewrite the GSG equation [Eq. (2.12)];

$$\begin{aligned} &2\gamma(u)\sin u + u_{tt}F_1(u) - u_{xx}F_2(u) + u_t^2 F_u \\ &\quad = H(u, u_{xx}, u_{tt}, \ldots) = 0. \end{aligned} \tag{4.1}$$

Here

$$F_1 = 1 + (\gamma_u/\gamma)\tan(u/2), \tag{4.2a}$$

$$F_2 = 1 - (\gamma_u/\gamma)\cot(u/2) \tag{4.2b}$$

$$\begin{aligned} \gamma &= e^{2\alpha}, \\ \theta &= u/2, \end{aligned} \tag{4.3}$$

and subscripts denote partial differentiation.

Let $u = \phi(x,t)$ be a solution of H. If H is invariant under the one-parameter (ϵ) group of transformations obtained from the infinitesimal transformation,

$$\begin{aligned} x' &= x + \epsilon\beta(x,t,u), \\ t' &= t + \epsilon\tau(x,t,u), \\ u' &= u + \epsilon\eta(x,t,u), \end{aligned} \tag{4.4}$$

through exponentiation, then

$$XH|_{u=\phi(x,t)} = 0, \tag{4.5}$$

where X is the operator

$$\begin{aligned} X = \beta\frac{\partial}{\partial x} + \tau\frac{\partial}{\partial t} + \eta\frac{\partial}{\partial u} + [\eta_x]\frac{\partial}{\partial u_x} \\ + [\eta_{xx}]\frac{\partial}{\partial u_{xx}} \cdots. \end{aligned} \tag{4.6}$$

Here $[\eta_x]$ and $[\eta_{xx}]$ are the infinitesimals for u_x and u_{xx}, respectively, whose explicit forms in terms of η, β, and τ can be obtained from Eq. (4.4). Furthermore if H is invariant under the transformation defined by Eq. (4.4) then the following equation must hold:

$$\frac{dx}{\beta} = \frac{dt}{\tau} = \frac{d\phi}{\eta}. \tag{4.7}$$

Using Eq. (4.5), we find the infinitesimals β, η, and τ. Then substituting these values in Eq. (4.7), we get a solution with two arbitrary constants. One of them, β, will be the similarity variable. The other one, $f(\beta)$ will be the independent variable.

For Eq. (4.1), when we impose the condition (4.5) and collect together the like-derivative terms in ϕ we get

$$\begin{aligned} \eta &= 0, \\ \tau &= \tau(x) + b, \\ \beta &= \beta(t) + b, \end{aligned} \tag{4.8}$$

restricted by the following equations:

$$\begin{aligned} \tau_x F_2 &= \beta_t F_1, \\ \tau_x F_2' &= \beta_t F_1'. \end{aligned} \tag{4.9}$$

Equation (4.9) tells us immediately that either

$$F_1 = a^2 F_2, \tag{4.10}$$

where a is just a constant, or

$$\tau_x = \xi_t = 0. \tag{4.11}$$

Equation (4.10) is the same as Eq. (3.13) in Ref. 1. If this relation holds, then the integrability condition of the GSM reduces to the sine–Gordon, the Euclidean sinh–Gordon, the time-independent, or the space-independent equations for different values of a^2. The invariance properties of the sine–Gordon equation has been analyzed in Ref. 8; the other three cases are related to the sine–Gordon equation by a variable transformation.[1]

If we do not have the special case described by Eq. (4.9), then using Eqs. (4.11), (4.8), and (4.7) we find the invariant variables β and $f(\beta)$:

$$\begin{aligned} &x - (a/b)t = \beta, \\ &u = f(\beta). \end{aligned} \tag{4.12}$$

This reduces the GSG equation to a particularly simple form

$$\{[F_2 - (a^2/b^2)F_1]f'\}' = 2\gamma \sin f, \tag{4.13}$$

where primes denote differentiation with respect to β.

ACKNOWLEDGMENTS

We would like to thank our colleagues in the Bogazici University Physics Department and Gebze Basic Sciences Research Institute for useful discussions and especially to Yavuz Nutku for his guidance and enthusiasm.

This research was supported in part by the Turkish Scientific and Technical Research Council.

[1]M. Arik and F. Neyzi, J. Math. Phys. **24**, 2648 (1983).
[2]K. Pohlmeyer, Commun. Math. Phys. **46**, 207 (1976).
[3]P. D. Lax, Commun. Pure Appl. Math. **21**, 467 (1968).
[4]M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Phys. Rev. Lett. **31**, 125 (1973).
[5]M. Gurses and Y. Nutku, J. Math. Phys. **22**, 1393 (1981).
[6]M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974).
[7]H. Chen, Phys. Rev. Lett. **33**, 925 (1974).
[8]M. Lakshmanan and P. Kaliappan, J. Math. Phys. **24**, 795 (1983).

Indefinite metric and other-than-Feynman propagators as tools in a relativistic two-body problem

Marian Günther
Department of Physics, University of Cincinnati, Cincinnati, Ohio 45221

(Received 14 November 1983; accepted for publication 16 December 1983)

By embedding the usual physical Fock space P of a simple field theory in a larger non-Hilbert representation space R, a formalism is obtained permitting the use of $\bar{\Delta}$ propagators instead of the Feynman propagators Δ_F to represent the internal particle lines in Feynman graphs. The new formalism is subsequently applied to a Bethe–Salpeter equation in the g^2 approximation of its kernel and an exact solution of the latter is obtained exhibiting much simpler analytical features compared to its standard counterpart—the so-called Wick equation.

PACS numbers: 11.10.St, 11.90. + t, 02.90. + p, 03.70. + k

I. INTRODUCTION

The modification of the standard field-theoretical algorithm proposed in this paper represents—in the nutshell—an attempt at the physical and mathematical validation of the use of other-than-Feynman types of propagators to represent *internal* particle lines in the (appropriately redefined) Feynman graphs. This wider (than just Feynman) variety of propagators admissible in this sense has its origin—by tracing back the meaning of such propagators to their generic definition as "contraction symbols" in the usage of the Wick's Ordering Theorem[1]—in an at first formal generalization of the definition of normal products of the field operators. This generalization must entail however the introduction of an auxiliary field or fields and thus an extension of the underlying representation space of the field algebra to a representation space R with indefinite metric (non-Hilbert). The physical subspace P of R, i.e., a subspace invariant under all operators representing physical quantities, remains nevertheless Hilbert and thus does *not* change the physical content of the theory, in which respect our new algorithm follows in fact exactly the footsteps of the well-known Bleuler–Gupta method in QED where the auxiliary ghost field representing the fourth component of the vector potential A_0 necessitates the introduction of the indefinite metric in a larger space ω and yet the physical space spanned by vectors $|\Psi\rangle$ corresponding to the absence of the quanta of the free field

$$\pi = \partial_\mu A^\mu$$

remains Hilbert.[2] The general field-theoretical aspects of the proposed formalism go therefore actually beyond the confines of *any* particular perturbative approach (i.e., whether one expands in terms of the standard or modified Feynman graphs) and it is in fact the much greater facility with which a certain type of the Bethe–Salpeter equation can be solved when the exchange particle propagator is represented by half the difference between the retarded and advanced propagators (i.e., the so-called $\bar{\Delta}$ propagator)[3] as compared to the well-known method[4] of solving the standard version of this equation (with all Feynman propagators) in which—in the author's opinion—also purely computational advantages of the new method can be seen. In order to make the paper self-contained, the rudiments of this new solution are given in Sec. VI, including the relativistic generalization of the Balmer–Ritz formula for two finite-mass particle bound states, which—to all appearances—cannot at least be as easily and unambiguously derived from the standard Bethe–Salpeter equation.

II. GENERALIZED CONTRACTION SYMBOLS (INTERNAL LINE PROPAGATORS)

We formulate our procedure for a scalar self-adjoint Klein–Gordon field $\varphi_a(x)$, where the subscript a is for a later distinction. The extension to other boson fields and to fermion fields is straightforward.

First, to establish notation, we briefly review the conventional formalism. In the interaction picture this field obeys the free field equation

$$(\Box + m_a^2)\varphi_a(x) = 0 \tag{1}$$

and the commutation relation

$$[\varphi_a(x),\varphi_a(x')] = i\Delta\,(x-x';m), \tag{2}$$

where $\Delta\,(x;m)$ is the well-known Pauli–Jordan function. The field $\varphi_a(x)$ may furthermore be represented, in a well-known way, as

$$\varphi_a(x) = \varphi_a^{(+)}(x) + \varphi_a^{(-)}(x), \tag{3}$$

where the positive and negative energy parts are

$$\varphi_a^{(+)}(x) = \sum_\alpha f_\alpha(x) a_\alpha \tag{4}$$

and

$$\varphi_a^{(-)}(x) = \sum_\alpha \overline{f_\alpha(x)} a_\alpha^*, \tag{5}$$

where $\{f_\alpha(x)\}$ is a set of positive energy solutions of the free Klein–Gordon equation, together with its complex-conjugate set $\{\overline{f_\alpha(x)}\}$ of negative energy solutions. These sets are assumed to be orthonormal in the covariant sense, namely,

$$-i\int d^3\mathbf{x}\left\{f_\alpha(x)\frac{\partial \overline{f_\alpha(x)}}{\partial x_0} - \frac{\partial f_\alpha(x)}{\partial x_0} f_\beta(x)\right\} = \delta_{\alpha\beta} \tag{6}$$

and to be complete in the sense that

$$\sum_\alpha f_\alpha(x)\ \overline{f_\alpha(x')} = i\Delta^{(+)}(x-x';m), \tag{7}$$

where $\Delta^{(+)}(x;m)$ stands for the positive energy part of the previously introduced Pauli–Jordan function $\Delta(x;m)$. Consequently, the creation and annihilation operators a_α^* and a_α satisfy[5]

$$[a_\alpha,a_\beta] = [a_\alpha^*,a_\beta^*] = 0, \tag{8}$$

$$[a_\alpha,a_\beta^*] = \delta_{\alpha\beta}. \tag{9}$$

In this context, the Wick ordering theorem concerns the expansion of the time-ordered products, $T(\varphi_a(x_1),...,\varphi_a(x_n))$ in terms of the normal products

$$N(\varphi_a(x_1),...,\varphi_a(x_r)), \quad r\leqslant n; \tag{10}$$

the *standard* definition of the latter being that the ordinary product of r φ_a operators

$$\varphi_a(x_1)\cdots\varphi_a(x_r), \tag{11}$$

should first be formally written as a sum of all the products of r factors of the type $\varphi_a^{(+)}$ or $\varphi_a^{(-)}$ resulting from substituting the expression (3) into (11) and then defining (10) by changing the order of those latter products by placing all $\varphi_a^{(-)}$ to the left of all $\varphi_a^{(+)}$.

For our purposes the pertinent part of Wick's theorem is that the coefficients of the normal products in this expansion are *solely* products of the c-number contraction symbols defined by

$$C(x,x')\equiv T(\varphi_a(x),\varphi_a(x')) - N(\varphi_a(x),\varphi_a(x')). \tag{12}$$

Equation (12) is the generic definition of the propagator representing the internal particle lines in Feynman graphs. The above definition of the normal product is what conveys the general impression that only Feynman propagators can be used to represent internal lines in a realistic theory.

We now endeavor to remove this last constraint by generalizing the normal product through the simple expedient of replacing $\varphi_a^{(+)}$ and $\varphi_a^{(-)}$ in the standard definition by φ_1 and φ_2, respectively, in a decomposition of φ_a,

$$\varphi_a(x) = \varphi_1(x) + \varphi_2(x), \tag{13}$$

different in general from (3); the (generalized) normal product N' being defined as the corresponding sum of products with all the φ_2 standing to the left of all φ_1. Caution must be exercised here because it is required that

$$[\varphi_1(x),\varphi_1(x')] = [\varphi_2(x),\varphi_2(x')] = 0 \tag{14}$$

for all x and x', for the normal product to be well defined, independently of the sequence of φ_1's among themselves or φ_2's among themselves.

No other linear combinations of $\varphi_a^{(+)}$ and $\varphi_a^{(-)}$ except the trivial $\varphi_1 = \varphi_a^{(+)}$, $\varphi_2 = \varphi_a^{(-)}$ (or vice versa) satisfies both (13) and (14) and we are thereby led to consider auxiliary fields. We introduce *two* independent "primary" self-adjoint scalar KG fields, $\varphi_b(x)$ and $\varphi_c(x)$, which satisfy the same field equations and commutation relations. Thus

$$(\Box + m^2)\varphi_b(x) = (\Box + m^2)\varphi_c(x) = 0 \tag{15}$$

and

$$[\varphi_b(x),\varphi_b(x')] = [\varphi_c(x),\varphi_c(x')] = i\Delta(x-x';m) \tag{16}$$

and

$$[\varphi_b(x),\varphi_c(x')] = 0. \tag{17}$$

The field $\varphi_a(x)$ satisfying (1) and (2) is then constructed as

$$\varphi_a(x) = (1+\lambda^2)^{-1/2}(\varphi_b(x) + \lambda\varphi_c(x)) \tag{18}$$

with an arbitrary real parameter λ. By putting either

$$\varphi_1(x) = (1+\lambda^2)^{-1/2}(\varphi_b^{(+)}(x) + \lambda\varphi_c^{(+)}(x)) \tag{19a}$$

and

$$\varphi_2(x) = (1+\lambda^2)^{-1/2}(\varphi_b^{(-)}(x) + \lambda\varphi_c^{(-)}(x)) \tag{19b}$$

or

$$\varphi_1(x) = (1+\lambda^2)^{-1/2}(\varphi_b^{(+)}(x) + \lambda\varphi_c^{(-)}(x)) \tag{20a}$$

and

$$\varphi_2(x) = (1+\lambda^2)^{-1/2}(\varphi_b^{(-)}(x) + \lambda\varphi_c^{(+)}(x)), \tag{20b}$$

where $(\pm)$ denote positive and negative energy parts as before, we obtain two decompositions of $\varphi_a(x)$ satisfying (13) and (14). Recognizing in (19) but another guise of the decomposition (3), the only nontrivially different decomposition is therefore (20). This decomposition *mixes positive and negative energy parts* and will lead to the necessity of introducing an indefinite metric into the formalism (see Sec. III).

We finally derive the formula for the generalized contraction symbol as follows: We rewrite (12) in the form

$$\begin{aligned}C'(x,x') = &-\theta(x_0'-x_0)[\varphi_a(x),\varphi_a(x')] \\ &+ \varphi_a(x)\varphi_a(x') - N'(\varphi_a(x),\varphi_a(x')).\end{aligned} \tag{21}$$

The last two terms in (21) can be rewritten as

$$\begin{aligned}&(\varphi_1(x)+\varphi_2(x))(\varphi_1(x')+\varphi_2(x')) \\ &\quad - N'(\varphi_1(x)+\varphi_2(x),\varphi_1(x')+\varphi_2(x')) = [\varphi_2(x),\varphi_1(x')]\end{aligned} \tag{22}$$

which by (20) becomes

$$\begin{aligned}&(1+\lambda^2)^{-1}\{[\varphi_b^{(+)}(x),\varphi_b^{(-)}(x')] + \lambda^2[\varphi_c^{(-)}(x),\varphi_c^{(+)}(x')]\} \\ &\quad = [i/(1+\lambda^2)]\{\Delta^{(+)}(x-x';m) + \lambda^2\Delta^{(-)}(x-x';m)\}.\end{aligned} \tag{23}$$

Combining (23) and (2) gives

$$\begin{aligned}C'(x,x') = i\{&-\theta(x_0'-x_0)\Delta(x-x';m) \\ &+ (1+\lambda^2)^{-1}\Delta^{(+)}(x-x';m) \\ &+ [\lambda^2/(1+\lambda^2)]\Delta^{(-)}(x-x';m)\}.\end{aligned} \tag{24}$$

The $\Delta^{(\pm)}$ functions have the well-known properties

$$\Delta^{(-)}(x;m) = (\overline{\Delta^{(+)}(x;m)}) = -\Delta^{(+)}(-x;m), \tag{25}$$

$$\Delta^{(+)}(x;m) + \Delta^{(-)}(x;m) = \Delta(x;m). \tag{26}$$

The Feynman propagator Δ_F, and $\bar{\Delta}$ are defined by

$$\Delta_F(x;m) = \theta(-x_0)\Delta^{(-)}(x;m) - \theta(x_0)\Delta^{(+)}(x;m), \tag{27}$$

$$\bar{\Delta}(x;m) = \tfrac{1}{2}(\theta(-x_0) - \theta(x_0))\Delta(x;m). \tag{28}$$

The overall proportionality coefficients in those definitions are so chosen that

$$(\Box + m^2)\Delta_F(x;m) = (\Box + m^2)\bar{\Delta}(x;m) = +\delta^4(x). \tag{29}$$

With these, Eq. (24) is reduced to

$$C'(x,x') = -i\left\{\frac{1-\lambda^2}{1+\lambda^2}\Delta_F(x-x';m) + \frac{2\lambda^2}{1+\lambda^2}\bar{\Delta}(x-x';m)\right\}. \tag{30}$$

Equation (30) represents—at least from the formal point of view—the desired generalized contraction symbol (internal line propagator). It indeed reproduces the standard result $-i\Delta_F(x-x';m)$ for $\lambda=0$, the main *novel* interest attaching in this paper to the case $\lambda=1$, where this propagator becomes $-i\bar{\Delta}(x-x';m)$ (and leads, as mentioned in the Sec. I to definite *computational* advantages when applied to the Bethe–Salpeter equation; see Sec. VI). It may be of some theoretical interest per se that this propagator is also relativistically "strictly causal" (i.e., that it vanishes for spacelike $x-x'$), that is to say that the class of propagators which we are endeavoring to show as admissible to represent internal particle lines contains in fact such "strictly causal" propagators.

Before proceeding to physical applications of the new algorithm, we devote the sections immediately following to a fuller physical and mathematical analysis of the new features of the formalism associated with the requirements (a) of increasing the number of the independent fields and (b) of mixing the positive and negative energy parts in the new definition of the creation and annihilation operators. This discussion will proceed for some time in terms of arbitrary λ.

III. FIELD-THEORETICAL CONSEQUENCES (ENLARGEMENT OF THE REPRESENTATION SPACE OF THE FIELD ALGEBRA; NEED FOR INDEFINITE METRIC)

The introduciton of *two* independent fields, $\varphi_b(x)$ and $\varphi_c(x)$, from which to construct $\varphi_a(x)$ makes it necessary to account in our formalism for a second linear combination of these fields *independent* of (18). The field

$$\varphi_d(x)=(1+\lambda^2)^{-1/2}(-\lambda\varphi_b(x)+\varphi_c(x)) \tag{31}$$

is independent of $\varphi_a(x)$ in the sense that

$$[\varphi_d(x),\varphi_a(x')]=0, \tag{32}$$

and also satisfies the commutation relation of the type (16)

$$[\varphi_d(x),\varphi_d(x')]=i\Delta(x-x';m). \tag{33}$$

This requires, in a well-known way, an extension of the original physical state vector space representation of the field algebra generated by $\varphi_a(x)$ [with the tacit inclusion of all other fields interacting with $\varphi_a(x)$ and comprising the original dynamical system] to a larger irreducible representation space R of an algebra generated by $\varphi_a(x)$ *and* $\varphi_d(x)$. In a *given* representation space R the original physical space P is to be identified with the subspace of R characterized by the absence of the d quanta, provided, however, that P (though *not* necessarily R) be Hilbert.

To further specify the most convenient choice of R in a unique way, we first notice that the Eqs. (18) and (31) are pseudounitary transforms—in an, in general, non-Hilbert R—of $\varphi_b(x)$ and $\varphi_c(x)$, respectively. Thus

$$\varphi_a(x)=U\varphi_b(x)U^*, \tag{34a}$$

$$\varphi_d(x)=U\varphi_c(x)U^*, \tag{34b}$$

$$U^*U=1; \tag{35}$$

where

$$U=\exp\left\{-i\tan^{-1}\lambda\int d^3\mathbf{x}(\varphi_b\partial_0\varphi_c-\varphi_c\partial_0\varphi_c)\right\}, \tag{36}$$

as can be seen from the well-known formula

$$e^ABe^{-A}=\sum_{n=0}^{\infty}\frac{1}{n!}\left[\overset{(n)}{A},B\right] \tag{37}$$

by the repeated use of the commutation relations

$$\left[\int(\varphi_b(x')\partial_0'\varphi_c(x')-\varphi_c(x')\partial_b'(x'))d^3\mathbf{x}',\varphi_b(x)\right]$$
$$=i\int(\Delta(x'-x;m)\partial_0'\varphi_c(x')-\varphi_c(x')\partial_0'\Delta(x'-x;m))d^3\mathbf{x}=i\varphi_c(x) \tag{38}$$

and similarly derived

$$\left[\int(\varphi_b(x')\partial_0'\varphi_c(x')-\varphi_c(x')\partial_0'\varphi_b(x'))d^3\mathbf{x}',\varphi_c(x)\right]$$
$$=-i\varphi_b(x). \tag{39}$$

This in turn leads us to expect that it should be possible to introduce in R *two* pseudounitarily equivalent bases of states, one consisting of the usual multiparticle occupation states of the fields $\varphi_a(x)$ and $\varphi_d(x)$ with its "natural" vacuum $|0\rangle$ being the extension to R of the physical vacuum of P, and the other basis consisting of an analogous set of multiparticle occupation states of the fields $\varphi_b(x)$ and $\varphi_c(x)$ with *its* natural vacuum $|\omega\rangle$ (referred to in the sequel as the *unphysical* vacuum). The *transition matrix* between the two bases is U, and, in particular

$$|0\rangle=U|\omega\rangle. \tag{40}$$

The above two bases will be said to define either the $a-d$ or the $b-c$ *pictures*, in obvious analogy to the Heisenberg, Schrödinger, or interaction *pictures* which likewise refer to specific *bases* of the same representation space of the field algebra.

We now define R uniquely by the requirement that it should contain a (presumably nonzero and nondegenerate) unphysical vacuum $|\omega\rangle$ of the type alluded to above, itself defined specifically through the requirement that

$$\varphi_1(x)|\omega\rangle=0 \tag{41}$$

or, equivalently, through the dual relation

$$\langle\omega|\varphi_2(x)=\langle\omega|(\varphi_1(x))^*=0 \tag{42}$$

with $\varphi_1(x)$ and $\varphi_2(x)$ given by (20), in strict analogy to, respectively,

$$\varphi_a^{(+)}(x)|0\rangle=0 \tag{43}$$

and

$$\langle 0|\varphi_a^{(-)}(x)=\langle 0|(\varphi_a^{(+)}(x))^*=0 \tag{44}$$

and pertaining to the decomposition (3). [Relation (43) is, as presently seen, also a *consequence* of (41).]

In terms of $|\omega\rangle$ the generalized contraction symbol is

$$C'(x,x')=\langle\omega|T(\varphi_a(x),\varphi_a(x'))|\omega\rangle \tag{45}$$

in strict analogy to the standard

$$C(x,x')=\langle 0|T(\varphi_a(x),\varphi_a(x'))|0\rangle. \tag{46}$$

Emphasizing the mixing of positive and negative energy parts in (20) we next notice that (41) holds only when simultaneously

$$\varphi_b^{(+)}(x)|\omega\rangle = 0 \tag{47}$$

and

$$\varphi_c^{(-)}(x)|\omega\rangle = 0 \tag{48}$$

so that $|\omega\rangle$ is actually defined—as it should in an enlarged representation space R—by *two* relations pertaining to two independent fields $\varphi_b(x)$ and $\varphi_c(x)$. As the relations of the type (34) obviously apply to the positive and negative energy parts of the corresponding fields separately and because of (47) and (48), we have for the physical vacuum

$$\varphi_a^{(+)}(x)|0\rangle = U\varphi_b^{(+)}(x)U^*U|\omega\rangle = U\varphi_b^{(+)}(x)|\omega\rangle = 0 \tag{49}$$

and

$$\varphi_d^{(-)}(x)|0\rangle = U\varphi_c^{(-)}(x)U^*U|\omega\rangle = U\varphi_c^{(-)}(x)|\omega\rangle = 0, \tag{50}$$

Eq. (49) furnishing the proof that indeed (43) follows from (41). Because of left-hand equation in (50) we can characterize the physical subspace P of R as spanned by vectors $|\Psi\rangle$ for which

$$\varphi_d^{(-)}(x)|\Psi\rangle = 0. \tag{51}$$

The above concludes in principle the formulation of our algorithm, but must still be augmented by metric considerations allowing us to conclude that P *is* Hilbert and thus is in fact the physical space of the standard theory, despite the fact that R *cannot* be Hilbert because of condition (48).

Condition (47) is equivalent to

$$b_\alpha|\omega\rangle = 0, \tag{52}$$

where the creation and annihilation operators b_α^* and b_α enter the expansion

$$\varphi_b(x) = \sum_\alpha (f_\alpha(x)b_\alpha + \overline{f_\alpha(x)}b_\alpha^*) \tag{53}$$

and $\{f_\alpha(x)\}$ and $\{\overline{f_\alpha(x)}\}$ are the same positive and negative energy solutions of the KG equation as in (4) and (5). Their commutation relations are

$$[b_\alpha, b_\beta] = 0; \quad [b_\alpha, b_\beta^*] = \delta_{\alpha\beta}. \tag{54}$$

These, in conjunction with (52), lead (for one degree of freedom) in a standard way to the "normal" hierarchy of eigenstates of the number-of-particles operator b^*b belonging to positive integer eigenvalues n,

$$|n;\omega\rangle = (1/\sqrt{n!})(b^*)^n|\omega\rangle \tag{55}$$

which satisfy

$$\langle n';\omega|n;\omega\rangle = \delta_{n'n} \tag{56}$$

and so have positive norm, and for which

$$b|n;\omega\rangle = |n-1;\omega\rangle\sqrt{n} \tag{57}$$

and

$$b^*|n;\omega\rangle = |n+1;\omega\rangle\sqrt{n+1}. \tag{58}$$

The space spanned by vectors (55) is thus Hilbert and the b-quanta are represented in R by "normal" particles.

By contrast, the condition (48) is equivalent to

$$s_\alpha^*|\omega\rangle = 0, \tag{59}$$

where s_α^* and s_α enter the expansion

$$\varphi_c(x) = \sum_\alpha (f_\alpha(x)s_\alpha + \overline{f_\alpha(x)}s_\alpha^*) \tag{60}$$

and

$$[s_\alpha, s_\beta] = 0; \quad [s_\alpha, s_\beta^*] = \delta_{\alpha\beta} \tag{61}$$

the same as (53) and (54). However, it is known[6] that it is impossible to find a *Hilbert* (positive metric) representation space of an algebra generated by operators satisfying (61) and yet containing a (nonzero) vector $|\omega\rangle$ satisfying (59). An indefinite metric irreducible representation satisfying these conditions is possible.[7] We rename the starred and unstarred operators thus,

$$c_\alpha = s_\alpha^*, \quad c_\alpha^* = s_\alpha, \tag{62}$$

so that condition (59) assumes the form

$$c_\alpha|\omega\rangle = 0 \tag{63}$$

superficially resembling (52), while the commutation relations (61) become "pathological,"

$$[c_\alpha, c_\beta] = 0; \quad [c_\alpha, c_\beta^*] = -\delta_{\alpha\beta} \tag{64}$$

and, finally, the expansion (60) assumes the unconventional form

$$\varphi_c(x) = \sum_\alpha (\overline{f_\alpha(x)}c_\alpha + f_\alpha(x)c_\alpha^*). \tag{65}$$

The representation space, for one degree of freedom, is spanned[7] by eigenvectors of the operator c^*c belonging now to *negative* integer eigenvalues, $-m$,

$$|m;\omega\rangle = (1/\sqrt{m!})(c^*)^m|\omega\rangle, \tag{66}$$

with alternating sign of the norm

$$\langle m';\omega|m;\omega\rangle = (-1)^m\delta_{m'm} \tag{67}$$

the necessity for which is seen from the reversed sign in the first of the following relations

$$c|m;\omega\rangle = -|m-1;\omega\rangle\sqrt{m} \tag{68}$$

and

$$c^*|m;\omega\rangle = |m+1;\omega\rangle\sqrt{m+1} \tag{69}$$

as compared with (57) and (58).

The c quanta must therefore be represented in R by negative norm "ghosts" [the state exemplified by (66) being an "m-ghost state"] and R must indeed be non-Hilbert.

The transition matrix U given by (36) will thus mix normal states and ghosts—a circumstance worth remembering especially when working in the $b-c$ picture. Since U is pseudounitary and preserves norm, relations (34) show that the a quanta inherit from the b field the property of being positive norm particles, while the d quanta will be ghosts like the c quanta. Since its defining conditon (51) is just the condition expressing the absence of those d ghosts, the subspace P is Hilbert and does not change the physical content and the probabilistic interpretation of the original field theory.

IV. RELATION BETWEEN VACUUM COMPONENTS IN $a-d$ AND $b-c$ PICTURES

In this section we will establish the important relation that, for any physical state $|\Psi\rangle$, i.e., for which there are no d ghosts

$$d|\Psi\rangle = 0, \tag{70}$$

there is the simple proportionality

$$\langle\omega|\Psi\rangle = \sqrt{1+\lambda^2}\,\langle 0|\Psi\rangle \tag{71}$$

between its components with respect to the unphysical and physical vacua. This relation will be the basis in the next section of the correspondence between wave functions in the two pictures of bound states involving the exchange of φ_a particles.

First we note that, in analogy with (65), the d field can be written

$$\varphi_d(x) = \sum_\alpha (\overline{f_\alpha(x)}d_\alpha + f_\alpha(x)d_\alpha^*), \tag{72}$$

where

$$[d_\alpha, d_\beta] = 0; \quad [d_\alpha, d_\beta^*] = -\delta_{\alpha\beta} \tag{73}$$

so that by (49), (50)

$$a_\alpha|0\rangle = d_\alpha|0\rangle = 0. \tag{74}$$

The matrix U can also be written in the form

$$\begin{aligned} U &= \exp\Big(\beta\sum_\alpha(b_\alpha c_c - b_\alpha^* c_\alpha^*)\Big) \\ &= \exp\Big(\beta\sum_\alpha(a_\alpha d_\alpha - a_\alpha^* d_\alpha^*)\Big), \end{aligned} \tag{75}$$

where $\beta = \tan^{-1}\lambda$. Restricting to one degree of freedom again, with

$$\begin{aligned} |n,m;0\rangle &= \frac{(a^*)^n(d^*)^m}{\sqrt{n!m!}}|0\rangle = U|n,m;\omega\rangle \\ &= U\Big(\frac{(b^*)^n(c^*)^m}{\sqrt{n!m!}}|\omega\rangle\Big), \end{aligned} \tag{76}$$

we can establish the relation

$$\begin{aligned} |n,m;0\rangle &= \sum_{s=-m}^{\infty}|n+s,m+s;\omega\rangle \\ &\times(-1)^s\frac{\sqrt{(n+s)!}}{\sqrt{n!m!(m+s)!}} \\ &\times\frac{d^m}{d\beta^m}\left\{\frac{(\sin\beta)^{m+s}}{(\cos\beta)^{n+s+1}}\right\} \end{aligned} \tag{77}$$

which explicitly exhibits the presence of additional (bc) *pairs* in a basis vector of the $a-d$ picture when viewed from the $b-c$ picture (the converse being also true about (ad) pairs when a basis vector from the $b-c$ picture is viewed from the $a-d$ picture). We omit the proof.

Physical states are composed of $|n,m=0;0\rangle$ states which span P. Specializing (77) to them gives

$$\begin{aligned} |n,m=0;0\rangle &= \sum_{s=0}^{\infty}|n+s,s;\omega\rangle \\ &\times(-1)^s\sqrt{\frac{(n+s)!}{n!s!}}\,\lambda^s(\sqrt{1+\lambda^2})^{n+1} \end{aligned} \tag{78}$$

and for the physical vacuum, $n=0$, we get[8]

$$|0\rangle = \sqrt{1+\lambda^2}\,e^{-\lambda b^*c^*}|\omega\rangle. \tag{79}$$

The inverse of U is obtained by changing the sign of λ in (75). This combined with the fact that (75) has exactly the same form when written in terms of either b and c or a and d operators allows us to repeat the above with the roles of the $a-d$ and $b-c$ pictures reversed, so that

$$|\omega\rangle = \sqrt{1+\lambda^2}\,e^{\lambda a^*d^*}|0\rangle. \tag{80}$$

If $|\Psi\rangle$ satisfies (70), the dual to (80) proves (71).

V. PHYSICAL APPLICATIONS OF THE ALGORITHM; BETHE-SALPETER EQUATION

The search for physical applications of the now fully formulated new algorithm can at this point be simply understood as a search for possible computational advantages of performing actual calculations in the $b-c$ picture and *then* interpreting the results physically from the point of view of the $a-d$ picture.

As no possible advantage can be foreseen at this early stage from at once attempting an application to a more complicated theory, we shall once more confine ourselves to a simple theory of two scalar, complex KG fields $\varphi_A(x)$ and $\varphi_B(x)$ describing "scattered particles" A and B, interacting through the exhange of a third particle a (referred to in the sequel as the "exchange particle" or simply as a "gluon") described by the real scalar KG field $\varphi_a(x)$ and identified with the meson field considered so far to illustrate our new formalism [i.e., notwithstanding the possibility of extending our new algorithm to *all* the fields $\varphi_A(x)$, $\varphi_B(x)$, and $\varphi_a(x)$ simultaneously, which we however shall *not* attempt here.]

Using bold type characters to denote Heisenberg operators, the equations of motion of such a theory are accordingly

$$(\Box + m_A^2 + g\boldsymbol{\varphi}_a)\boldsymbol{\varphi}_A + 0, \tag{81}$$

$$(\Box + m_B^2 + g\boldsymbol{\varphi}_a)\boldsymbol{\varphi}_B = 0, \tag{82}$$

and

$$(\Box + m_a^2)\boldsymbol{\varphi}_a = -g(\boldsymbol{\varphi}_A^*\boldsymbol{\varphi}_A + \boldsymbol{\varphi}_B^*\boldsymbol{\varphi}_B). \tag{83}$$

These equations are derivable from the Lagrangian density

$$\mathbb{L} = \mathbb{L}_A + \mathbb{L}_B + \mathbb{L}_a + \mathbb{L}_{\text{int}}, \tag{84}$$

where

$$\mathbb{L}_{\substack{A\\B}} = (\partial_v\varphi^*_{\substack{A\\B}})(\partial^v\varphi_{\substack{A\\B}}) - m^2_{\substack{A\\B}}\varphi^*_{\substack{A\\B}}\varphi_{\substack{A\\B}}, \tag{85}$$

$$\mathbb{L}_a = \tfrac{1}{2}(\partial_v\varphi_a)(\partial^v\varphi_a) - \tfrac{1}{2}m_a^2\varphi_a^2, \tag{86}$$

and the interaction Lagrangian

$$\mathbb{L}_{\text{int}} = -\mathbb{H}_{\text{int}} = -g(\varphi_A^*\varphi_a\varphi_A + \varphi_B^*\varphi_a\varphi_B). \tag{87}$$

It is now necessary to augment the Lagrangian (84) by addition of the dynamics of the ghost field φ_d, namely,

$$\mathbb{L}_R = \mathbb{L} + \mathbb{L}_d, \tag{88}$$

with

$$\mathbb{L}_d = \tfrac{1}{2}(\partial_v\varphi_d)(\partial^v\varphi_d) - \tfrac{1}{2}m_a^2\varphi_d^2. \tag{89}$$

This comprises the $a-d$ picture in which the d-particle field operator $\varphi_d(x)$, being free, is identical with its Heisenberg counterpart $\boldsymbol{\varphi}_d(x)$; it should be however strongly emphasized that in the $b-c$ picture the fields $\boldsymbol{\varphi}_b(x)$ and $\boldsymbol{\varphi}_c(x)$, still related to $\boldsymbol{\varphi}_a(x)$ and $\varphi_d(x)$ through (18) and (31)[9] both *do* interact with the scattered particles A and B.

Starting with the well-known expansion for the S operator

$$S=\sum_n \frac{(-i)^n}{n!}\int_{-\infty}^{+\infty} d^4x_1\cdots\int_{-\infty}^{+\infty} d^4x_n \times T(\mathbf{H}_{\rm int}(x_1)\cdots\mathbf{H}_{\rm int}(x_n)) \tag{90}$$

but working in the $b-c$ picture—i.e., expressing the processes relating to the exchange particles in terms of emissions and absorptions of b and c particles rather than directly in terms of the a particles—we are then indeed confronted with the task of expressing the chronological products appearing in (90) in terms of generalized normal products defined in Sec. II (of course, only as far as the exchange particles are concerned); in short, the Feynman expansion for any particular scattering process is now in terms of graphs where internal exchange particle lines are represented by (30).

We shall confine ourselves in this paper to the problem of bound states of A and B through the exchange of φ_a particles, since not only the reinterpretation of *external* exchange particle lines in terms of $\varphi_a^{(+)}(x)$ and $\varphi_a^{(-)}(x)$, proper for the $a-d$ picture from that in terms of $\varphi_1(x)$ and $\varphi_2(x)$, given by (20) and proper for the $b-c$ picture, is here completely avoided, but mainly because of the seemingly significant computational advantage of our new algorithm in this particular application, which we shall endeavor to illustrate in the next section at least for the case $m_a=0$.

Since nonperturbative methods are essential for bound states, it should be realized from the onset that the scheme of embedding the physical space P in a larger representation space R, as well as performing the change of basis in R, is completely independent of a perturbative approach. The Bethe–Salpeter equation, being nonperturbative, allows a good test of this algorithm. The two steps in solving an AB bound state problem will be (1) to set up a Bethe–Salpeter equation in the $b-c$ picture and solve it in the desired approximation of its kernel for the particular value of λ for which the solution is easiest computationally, and (2) to interpret the wave function so obtained (from the point of view of the physical $a-d$ picture).

Stressing the importance of the second step, we shall actually *define* here the Bethe–Salpeter equation as that satisfied by the wave function

$$\psi(x_1,x_2;\Omega)=\langle\Omega\,|T(\varphi_A(x_1),\varphi_B(x_2))|\Psi\rangle, \tag{91}$$

where the vacuum vector $|\Omega\rangle$ can now stand for either $|0\rangle$ or $|\omega\rangle$ and where $|\Psi\rangle$ is any (Heisenberg) state from R containing one particle A and one particle B.

Special significance attaches then to the case of $|\Psi\rangle$ being a physical vector from P, since the vector

$$T(\varphi_A(x_1)\varphi_B(x_2))|\Psi\rangle \tag{92}$$

must then also lie in P, and so, by the discussion of Sec. IV, the amplitude $\psi(x_1,x_2;0)$ and $\psi(x_1,x_2;\omega)$ must be *proportional*. This means that *every physical solution of the standard Bethe–Salpeter equation* (i.e., of an equation written only for the physical sector and thus *a priori* exhibiting no d ghost, but *also* interpretable as the Bethe–Salpeter equation in the $a-d$ picture) *must be found among the solutions of the corresponding Bethe–Salpeter equation pertaining to the $b-c$ picture*. In particular, the bound state spectrum, as defined, e.g., by the energy poles in the $\psi(x_1,x_2;0)$ amplitudes, must be found in the bound state spectrum resulting from solving the Bethe–Salpeter equation in the $b-c$ picture, i.e, among the poles of $\psi(x_1,x_2;\omega)$.

The above is of course valid—in the strict sense of a mathematical theorem—only in the context of *exact* Bethe–Salpeter equations in the $a-d$ and $b-c$ pictures, respectively. Nevertheless—assuming that there is no *a priori* reason to trust the computational accuracy based on the commonly used g^2 approximation in the kernel (referred to mostly as the ladder approximation) more in one case than in the other—it *does* follow that the spectrum of the standard Bethe–Salpeter equation

$$\{(\Box_1+m_A^2)(\Box_2+m_B^2)+ig^2\Delta_{\rm F}(x_1-x_2;m_a)\}\times\psi(x_1,x_2;0)=0 \tag{93}$$

must coincide—within the numerical accuracy inherent only in the overall reliability of that approximation—with a subset of the spectrum of the corresponding Bethe–Salpeter equation in the $b-c$ picture, which for (93) is

$$\{(\Box_1+m_A^2)(\Box_2+m_B^2)+ig^2\overline{\Delta}(x_1-x_2;m_a)\}\times\psi(x_1,x_2;\omega)=0. \tag{94}$$

Equations (93) and (94) follow, by virtue of (45), (46), and (30), from the common generic origin

$$\{(\Box_1+m_A^2)(\Box_2+m_B^2)-g^2\langle\Omega\,|T(\varphi_a(x_1),\varphi_a(x_2))|\Omega\rangle\}\times\psi(x_1,x_2;\Omega)=0, \tag{95}$$

where again $|\Omega\rangle$ stands for either $|0\rangle$ or $|\omega\rangle$, rederived from field theory in the Appendix (in order to provide additional emphasis on the fact that the wave function of the Bethe–Salpeter equation *is* interpretable as (91)—a crucial link in the above discussion but somewhat deemphasized in the standard[10] derivation of the Bethe-Salpeter equation).

VI. EXACT SOLUTION OF (94) FOR $m_A=0$

We finally illustrate the computational advantages of our new algorithm by solving (94) exactly in the case $m_a=0$, which, as we shall see, is actually much easier to do than in a corresponding case of (93) (known as Wick equation) and which—as mentioned in Sec. I—provided in fact the main-motivation to develop this algorithm.

This presentation is a somewhat shortened version of this solution contained in the already mentioned paper,[3] but is now compiled in an entirely self-motivated way, i.e., not as a mere illustration of similar methods applied to Wick equation (where it was treated only as an ad hoc exactly soluble model not yet connected to physical reality by arguments proceeding from the basic principles).

We start with the well-known plane wave representation for $\overline{\Delta}$ defined by (28)

$$\overline{\Delta}(x;m)=\frac{1}{(2\pi)^4}\int d^4k\,e^{-ik\cdot x}P\left(\frac{1}{m^2-k^2}\right). \tag{96}$$

Equation (94) written in terms of Fourier transforms is

$$(m_A^2 - p_A^2)(m_B^2 - p_B^2)\psi(p_A, p_B) + \frac{ig^2}{(2\pi)^2}\int d^4k\, \psi(p_A - k, p_B - k) P\left(\frac{1}{m_a^2 - k^2}\right) = 0, \tag{97}$$

where $\psi(p_A, p_B)$ is the Fourier transform of the previously used wave function $\psi(x_1,x_2;\omega)$. The symbol P appearing in (96) and (97) means the principal part integration in the zeroth component of the four-momentum k, and it is in fact this particular feature that allows us to solve our problem exactly.

We now specifically consider the case $m_a = 0$. It is convenient to work in terms of the function

$$u(p;r) = (m_A^2 - p_A^2)(m_B^2 - p_B^2)\psi(p_A, p_B) \tag{98}$$

rather than $\psi(p_A, p_B)$ itself, expressing at the same time the four-momenta p_A and p_B in terms of a constant four-vector r, representing the total energy-momentum of the bound state, and a variable four-momentum p, thus

$$p_A = \beta_+ r + p, \tag{99}$$

$$p_B = \beta_- r - p, \tag{100}$$

where the as yet undetermined constants β_+ and β_- satisfy

$$\beta_+ + \beta_- = 1. \tag{101}$$

In terms of this notation, Eq. (97) becomes

$$\frac{(2\pi)^4 i}{g^2} u(p;r) + \int d^4p'' \frac{u(p'';r)}{[(p'' + \beta_+ r)^2 - m_A^2][(p'' - \beta_- r)^2 - m_B^2]} \times P\left(\frac{1}{(p'' - p)^2}\right) = 0. \tag{102}$$

In order to arrive at a synthetic survey of the solutions of (102), we next propose to replace $u(p;r)$ by a generating function (chosen in the form of a "four-point function" pertaining to the scattering process $p_A + p_B \to + p_A' + p_B'$), $(p|T(r)|p')$, and in fact endeavor to solve the following *inhomogeneous* version of (102),

$$\frac{(2\pi)^4 i}{g^2}(p|T(r)|p') + \int d^4p'' \frac{(p''|T(r)|p')}{[(p'' + \beta_+ r)^2 - m_A^2][(p'' - \beta_- r)^2 - m_B^2]} \times P\left(\frac{1}{(p'' - p)^2}\right) = \frac{1}{(p - p')^2}. \tag{103}$$

Working specifically in the rest frame of the bound state of A and B,

$$\mathbf{r} = 0; \quad r_0 = E,$$

and defining a new generating function Λ by

$$(p_0, |\mathbf{p}|\,|\Lambda(E,z'')|\, p_0', |\mathbf{p}|) = |\mathbf{p}|\,|\mathbf{p}'|(p|T(r)|p'), \quad \text{etc.}, \tag{104}$$

Eq. (103) gives

$$\frac{2(2\pi)^4 i}{g^2}(p_0, |\mathbf{p}|\,|\Lambda(E,z'')|\, p_0', |\mathbf{p}|) + \int d^2\Omega'' \int_0^{+\infty} d|\mathbf{p}| \int_{-\infty}^{+\infty} dp_0'' \times \frac{(p_0', |\mathbf{p}''|\,|\Lambda(E,z)|\, p_0', |\mathbf{p}'|)}{[(p_0'' + \beta_+ E)^2 - |\mathbf{p}''|^2 - m_A^2][(p_0'' - \beta_- E)^2 - |\mathbf{p}''|^2 - m_B^2]} P\left(\frac{1}{z' - y'}\right) = \frac{1}{z'' - y''}, \tag{105}$$

where the variables z, z', and z'' denote cosines of the scattering angles (i.e., cosines of the angles between the vectors: $\mathbf{p}'$ and $\mathbf{p}''$; $\mathbf{p}''$ and $\mathbf{p}$, and $\mathbf{p}$ and $\mathbf{p}'$, respectively), $\int d^2\Omega''(\cdots)$ stands for angular integration in the p'' variable, and where, finally, the variables y, y' and y'' are defined by

$$y = \frac{|\mathbf{p}'|^2 + |\mathbf{p}''|^2 - (p_0' - p_0'')^2}{2|\mathbf{p}'|\,|\mathbf{p}''|} \quad \text{etc.}, \tag{106}$$

the definitions of y' and y'' obtaining by cyclic permutations of unprimed, primed, and double-primed quantities.

Included in the next step will be an extension of the integration in the $|\mathbf{p}''|$ variable from $-\infty$ to $+\infty$ instead of from 0 to $+\infty$ (which we shall nevertheless continue denoting by $|\mathbf{p}''|$ to avoid unnecessary confusion) by the simple expedient of extending the function Λ to negative values of $|\mathbf{p}''|$ through

$$(p_0, -|\mathbf{p}|\,|\Lambda(E,z'')|\, p_0', |\mathbf{p}'|) = -(p_0, |\mathbf{p}|\,|\Lambda(E,z'')|\, p_0' |\mathbf{p}'|), \tag{107}$$

compatible with (105) (bringing about the appearance of an additional factor $\frac{1}{2}$ before the integral).

We reduce the number of integration variables from four to three by (a) expressing the angular integration in terms of the variables z and z' and (b) performing the z' integration explicitly, since the unknown function $(p_0'', |\mathbf{p}''|\,|\Lambda(E,z)|\, p_0, |\mathbf{p}'|)$ does not depend on z'. Thus

$$\frac{2(2\pi)^4 i}{g^2}(p_0, |\mathbf{p}|\,|\Lambda(E,z'')|\, p_0', |\mathbf{p}'|) + \frac{2\pi}{2}\int_{-\infty}^{+\infty} d|\mathbf{p}''| \int_{-\infty}^{+\infty} dp_0'' \int_{-1}^{+1} dz \times \frac{(p_0''|\mathbf{p}''|\,|\Lambda(E,z)|\, p_0', |\mathbf{p}'|)}{[(p_0'' + \beta_+ E)^2 - |\mathbf{p}''|^2 - m_A^2][(p_0'' - \beta_- E)^2 - |\mathbf{p}''|^2 - m_B^2]} \operatorname{Im}\left(\frac{1}{1 - z^2 - y'^2 - z''^2 + 2zy'z'')^{1/2}}\right) = \frac{1}{z'' - y''}, \tag{108}$$

where $(1 - z^2 - y'^2 - z''^2 + 2zy'z'')^{1/2}$, understood as a function of y'; denotes that branch of this function, which, in the cut y' plane, assumes real positive values on the lower lip of the cut. The appearance of the imaginary part of the inverse of this function is again a significant feature in solving (94) instead of (93), for in the latter case it is replaced by a much more complicated expression, practically blocking further progress towards an exact solution.[11]

Solving (108) involves a proper choice of auxiliary variables, which we recapitulate briefly as follows:

Assigning the values

$$\beta_+ = (E^2 \pm m_A^2 \mp m_B^2)/2E \tag{109}$$

to the previously undetermined $\beta_\pm$ satisfying (101) and going over from the p_0 and $|\mathbf{p}|$ variables (analogously for primed and double-primed quantities) to

$$v_\pm = \frac{p_0 \pm |\mathbf{p}| + l}{p_0 \pm |\mathbf{p}| - l}, \tag{110}$$

where

$$l = (1/2E)[(E - m_A - m_B)(E - m_A + m_B)(E + m_A - m_B)(E + m_A + m_B)]^{1/2}, \tag{111}$$

we achieve that

$$\frac{d p_0'' \, d|\mathbf{p}''|}{[(p_0'' + \beta_+ E)^2 - |\mathbf{p}''|^2 - m_A^2][(p_0'' - \beta_- E)^2 - |\mathbf{p}''|^2 - m_B^2]}$$

$$= -\frac{1}{2}\frac{1}{(\beta_+ E + l)(\beta_- E - l)}\frac{dv''_+ \, dv''_-}{(v''_+ v''_- - u^2_+)(v''_+ v''_- - u^2_-)}, \tag{112}$$

where

$$u^2_\pm = \frac{\beta_\pm E \mp l}{\beta_\pm E \pm l}, \tag{113}$$

and where the stress is on the fact that $dv''_+ \, dv''_-$ is multiplied in (112) by a function of the *product* $v''_+ v''_-$ only. We also opt at this point to perform our calculations for E in the scattering region ($E \geqslant m_A + m_B$) which makes l of (111) as well as the $v_\pm$ variables real (eventually "reaching" the bound state values of E by an analytic continuation of the results in the physical E sheet).

To represent *angular* variables, we first notice that as a consequence of (106) and (110)

$$\frac{y-1}{y+1} = \frac{(v''_+ - v'_+)(v''_- - v'_-)}{(v''_+ - v'_-)(v''_- - v'_+)} \text{ etc.}, \tag{114}$$

and then *define* the t parameters through

$$\frac{z-1}{z+1} = \frac{(v''_+ - tv'_+)(v''_- - tv'_-)}{(v''_+ - tv'_-)(v''_- - tv'_+)} \tag{115}$$

and

$$\frac{z''-1}{z''+1} = \frac{(v_+ - t''v'_+)(v_- - t''v'_-)}{(v_+ - t''v'_-)(v_- - t''v'_+)}, \tag{116}$$

respectively [the introduction of t' is not necessary since the variable z' was eliminated in the step leading from (105) to (108)]. Viewing (115) and (116) as quadratic equations for the determination of t and t'' for given z and z'', respectively, we then see that another pair of solutions, s and s'' also exists, related to the original t and t'' through

$$v''_+ v''_- = stv'_+ v'_- \tag{117}$$

and

$$v_+ v_- = s''t''v'_+ v'_-, \tag{118}$$

respectively. Under these circumstances we can actually *replace* other variables by *t and s* variables, which, as we shall presently see is actually the most significant step in our solution.

The most advantageous choice at this point is actually s, t, and

$$w'' = \cosh \psi'', \tag{119}$$

where the hyperbolic angle ψ'' is defined by

$$\tanh \psi'' = (v''_+ - v''_-)/(v''_+ + v''_-) \tag{120}$$

and the corresponding triple s'', t'', and w, in terms of which

$$\frac{dv''_+ \, dv''_- \, dz}{(1 - z^2 - y'^2 - z''^2 + 2zy'z'')^{1/2}}$$

$$= v'_+ v'_- \sinh \xi \frac{\sinh \psi}{\sinh \psi''}\frac{ds\, dt\, dw''}{(-Q)^{1/2}}. \tag{121}$$

Q is given by

$$\begin{aligned} Q = &-x^2 - x'^2 - x''^2 + 2xx'x'' + 2ww'(x'' - xx') \\ &+ 2w'w''(x - x'x'') + 2w''w(x' - x''x) \\ &+ w^2(x^2 - 1) + w'^2(x'^2 - 1) + w''^2(x''^2 - 1) \\ \equiv &Aw''^2 + Bw'' + C, \end{aligned} \tag{122}$$

where the further abbreviations mean

$$x = \cosh \xi, \tag{123}$$

$$x' = \cosh \xi', \tag{124}$$

$$x'' = \cosh \xi'', \tag{125}$$

with the hyperbolic angles ξ, ξ', and ξ'' defined by

$$\tanh \xi = (s - t)/(s + t), \tag{126}$$

$$\tanh \xi' = (st - s''t'')/(st + s''t''), \tag{127}$$

$$\tanh \xi'' = (s'' - t'')/(s'' + t'') \tag{128}$$

[note an asymmetry in the definition (127) caused by the absence of t'].

Realizing that

$$\frac{1}{\sqrt{Q}} = \frac{1}{\sinh \xi''}\frac{\partial}{\partial w''}\ln W, \tag{129}$$

where

$$W = \sqrt{A}\,(w'' + B/2A) + \sqrt{Q} \tag{130}$$

with the expressions A and B identified through (122), that

$$\frac{1}{z''-y''}=\frac{\sinh\psi\sinh\psi'}{\sinh\xi''}\Big(-\frac{1}{s''-1}+\frac{1}{t''-1}\Big), \tag{131}$$

defining a new generating function Φ through

$$\Phi(s'',t'',w)=\frac{\sinh\xi''}{\sinh\psi\sinh\psi'} \times(v_+,v_-|A(E,z'')|v'_+\,v'_-)\ \text{etc.}, \tag{132}$$

and finally invoking (112) and (121), the Eq. (108) becomes

$$\frac{2(2\pi)^4 i}{g^2}\Phi(s'',t'',w) + \frac{2\pi v'_+\,v'_-}{4}\frac{1}{(\beta_+E+l)(\beta_-E-l)} \times\int_{-\infty}^{+\infty}ds\int_{-\infty}^{+\infty}dt\int_{w''_-}^{w''_+}dw''\,\Phi(s,t,w'') \times\mathrm{Re}\Big(\frac{\partial\ln W}{\partial w''}\Big)\frac{1}{(v''_+\,v''_- -u^2_+)(v''_+\,v''_- -u^2_-)} =-\frac{1}{s''-1}+\frac{1}{t''-1}, \tag{133}$$

where the integration limits $w'_\pm$ correspond to the integration limits $z=\pm1$ in (108) and are

$$w''_\pm=xw'\pm\sqrt{x^2-1}\,\sqrt{w'^2-1}\,. \tag{134}$$

The exact solubility of (133) follows from the fact that, as can be shown,[12]

$$\frac{W(w''=w''_+)}{W(w''=w''_-)}=\frac{(s-s'')(t-t'')}{(s-t'')(t-s'')}, \tag{135}$$

enabling us to make the solving ansatz that Φ is independent of the w parameter, so that, performing the w'' integration, we obtain, consistently

$$\frac{2(2\pi)^4 i}{g^2}\Phi(s'',t'')+\frac{2\pi}{4}\frac{v'_+\,v'_-}{(\beta_+E+l)(\beta_-E-l)} \times\int_{-\infty}^{+\infty}ds\int_{-\infty}^{+\infty}dt \times\frac{\Phi(s,t)\ln\Big|\frac{(s-s'')(t-t'')}{(s-t'')(t-s'')}\Big|}{(v'_+\,v'_-\,st-u^2_+)(v'_+\,v'_-\,st-u^2_-)} =-\frac{1}{s''-1}+\frac{1}{t''-1}. \tag{136}$$

A glance at (136) allows us finally to conclude that it is soluble by the further ansatz

$$\phi(s,t)=G(s)-G(t), \tag{137}$$

reducing in fact the whole problem of finding our generating function to that of solving one integro-differential equation in one variable only.

Before proceeding with the solution it is worthwhile to again compare the present situation with the corresponding one for the Bethe–Salpeter equation (93). These remarks, though rather technical, are intended to illustrate the much greater simplicity associated with Eq. (94).

(a) The ansatz (137) is *not* allowed at the corresponding stage, for the Bethe–Salpeter equation (93) even if the previous ansatz about the w independence of Φ were allowed.[13] Instead, Φ would have to satisfy a second-order partial differential equation in both s and t.[14]

(b) This partial differential equation *is* separable, but leads to second-order ordinary differential equations of a complicated Heun type, a Fuchsian with *four* singular points, from which to determine the energy spectrum.

Continuing with the present solution, the integro-differential equation for $G(t)$ is

$$-\frac{2(2\pi)^4}{g^2}\frac{dG(t'')}{dt''}+\frac{2\pi v'_+\,v'_-}{4lE}\int_{-\infty}^{+\infty}K(t'',t)G(t)dt =-\frac{1}{(t''-1)^2} \tag{138}$$

with the kernel

$$K(t'',t)=\int_{-\infty}^{+\infty}ds\Big(\frac{1}{v'_+\,v'_-\,st-u^2_+}-\frac{1}{v'_+\,v'_-\,st-u^2_-}\Big) \times P\Big(\frac{1}{t''-t}-\frac{1}{t''-s}\Big) \tag{139}$$

and is actually obtained by substituting (137) into (136) and then additionally differentiating the result partially with respect to t'' in order to get rid of the logarithmic terms. If the expressions $(m_A^2+p_A^2)$ and $(m_B^2+p_B^2)$ appearing in (97) are interpreted as the reciprocals of the Fourier transforms of retarded propagators,[15] then

$$\frac{1}{v'_+\,v'_-\,st-u^2_\pm}\equiv\frac{1}{tv'_+\,v'_-} \times\lim_{\epsilon\to0}\frac{1}{s-(u^2_\pm/tv'_+\,v'_-)\pm i\epsilon} \tag{140}$$

and (138) becomes

$$\frac{-2(2\pi)^4 i}{g^2}\frac{dG(t'')}{dt''}+\frac{(2\pi)^2}{8lE} \times\int_{-\infty}^{+\infty}\frac{dt}{t}\Bigg(\frac{1}{t-t''+i\epsilon}+\frac{1}{t-t''-i\epsilon} +\frac{1}{t''-\frac{u^2_+}{v'_+\,v'_-\,t}+i\epsilon}+\frac{1}{t''-\frac{u^2_-}{v'_+\,v'_-\,t}-i\epsilon}\Bigg) \times G(t)=-\frac{1}{(t''-1)^2}. \tag{141}$$

Modelled on the well-known representation for the inhomogeneity

$$-\frac{1}{(t''-1)^2}=\frac{i}{2}\int_{-i\infty}^{+i\infty}\frac{\nu d\nu}{\sin\pi\nu}(-t)^{\nu-1}d\nu \tag{142}$$

the closed-form solution of (141) can be sought in the form

$$G(t)=\frac{i}{2}\int_{-i\infty}^{+i\infty}\frac{\nu d\nu}{\sin\pi\nu}\times\left[f(\nu)(-t)^{\nu}+h(\nu)\left(-\frac{1}{tv'_{+}v'_{-}}\right)^{v}\right]. \quad (143)$$

Owing to the circumstance that for noninteger ν the functions $(-t)^{-\nu}$ and $(-1/tv'_{+}v'_{-})^{\nu}$ must be considered independent, the functions $f(\nu)$ and $h(\nu)$ are finally obtained as solutions of

$$1=(M\nu+N)f(\nu)+(u_{+})^{-2\nu}Nh(\nu) \quad (144)$$

and

$$0=(u_{-})^{2\nu}Nf(\nu)+(M\nu+N)h(\nu), \quad (145)$$

with

$$M=2(2\pi)^4 i/g^2 \quad (146)$$

and

$$N=(2\pi)^3/8lE. \quad (147)$$

To obtain the energy spectrum of (94) we first determine the positions of poles (Regge poles) of $f(\nu)$ and $h(\nu)$ from the determinantal equation

$$\det\begin{pmatrix}M\nu+N & (u_{+})^{-2\nu}N\\ (u_{-})^{2\nu}N & Mv+N\end{pmatrix}=\left\{M\nu+N\left[1+\left(\frac{u_{-}}{u_{+}}\right)^{v}\right]\right\}\times\left\{M\nu+N\left[1-\left(\frac{u_{-}}{u_{+}}\right)^{\nu}\right]\right\}=0. \quad (148)$$

The bound state poles in $G(t)$—and therefore in all the previously defined amplitudes—obtain where, for integer $\nu=n$, these Regge poles "pinch" against the poles stemming from $1/\sin\pi\nu$ in (143). The bound state formula is thus obtained by equating to zero the { } brackets in (148) for $\nu=n$, which, recalling the definitions of l and $u_{\pm}$ given by (111) and (113) and (109), respectively, and finally introducing a very convenient parameter

$$Z=u_{-}/u_{+} \quad (149)$$

reads

$$\frac{4\pi in}{g^2}+\frac{1\pm Z^n}{4m_A m_B(Z-1/Z)}=0 \quad (150)$$

concurrent with

$$E=\sqrt{m_A^2+m_B^2+m_A m_B(Z+1/Z)}\,. \quad (151)$$

The bound state spectrum of the Bethe–Salpeter equation in the $b-c$ picture (94)—which, according to the discussion at the end of the previous section should contain the physical spectrum in an approximation as good as that following from the solution of the Wick equation (93)—is therefore obtained by the simple expedient of solving the algebraic equation of the $(n+1)$th degree (150) and then computing the bound state energies from (151).

It is interesting to note that—whatever the physical meaning of lower lying energy states—the "plus" branch of (150) reproduces easily at least the nonrelativistic Balmer–Ritz formula

$$E-m_A-m_B=-\left(\frac{g^2}{16\pi m_A m_B}\right)^2\frac{m_A m_B}{m_A+m_B}\frac{1}{2n^2}, \quad (152)$$

i.e., for energies E just below the scattering threshold m_A+m_B [and therefore Z in the vicinity of 1, which can be represented there by a power series in n through (150), the leading term leading to (152)].

By contrast, in the case of the Wick equation (93), the energy spectrum derived from the second-order differential equations mentioned under (b) above is known to exhibit many pathological analytical features, and is not obtainable in an analytically closed form to the extent that even the clear-cut correspondence with the nonrelativistic Ritz–Balmer formula is somewhat dubious.

Finally, unlike (93), the solution of (94) does not necessitate the so called Wick rotation (auxiliary transition to the Euclidean metric of the original four-momentum space by rotating the integration path in k_0) and is performed entirely in the original non-Euclidean space [in fact the Wick rotation is certainly not allowed in the case of (94)].

ACKNOWLEDGMENTS

It is with great pleasure that the author would like to acknowledge and thank for many stimulating and thorough discussions pertaining to this work Professor A. S. Wighman of Princeton University and Professor Bernard Goodman of the University of Cincinnati; the latter also for his help in editing the manuscript both linguistically and for clarity.

The author would also like to thank Professor Res Jost of ETH Zürich for his interest in the exact solution of Sec. VI, especially in the physical meaning of the spectrum given by Eqs. (149)–(151).

APPENDIX

In order to derive Eq. (95) of the main text, we shall first prove the theorem, valid *exactly* in terms of the operators involved, to the effect that

$$\begin{aligned}\varphi_0(x_1,x_2)&=\varphi(x_1,x_2)\\&+g\int_{-\infty}^{+\infty}\Delta_{\rm Ret}(x_1-x_1';m_A)\varphi_a(x_1')\varphi(x_1',x_2)d^4x_1'\\&+g\int_{-\infty}^{+\infty}\Delta_{\rm Ret}(x_2-x_2';m_B)\varphi_a(x_2')\varphi(x_1,x_2')d^4x_2'\\&+g^2\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\Delta_{\rm Ret}(x_1-x_1';m_A)\\&\times\Delta_{\rm Ret}(x_2-x_2';m_B)\\&\times T'(\varphi_a(x_1'),\varphi_a(x_2'))\varphi(x_1',x_2')d^4x_1'd^4x_2'\end{aligned} \quad (A1)$$

is a *free* field in both arguments x_1 and x_2, thus

$$(\Box_1+m_A^2)\varphi_0(x_1,x_2)=(\Box_2+m_B^2)\varphi_0(x_1,x_2)=0. \quad (A2)$$

Here $\varphi(x_1',x_2')$ is defined by

$$\varphi(x_1,x_2)=T(\varphi_A(x_1),\varphi_B(x_2)). \quad (A3)$$

T and T' stand for chronological and antichronological ordering, respectively, and

$$\Delta_{\rm Ret}(x;m)=-\theta(x_0)\Delta(x;m) \quad (A4)$$

and is so defined that

$$(\Box + m^2)\Delta_{\rm Ret}(x;m) = +\delta^4(x) \tag{A5}$$

[compare (29) of the main text].

The above theorem, its proof and the subsequent derivation of the Bethe–Salpeter equation, represent a simplified version of a more general procedure involving n interacting particles and applied to QED, published elsewhere.[16]

In the case of two bodies the proof of the theorem proceeds most simply from the lemma

$$T'((\Box_1 + m_A^2 + g\varphi_a(x_1)) \times(\Box_2 + m_B^2 + g\varphi_a(x_2)))\varphi(x_1,x_2) = 0 \tag{A6}$$

tantamount essentially[17] to the validity of the following two statements:

$$(\Box_1 + m_A^2 + g\varphi_a(x_1)) \times(\Box_2 + m_B^2 + g\varphi_a(x_2))\varphi_B(x_2)\varphi_A(x_1) = 0 \tag{A7}$$

valid for point events x_2 later than x_1 because of the field equation (82), and

$$(\Box_2 + m_B^2 + g\varphi_a(x_2)) \times(\Box_1 + m_A^2 + g\varphi_a(x_1))\varphi_A(x_1)\varphi_B(x_2) = 0 \tag{A8}$$

valid in the opposite case of x_1 being later than x_2, because of (81). We actually prove the first equation in (A2)

$$(\Box_1 + m_A^2 + g\varphi_a(x_1))\varphi(x_1,x_2) + g\int_{-\infty}^{+\infty} d^4x_2'\Delta_{\rm Ret}(x_2 - x_2';m_B) \times T'(\varphi_a(x_2'),(\Box_1 + m_A^2 + g\varphi_a(x_1))) \times\varphi(x_1,x_2') = 0, \tag{A9}$$

the second following by reversing the roles of particles A and B. Introducing the abbreviation

$$K(x_1,x_2)\equiv(\Box_1 + m_A^2 + g\varphi_a(x_1))\varphi(x_1,x_2) \tag{A10}$$

we have

$$K(x_1,x_2) = 0 \quad \text{for } x_1 \text{ later than } x_2, \tag{A11}$$

whereas the lemma [(A6)] can be rewritten in the form

$$(\Box_2 + m_B^2)K(x_1,x_2) = -gT(\varphi_a(x_2),(\Box_1 + m_A^2 + g\varphi_a(x_1))\varphi(x_1,x_2). \tag{A12}$$

Consequently,

$$\begin{aligned} g&\int_{-\infty}^{+\infty} d^4x_2'\Delta_{\rm Ret}(x_1 - x_2';m_B)T'(\varphi_a(x_2'),(\Box_1 + m_A^2 + g\varphi_a(x_1)))\varphi(x_1,x_2') \\ &= -\int_{-\infty}^{+\infty} d^4x_2'\Delta_{\rm Ret}(x_2 - x_2';m_B)(\Box_2 + m_B^2)K(x_1,x_2') \\ &= \int_{-\infty}^{x_{02}} d^4x_2'\Delta(x_2 - x_2';m_B)(\Box_2 + m_B^2)K(x_1,x_2') \\ &= K(x_1,x_2) + \lim_{x_{02}\to-\infty}\int d^3\mathbf{x}_2'\Delta(x_2 - x_2';m_B)(\overrightarrow{\partial}_{02}' - \overleftarrow{\partial}_{02}')K(x_1,x_2'), \end{aligned} \tag{A13}$$

the proof of (A9) accomplished by the vanishing of the last surface integral as a consequence of (A11).

To derive the Bethe–Salpeter equation in the g^2 approximation in its kernel (in either of the $a-d$ or $b-c$ pictures) we first notice that taking the equality (A1) between the brackets $\langle\Omega\,|$ and $|\Psi\rangle$ *as it stands* yields a linear relationship not only between the wave function (91) (through the zeroth and second-order terms in g) and two-gluon terms (through the terms quadratic in g), but also, through the linear term, between one-gluon amplitudes

$$\langle\Omega\,|\varphi_a(y)T(\varphi_a(x_1),\varphi_B(x_2))|\Psi\rangle,$$

which have therefore to be eliminated in order to obtain a self-consistent equation containing only terms of the type (91) in *all* terms up to the second order in g. This is accomplished by iterating (A1) *prior* to putting it between the brackets $\langle\Omega\,|$ and $|\Psi\rangle$ by expressing the zero-order g term $\psi(x_1,x_2)$, by the inhomogeneity $\varphi_0(x_1,x_2)$, and the higher-order g terms, and substituting it back into the linear g terms. Thus

$$\begin{aligned} I(x_1,x_2) = \varphi(x_1,x_2) &- g^2\int_{-\infty}^{+\infty} d^4x_1'\int_{-\infty}^{+\infty} d^4x_2'\Delta_{\rm Ret}(x_1 - x_1';m_A)\Delta_{\rm Ret}(x_2 - x_2';m_B)T(\varphi_a(x_1'),\varphi_a(x_2'))\varphi(x_1',x_2') \\ &+ g^2\int_{-\infty}^{+\infty} d^4x_1'\int_{-\infty}^{+\infty} d^2x_1''\Delta_{\rm Ret}(x_1 - x_1';m_A)\varphi_a(x_1')\Delta_{\rm Ret}(x_1' - x_1'';m_A)\varphi_a(x_1'')\varphi(x_1'',x_2) \\ &+ g^2\int_{-\infty}^{+\infty} d^4x_2'\int_{-\infty}^{+\infty} d^4x_2''\Delta_{\rm Ret}(x_2 - x_2';m_B)\varphi_a(x_2')\Delta_{\rm Ret}(x_2' - x_2'';m_B)\varphi_a(x_2'')\varphi(x_1,x_2'') + O(g^3), \end{aligned} \tag{A14}$$

where use was made of

$$T(\varphi_a(x),\varphi_a(y)) + T'(\varphi_a(x),\varphi_a(y)) = \varphi_a(x)\varphi_a(y) + \varphi_a(y)\varphi_a(x) \tag{A15}$$

and where the inhomogeneity

$$I(x_1,x_2) = \varphi_0(x_1,x_2) + g\int_{-\infty}^{+\infty}\Delta_{\rm Ret}(x_1 - x_1';m_A) \times\varphi_a(x_1')\varphi_0(x_1',x_2)d^4x_1' + g\int_{-\infty}^{+\infty}\Delta_{\rm Ret}(x_2 - x_2';m_B) \times\varphi_a(x_2')\varphi_0(x_1,x_2')d^4x_2' \tag{A16}$$

is still a sum of free fields at least in one of the arguments x_1 and x_2, so that

$$(\Box_1 + m_a^2)(\Box_2 + m_B^2)I(x_1,x_2) = 0. \tag{A17}$$

By repeating the above iteration once more in the last two g^2 terms in (A14)—tantamount, in the g^2 approximation, to simply replacing them by terms containing $\varphi_0(x_1,x_2)$ instead of $\boldsymbol{\varphi}(x_1,x_2)$ and then incorporating them into a new inhomogeneity, $I'(x_1,x_2)$, still satisfying (A17)—we get (after also placing the pertinent equation between the brackets $\langle\Omega\,|$ and $|\Psi\rangle$)

$$\begin{aligned}\langle\Omega\,|I'(x_1,x_2)|\Psi\rangle &\\ = \psi(x_1,x_2;\Omega) &+ g^2\int_{-\infty}^{+\infty} d^4x_1'\int_{-\infty}^{+\infty} d^4x_2' \\ &\times\Delta_{\rm Ret}(x_1-x_1';m_A)\Delta_{\rm Ret}(x_2-x_2';m_B)\\ &\times\langle\Omega\,|T(\boldsymbol{\varphi}_a(x_1'),\boldsymbol{\varphi}_a(x_2'))\boldsymbol{\varphi}(x_1',x_2')|\Psi\rangle\end{aligned} \tag{A18}$$

or, in the differential form

$$\begin{aligned}&(\Box_1 + m_A^2)(\Box_2 + m_B^2)\psi(x_1,x_2;\Omega)\\ &\quad + g^2\langle\Omega\,|T(\boldsymbol{\varphi}_a(x_{1+}),\boldsymbol{\varphi}_a(x_2))\boldsymbol{\varphi}(x_1,x_2)|\Psi\rangle = 0.\end{aligned} \tag{A19}$$

The last step leading to Eq. (95) of the main text is now only the truncation of the expression

$$\langle\Omega\,|T(\boldsymbol{\varphi}_a(x_1),\boldsymbol{\varphi}_a(x_2))\boldsymbol{\varphi}(x_1,x_2)|\Psi\rangle \tag{A20}$$

into

$$\langle\Omega\,|T(\boldsymbol{\varphi}_a(x_1),\boldsymbol{\varphi}_a(x_2))|\Omega\rangle\psi(x_1,x_2;\Omega) \tag{A21}$$

in which we finally also replace the Heisenberg operators $\boldsymbol{\varphi}_a$ by their interaction counterparts φ_a, and two-gluon terms, which we discard (since *their* elimination effects the kernel only in the g^4-approximation).

In connection with the discussion at the end of Sec. V it is finally worth while to remark that it is only the last truncation that differentiates between (93) and (94) of the main text. By invoking the important Eq. (71) of Sec. IV for the expression (A20) in the case of $|\Psi\rangle\in P$, we see that, for physical $|\Psi\rangle$, the forms (A18) and (A19) are still simply identical in the $a-d$ and $b-c$ pictures (i.e., proportional with the common proportionality factor when written for $|\Omega\rangle = |0\rangle$ and $|\Omega\rangle = |\omega\rangle$, respectively).

[1]Compare, e.g., J. M. Jauch and F. Rohrlich, *The Theory of Photons and Electrons* (Springer-Verlag, Berlin, 1976), Appendix A4, pp. 445–453.

[2]Compare a short discussion of the methods of Bleuler and Gupta, in J. M. Jauch and F. Rohrlich, Ref. 1, pp. 100–110; especially of indefinite metric, p. 105. For a more extensive source dealing with the applications of indefinite metric to quantum field theory see, e.g., K. L. Nagy, *State Vector Spaces with Indefinite Metric in Quantum Field Theory* (Noordhoff, Groningen, The Netherlands, 1966).

[3]M. Günther, Phys. Rev. D **9**, 2411–2450 (1974); especially pp. 2436–2441.

[4]G. C. Wick, Phys. Rev. **96**, 1124 (1954); R. E. Cutkosky, Phys. Rev. **96**, 1135 (1954); H. S. Green, Nuovo Cimento **5**, 866 (1957); S. N. Biswas, *ibid.* **7**, 557 (1958) etc. For fuller literature see Ref. 3.

[5]Compare, e.g., S. S. Schweber and H. A. Bethe, *An Introduction to Relativistic Quantum Field Theory* (Row, Peterson, New York, 1961), pp. 745–6.

[6]Compare the discussion, K. L. Nagy, Ref. 2, pp. 22–3 and J. M. Jauch and F. Rohrlich, Ref. 1, pp. 104–5. We are here *forced* into the alternative (2.5)–(2.6) of K. L. Nagy which are our Eqs. (64) and (63), respectively [after the change of notation (62)].

[7]Our Eqs. (66), (67), (68), and (69) are Eqs. (2.7), (2.8), (2.12) of K. L. Nagy, Ref. 2, and Eqs. (6.56), (6.57), (6.63), and (6.63*) of J. M. Jauch and F. Rohrlich, Ref. 1, respectively.

[8]Compare formula (6-41), p. 102, J. M. Jauch and F. Rohrlich, Ref. 1. Note also that when acting on the particular vector $|\omega\rangle$ the more complicated operator $\exp\{\tan^{-1}\lambda\,(bc - b^*c)\}$ has the same effect as $\exp(-\lambda b^*c^*)$.

[9]Though the pseudounitary operator U of (34) must be replaced by the Heisenberg counterpart of (36). Since $\int d^3\mathbf{x}(\boldsymbol{\varphi}_b\partial_0\boldsymbol{\varphi}_c - \boldsymbol{\varphi}_c\partial_0\boldsymbol{\varphi}_b)$ is no longer time independent, the time at which this surface integral has to be taken can be arbitrarily fixed, e.g., as at $t = -\infty$.

[10]E. E. Salpeter and H. A. Bethe, Phys. Rev. **84**, 1232 (1951).

[11]Compare the detailed discussion of the same method when applied to the standard Bethe–Salpeter equation (93) (in its original non-Euclidean metric, prior to the Wick rotation), Ref. 3, pp. 2422–2436.

[12]See for detailed proof of (135), Ref. 3, Appendix A, pp. 2441–2445 where W was denoted by ξ or (up to a common factor) by $\bar{\xi}$. To accommodate both cases of Wick's rotated and unrotated *standard* Bethe–Salpeter equation, the formulas were also written in terms of ordinary rather than hyperbolic angles, thus, e.g., the previous ψ means now $i\psi$ etc. This change does *not* however effect the definitions of the s and t parameters which are the same as here.

[13]The question of whether or not the ansatz (in the case of the standard BS equation) of the w-independence of Φ can be made seems to be intimately connected with the question of the permissibility of the so called Wick rotation and was actually used in the paper (Ref. 3) as a test of the latter.

[14]The partial differential equation referred to is Eq. (160), p. 2432, Ref. 3.

[15]The question of what propagators to use for the *scattered* particles A and B is—in the particular context of the Bethe–Salpeter equation—of much lesser consequence than for the exchange particle since their choice depends entirely on the type of initial conditions imposed on the solutions of, respectively, the differential equations (93) or (94)—the initial conditions of particularly little importance for bound states. Still, the derivation of the Bethe–Salpeter equation of the Appendix *does* suggest the use of retarded propagators.

[16]M. Günther, Phys. Rev. **88**, 1411 (1952); **94**, 1347 (1954).

[17]However, a more detailed proof of the important premise (A6), with particular attention payed to the *step functions* involved, can be given as follows: Eq. (A3) can be written as

$$\begin{aligned}\boldsymbol{\varphi}(x_1,x_2) &= \theta(t_1-t_2)\boldsymbol{\varphi}_A(x_1)\boldsymbol{\varphi}_B(x_2) + \theta(t_2-t_1)\boldsymbol{\varphi}_B(x_2)\boldsymbol{\varphi}_A(x_1)\\ &= \theta(t_1-t_2)[\boldsymbol{\varphi}_A(x_1),\boldsymbol{\varphi}_B(x_2)] + \boldsymbol{\varphi}_B(x_2)\boldsymbol{\varphi}_A(x_1)\\ &= -\theta(t_2-t_1)[\boldsymbol{\varphi}_A(x_1),\boldsymbol{\varphi}_B(x_2)] + \boldsymbol{\varphi}_A(x_1)\boldsymbol{\varphi}_B(x_2).\end{aligned} \tag{R1}$$

Consequently, because of the field equations (81) and (82), the left-hand side of (A6) is

$$\begin{aligned}&\{\theta(t_2-t_1)(\Box_1 + m_A^2 + g\boldsymbol{\varphi}_a(x_1))(\Box_2 + m_B^2 + g\boldsymbol{\varphi}_a(x))\\ &\quad\times\theta(t_1-t_2) - \theta(t_1-t_2)(\Box_2 + m_B^2 + g\boldsymbol{\varphi}_a(x_2))\\ &\quad\times(\Box_1 + m_A^2 + g\boldsymbol{\varphi}_a(x_1))\theta(t_2-t_1)\}[\boldsymbol{\varphi}_A(x_1),\boldsymbol{\varphi}_B(x_2)].\end{aligned} \tag{R2}$$

Here, the significant fact is that (R2) can be written in the form of an operator acting only on a *commutator* of the Heisenberg fields $\boldsymbol{\varphi}_A(x_1)$ and $\boldsymbol{\varphi}_B(x_2)$. The proof that (R2) is zero follows mainly from

$$\begin{aligned}&(\Box_1 + m_A^2 + g\boldsymbol{\varphi}_a(x_1))\theta(t_1-t_2)[\boldsymbol{\varphi}_A(x_1),\boldsymbol{\varphi}_B(x_2)]\\ &\quad = \theta(t_1-t_2)(\Box_1 + m_A^2 + g\boldsymbol{\varphi}_a(x_1))[\boldsymbol{\varphi}_A(x_1),\boldsymbol{\varphi}_B(x_2)],\ \text{etc.}\end{aligned} \tag{R3}$$

and

$$\begin{aligned}&(\Box_1 + m_A^2 + g\boldsymbol{\varphi}_a(x_1))\theta(t_1-t_2)[\boldsymbol{\varphi}_A(x_1),\boldsymbol{\varphi}_a(x_2)]\\ &\quad = \theta(t_1-t_2)(\Box_1 + m_A^2 + g\boldsymbol{\varphi}_a(x_1))[\boldsymbol{\varphi}_A(x_1),\boldsymbol{\varphi}_a(x_2)],\ \text{etc.}\end{aligned} \tag{R4}$$

which are in turn true because, even in the Heisenberg picture, both the zeroth and the first time derivative of a commutator between dynamically different fields are zero in the limit $x_1 \rightarrow x_2$. The application of (R3) to (R2) gives first

$$\begin{aligned}&\theta(t_2-t_1)(\Box_1 + m_A^2 + g\boldsymbol{\varphi}_a(x_1))\theta(t_1-t_2)\\ &\quad\times(\Box_2 + m_B^2 + g\boldsymbol{\varphi}_a(x_2))\boldsymbol{\varphi}_A(x_1)\boldsymbol{\varphi}_B(x_2)\\ &\quad\times\theta(t_1-t_2)(\Box_2 + m_B^2 + g\boldsymbol{\varphi}_a(x_2))\theta(t_2-t_1)\\ &\quad\times(\Box_1 + m_A^2 + g\boldsymbol{\varphi}_a(x_1))\boldsymbol{\varphi}_B(x_2)\boldsymbol{\varphi}_A(x_1).\end{aligned} \tag{R5}$$

The expression (R5) is then seen to be zero by realizing that

$$\begin{aligned}&(\Box_1 + m_A^2 + g\boldsymbol{\varphi}_a(x_1))\boldsymbol{\varphi}_B(x_2)\boldsymbol{\varphi}_A(x_1)\\ &\quad = g[\boldsymbol{\varphi}_a(x_1),\boldsymbol{\varphi}_B(x_2)]\boldsymbol{\varphi}_A(x_1),\quad \text{etc.,}\end{aligned} \tag{R6}$$

followed by the application of (R4).

Integration conditions for first-order differential linear equations in higher-dimensional gauge theories

D. B. Fairlie
University of Durham, DH13LE, Durham, Great Britain

J. Nuyts
University of Mons, 7000, Mons, Belgium

(Received 13 June 1983; accepted for publication 9 November 1983)

Generalized self-duality type relations for gauge fields in higher-dimensional spaces are discussed from the point of view of orbits, strata, and stability groups. It is explained how to obtain them as integrability conditions for first-order gauge-invariant linear differential equations.

PACS numbers: 11.15. − q, 02.30.Hq

1. INTRODUCTION

In a recent article,[1] we have started to study the generalization of the usual self-dual (or anti-self-dual) relations in four dimensions to gauge theories in dimension d higher than four. In that article, we asked the question of how to generalize the Belavin and Zakharov equations[2] whose integration conditions are precisely the duality relations for the gauge fields in four dimensions. We present here in Sec. 3 a solution to this problem after a somewhat more global presentation of the duality equations given in Sec. 2. As we already showed, the dimension $d = 8$ possesses very specific properties. We will exhibit this case with more detail in Sec. 4. Our conclusions are drawn in Sec. 5.

2. SELF-DUALITY IN HIGHER DIMENSIONS

We start with the potentials A_μ of a gauge theory in dimension d, and for an arbitrary gauge group. The equations of motion for the gauge fields $F_{\mu\nu}$ derived from the covariant derivatives D_μ

$$D_\mu = \partial_\mu + A_\mu\ , \tag{2.1}$$

$$F_{\mu\nu} = [D_\mu, D_\nu] \tag{2.2}$$

are

$$D_\mu F_{\mu\nu} = 0 \tag{2.3}$$

while the Bianchi identities are

$$D_\mu \wedge F_{\rho\sigma} = 0\,, \tag{2.4}$$

where $\wedge$ designates the full antisymmetrization of μ, σ, and ρ.

It is obvious that, if F satisfies the linear relations

$$\lambda F_{\mu\nu} = T_{\mu\nu\rho\sigma} F_{\rho\sigma}, \tag{2.5}$$

where T is an arbitrary completely antisymmetric constant tensor, the equations of motion (2.3) are satisfied in virtue of the identities (2.4), provided T and λ are nonzero. While (2.3) are second-order equations for the potentials A, the "secular" relations (2.5) are simpler since only of first order in derivatives though still nonlinear.

The secular Eq. (2.5) has many interesting properties, some of which we now outline. First, for T fixed (2.5) can be considered as an eigenvalue equation for the $d(d-1)/2$ vectors F with the symmetric $d(d-1)/2 * d(d-1)/2$ matrix T (lines being indexed by μ,ν and columns by ρ,σ). The matrix T has zero trace

$$\mathrm{Tr}\, T = T_{\mu\nu\mu\nu} = 0\,. \tag{2.6}$$

Hence, there are $d(d-1)/2$ eigenvalues and their sum is zero. Distinct eigenvalues correspond to orthogonal subspaces of the F space. The relevant metric is

$$F^{(1)}F^{(2)} = F^{(1)}_{\mu\nu} F^{(2)}_{\mu\nu}\,. \tag{2.7}$$

On the other hand, the SO(d) space group acts on the F as the $d(d-1)/2$ dimensional adjoint representation, irreducible except for $d = 4$, where

$$6 = (3,1) + (1,3) \tag{2.8}$$

under the SU(2)*SU(2)/Z(2) decomposition of SO(4). Under these transformations, T behaves as a C^d_4 dimensional repesentation irreducible except for $d = 8$, where

$$70 = 35_s + 35_a \tag{2.9}$$

a SO(8) self-dual and anti-self-dual decomposition; see (2.16).

Since T is a fixed vector, it has a little group or stability group L, i.e., the group of those transformations of SO(d) which leave T unchanged

$$L \subset (\mathrm{SO}(d)\,. \tag{2.10}$$

For $h \in L$

$$T = hT\,. \tag{2.11}$$

For any $s \in \mathrm{SO}(d)$, if

$$\tilde{T} = sT \tag{2.12}$$

the stability group $\tilde{L}$ of $\tilde{T}$, the transform of T under s, is conjugated (and hence isomorphic) to L with

$$\tilde{h} = shs^{-1}\,. \tag{2.13}$$

When T and $\tilde{T}$ are related by (2.12) they are said to be on the same orbit. Two T not on the same orbit can also have isomorphic stability groups (for example: T and αT for any constant α different from zero). They are then said to be on the same stratum. The strata are thus the unions of the orbits with isomorphic stability groups. [The uninteresting vector $T = 0$ is a stratum on its own, its stability group being the full SO(d).]

Before applying these remarks to the secular Eq. (2.5), let us still note that the T space can be split completely into disjoint orbits and disjoint strata and that to any T there corresponds a stability group. Conversely we have not proved that any given subgroup L of SO(d) is automatically the stability group of some vector T. Given L, the simplest way to see if a T exists is to look if the decomposition of the C^d_4 dimensional antisymmetric representation of SO(d) contains one or more singlets for the subgroup.

If we apply the transformations of L on (2.5) we see that F and hF correspond to the same eigenvalue. Hence, the orbits of F under h lead to vectors of the same eigenvalue λ subspace.

The number of different eigenvalues in (2.5) will clearly tend to be smaller if the stability group is larger and decomposes F in fewer pieces. This stresses the interest of trying to find T with maximal subgroups of SO(d) as stability groups.

Let us now briefly recall the application of these ideas to the cases $d = 4$ and $d = 8$ but refer to our first paper for details.

The SO(4) case

There exists only one allowed T tensor (up to a multiplicative constant α)

$$T_{\mu\nu\rho\sigma} = \tfrac{1}{2}\epsilon_{\mu\nu\rho\sigma} \,. \tag{2.14}$$

Its stability group is SO(4) itself. The orbit has only one point and the stratum is given by αT. No true SO(4) subgroup is the stability group of a T.

Under SO(4), as already explained (2.8), the 6-dimensional space F splits into $3 + 3$, corresponding to the eigenvalues $+1$ and -1 of T and the usual self-dual and anti-self-dual decomposition. In this case, SO(4) self-duality coincides with the secular equation. This is not true for $d > 4$, where the generalized self-duality for the gauge fields F, i.e., the secular equation, is not identical, obviously, to SO(d) duality which relates an n-index ($n < d$) antisymmetric tensor with a $(d - n)$-index antisymmetric tensor.

The SO(8) case (see Appendix)

As is well known SO(8) contains SO(7) locally in two distinct ways. Here we are interested in the $\widetilde{\text{SO}}(7)$ embedding, where SO(8) contains the covering group of SO(7). The eight-dimensional vector of SO(8) is the spinor repesentation of $\widetilde{\text{SO}}(7)$. Moreover the decompositions are as follows (2.9):

$$A \;\; 8\rightarrow 8 \,, \tag{2.15a}$$

$$T_a \;\; 35\rightarrow 35 \,, \tag{2.15b}$$

$$T_s \;\; 35\rightarrow 1 + 7 + 27 \,, \tag{2.15c}$$

$$F \;\; 28\rightarrow 7 + 21 \,. \tag{2.15d}$$

We see that the self-dual part of T (2.15c),

$$T_{\mu\nu\rho\sigma} = \tfrac{1}{24}\,\epsilon_{\mu\nu\rho\sigma\alpha\beta\gamma\delta}\; T_{\alpha\beta\gamma\delta} \tag{2.16}$$

contains a singlet. This T has $\widetilde{\text{SO}}(7)$ as stability group. There are two orbits for F, one seven dimensional, the other 21 dimensional (strata are obtained by simple scale multiplication) with two corresponding eigenvalues of (2.5), namely λ and λ_{21}. Since the trace of the T matrix is zero (2.6)

$$7\lambda_7 + 21\lambda_{21} = 0 \,, \tag{2.17}$$

which shows that

$$\lambda_7 = -3 \tag{2.18}$$

and

$$\lambda_{21} = 1$$

by suitably normalizing T.

3. LINEAR DIFFERENTIAL EQUATIONS. INTEGRABILITY CONDITIONS

Let us now show how to obtain, in general, first-order gauge-invariant conditions whose integrability conditions are the generalized self-dual relations. Consider $G_{\mu\nu}$ a general antisymmetric tensor spanning the space of all the eigenvalues of (2.5) except one, say λ. The linear differential equations for G

$$D_\mu G_{\mu\nu} = 0 \tag{3.1}$$

imply, as integrability condition for all G, that F belongs to the space corresponding to the eigenvalue λ of (2.5). Indeed (3.1) imply

$$D_\nu D_\mu G_{\mu\nu} = \tfrac{1}{2}F_{\nu\mu}\, G_{\mu\nu} = 0. \tag{3.2}$$

Making the gauge indices more explicit, (3.2) is equivalent to

$$F^a_{\mu\nu}\lambda^{\;a}_{\alpha\beta} G^{\;\beta}_{\mu\nu} = 0, \tag{3.3}$$

where the λ's are the generators of the gauge group for the representation acting on G. Multiplying (3.3) by an arbitrary vector c and defining

$$\rho^a_{\mu\nu} = c^\alpha \lambda^{\;a}_{\alpha\beta} G^{\;\beta}_{\mu\nu}, \tag{3.4}$$

one sees that, when G and c vary arbitrarily, ρ ranges over the adjoint representation. Since F has to be orthogonal to the space spanned by ρ,

$$F^a_{\mu\nu}\,\rho^a_{\mu\nu} = 0, \tag{3.5}$$

it has to belong, for every a, to the space of the missing eigenvalue λ.

Equations (3.2) are the integration conditions of (3.1).

Let us apply this first to the four-dimensional case and defer the eight-dimensional case to the next section. Let

$$G_{\mu\nu} = \eta^a_{\mu\nu}\, g^a \;(\text{or } \eta'^a_{\mu\nu}\, g'^a) \,, \tag{3.6}$$

where $\eta^a_{\mu\nu}$ ($a = 1,2,3$) are the self-dual (and $\eta'^a_{\mu\nu}$ the anti-self-dual) projectors.

$$\eta^a_{\mu\nu} = \epsilon_{0a\mu\nu} + \delta_{0\mu}\delta_{a\nu} - \delta_{0\nu}\delta_{a\mu} \,, \tag{3.7}$$

$$\eta'^a_{\mu\nu} = \epsilon_{0a\mu\nu} - (\delta_{0\mu}\delta_{a\nu} - \delta_{0\nu}\delta_{a\mu}) \,. \tag{3.8}$$

They satisfy, see (2.14)

$$T_{\mu\nu\rho\sigma}\eta^a_{\rho\sigma} = \eta^a_{\mu\nu} \,, \tag{3.9}$$

$$T_{\mu\nu\rho\sigma}\eta'^a_{\rho\sigma} = -\eta'^a_{\mu\nu} \,, \tag{3.10}$$

$$\eta'^a_{\mu\nu}\eta^b_{\mu\nu} = 0 \,, \tag{3.11}$$

i.e., η belongs to the (3.1) space and η' to the (1.3) space. The linear gauge-invariant differential Eqs. (3.1) with (3.6) imply

$$F_{\mu\nu}\eta^a_{\mu\nu}g^a = 0 \;(\text{or } F_{\mu\nu}\eta'^a_{\mu\nu}g'^a = 0) \,, \tag{3.12}$$

which shows in virtue of (3.10) that F is of the form

$$F_{\mu\nu} = \eta'^a_{\mu\nu}F^a \quad (\text{or } F_{\mu\nu} = \eta^a_{\mu\nu}F^a) \,. \tag{3.13}$$

In this simple case since the space (3.1) [or (1,3)] can be spanned by squaring the two-dimensional spinors (2,1) [or (1,2)]

$$(2,1)\times(2,1) = (1,1) + (3,1) \tag{3.14}$$

and taking the symmetric part. One can write the 2×2 matrix

$$D = D_0 + i\sigma_a D_a \tag{3.15}$$

and take as differential equation[2]

$$D\Psi = 0 \quad (\text{or } D^t\Psi' = 0), \tag{3.16}$$

where the spinor Ψ (or Ψ') has to be taken of the form

$$\Psi = \chi(x)\Pi \tag{3.17}$$

proportional to a constant but arbitrary spinor Π. The connection with our general approach is obtained by considering the 2×2 matrices

$$G = \Psi * \Pi^t (i\sigma_2) \tag{3.18}$$

(and analogously for G') which span the (3,1) space [or (1,3)]. We also see here the requirement of constant Π since (3.16) then implies

$$DG = 0 \ (\text{or } D^t G' = 0) \tag{3.19}$$

the same equation as (3.1) with (3.6), except for notation.

4. LINEAR EQUATION IN THE SO(8) CASE. $\widetilde{\mathrm{SO}}(7)$ EMBEDDING

Since the eight-dimensional case has very important properties we now present it with more details.

We first present the general case using the results and notation of Appendix A, for the $\Lambda^{A}_{\mu\nu}$ and $M^{AB}_{\mu\nu}$ ($A,B = 1,...,7$; $\mu,\nu = 1,...,8$) 7- and 21-dimensional solutions of the secular equations.

The linear differential equations can be written

$$(\text{Case I})\ D_\mu \Lambda^{A}_{\mu\nu}\Psi^A = 0, \tag{4.1}$$

$$(\text{Case II})\ D_\mu M^{AB}_{\mu\nu}\Psi^{AB} = 0. \tag{4.2}$$

In the first case Ψ^A span the seven-dimensional (eigenvalue -3) space and in the second case $\Psi^{AB} = -\Psi^{BA}$ the 21-dimensional (eigenvalue $+1$) space. In the first case (4.1) implies

$$(\text{Case I})\ F_{\mu\nu} = M^{AB}_{\mu\nu}F^{AB}, \tag{4.3}$$

while in the second case (4.2) implies

$$(\text{Case II})\ F_{\mu\nu} = \Lambda^{A}_{\mu\nu}F^{A} \tag{4.4}$$

so that F is restricted to belong to the 21- (first case) or 7-dimensional space (second case). Equations (A.1)–(A.12) of the Appendix justify these results.

5. CONCLUSIONS

In this paper we have shown the relevance of the stability groups of the $T_{\mu\nu\rho\sigma}$ tensor which generalizes the four-dimensional $\epsilon_{\mu\nu\rho\sigma}$, and of the decomposition into orbits and strata of the spaces of antisymmetric tensors with two and four indices for the discussion of the generalized self-dual relations for the gauge fields in dimensions greater than four.

These general results have been made more explicit when $d = 8$ and when T is stable under an $\widetilde{\mathrm{SO}}(7)$ subgroup of SO(8), a particularly interesting case.

We have also shown how to obtain, in general, the self-duality relations as integration conditions for first-order gauge-invariant linear-differential equations, giving again the $d = 8$ case explicitly.

These results will be used in a subsequent article[3] to obtain spherically symmetric solutions for eight-dimensional gauge theories.

ACKNOWLEDGMENT

One of the authors (D.B.F) thanks the University of Mons for hospitality.

APPENDIX: $\widetilde{\mathrm{SO}}(7)$ EMBEDDING IN SO(8)

In our first paper we have shown that the $\widetilde{\mathrm{SO}}(7)$ embedding in SO(8) leads to the set of seven democratic equations

$$F_{81} + F_{72} + F_{45} + F_{36} = 0, \tag{A1}$$

$$F_{82} + F_{17} + F_{35} + F_{64} = 0, \tag{A2}$$

$$F_{83} + F_{74} + F_{52} + F_{61} = 0, \tag{A3}$$

$$F_{84} + F_{37} + F_{51} + F_{26} = 0, \tag{A4}$$

$$F_{85} + F_{76} + F_{14} + F_{23} = 0, \tag{A5}$$

$$F_{86} + F_{57} + F_{13} + F_{42} = 0, \tag{A6}$$

$$F_{87} + F_{65} + F_{43} + F_{21} = 0. \tag{A7}$$

Corresponding to these seven equations it is convenient to define the seven (antisymmetric) 8×8 $\Lambda^{A}_{\mu\nu}$ matrices ($A = 1,...7; \mu,\nu = 1,...,8$). [$\Lambda^{A}_{\mu\nu}$ has elements $+1$ only for the indices appearing in Eqs. (A1)–(A7), -1 by antisymmetry and 0 everywhere else]. These $\Lambda^{A}_{\mu\nu}$ are the Clebsch–Gordan coefficients of the $8\times 8\rightarrow 7$ product in $\widetilde{\mathrm{SO}}(7)$. Define then the 21 matrices ($M^{AB} = -M^{BA}; A,B = 1,...,7$)

$$M^{AB}_{\mu\nu} = -\tfrac{1}{4}(\Lambda^{A}_{\mu\sigma}\Lambda^{B}_{\sigma\nu} - \Lambda^{B}_{\mu\sigma}\Lambda^{A}_{\sigma\nu}), \tag{A8}$$

generators of the $\widetilde{\mathrm{SO}}(7)$ transformations

$$[M^{AB},M^{CD}] = \delta^{BC}M^{AD} + \delta^{AD}M^{BC} - \delta^{AC}M^{BD} - \delta^{BD}M^{AC} \tag{A9}$$

and also Clebsch–Gordan coefficients of $8\times 8\rightarrow 21$.

They satisfy the properties

$$\Lambda^{A}_{\mu\nu}\Lambda^{B}_{\mu\nu} = 8\delta^{AB}, \tag{A10}$$

$$M^{AB}_{\mu\nu}M^{CD}_{\mu\nu} = 4(\delta^{AC}\delta^{BD} - \delta^{AD}\delta^{BC}), \tag{A11}$$

$$\Lambda^{A}_{\mu\nu}M^{BC}_{\mu\nu} = 0, \tag{A12}$$

$$T_{\mu\nu\rho\sigma}\Lambda^{A}_{\rho\sigma} = -3\Lambda^{A}_{\mu\nu}, \tag{A13}$$

$$T_{\mu\nu\rho\sigma}M^{AB}_{\rho\sigma} = M^{AB}_{\mu\nu}. \tag{A14}$$

[1] E. Corrigan, C. Devchand, D. B. Fairlie, and J. Nuyts, "First order equations for gauge fields in spaces of dimensions greater than four," Nucl. Phys. B **214**, 452 (1983).
[2] A. Belavin and V. Zakharov, Phys. Lett. B **73**, 53 (1978).
[3] D. B. Fairlie and J. Nuyts, "Spherically symmetric solutions for gauge fields in eight dimensions" (to be published).

Character expansion for U(N) groups and U(N/M) supergroups

A. B. Balantekin
Center for Theoretical Physics, Laboratory for Nuclear Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 20 September 1983; accepted for publication 3 February 1984)

A formula is presented to generate expansions over U(N) characters. The procedure to obtain the corresponding U(N/M) character expansions is also described. Several examples are worked out in both cases.

PACS numbers: 11.30.Pb, 11.30.Ly, 21.60.Fw

I. INTRODUCTION

Character expansion techniques have been widely utilized in both particle and nuclear physics. In the context of particle physics, they have been used to evaluate certain U(N) group integrals[1,2] appearing in lattice gauge theories and the lattice version of the U(N)×U(N) sigma model, to investigate the role of internal symmetries in proton–antiproton annihilation,[3] to study phase transitions in hadronic matter with internal symmetries,[4] and to impose the colorlessness condition in the calculation of a quark-gluon gas.[5] In the context of nuclear physics they have been used to calculate the partition function[6] associated with a unitary ensemble[7] of random scattering matrices in a statistical theory of nuclear reactions, and to calculate the density of levels of a spherical nucleus as a function of angular momentum.[8]

The purpose of this paper is to present a formula for the expansion of a product of arbitrary functions of the eigenvalues of the fundamental representation matrix of the U(N) group in terms of its characters, and to obtain a similar result for the U(N/M) supergroup. Previous results[1,2,6] are shown to be special cases of the present general formulas.

In Sec. II the main result for the U(N) case is described. A number of examples for this case, which are potentially useful in physical applications, are explicitly worked out in Sec. III.

In Sec. IV, we investigate how similar results can be obtained for expansion in terms of U(N/M) supergroup characters and give some examples. Finally Sec. V contains a brief discussion of the results.

II. U(N) CHARACTER EXPANSION

In the rest of this paper the notation det (A_{ij}) denotes the determinant of the matrix whose ijth element is A_{ij}. The representations of the group U(N) are labeled by a partition into N parts: $(n_1,n_2,\ldots,n_N)$, where $n_1 \geqslant n_2 \geqslant \ldots \geqslant n_N \geqslant 0$ (see Ref. 9). The character of the irreducible representation corresponding to the partition $(n_1,n_2,\ldots,n_N)$ is given by Weyl's formula[10]

$$\chi_{(n_1,n_2,\ldots,n_N)}(U) = \frac{\det(t_i^{n_j+N-j})}{\Delta\,(t_1,\ldots,t_N)}, \tag{2.1}$$

where $t_i, i = 1,\ldots,N$, are the eigenvalues of the group element U in the fundamental representation, and $\Delta\,(t_1,t_2,\ldots,t_N)$ is the Vandermonde determinant in the arguments $t_1,\ldots,t_N$:

$$\Delta\,(t_1,t_2,\ldots,t_N) = \det(t_j^{N-i}). \tag{2.2}$$

Equation (2.1) can be written as

$$\chi_{(n_1,n_2,\ldots,n_N)}(U) = \det(h_{n_i+i-j}), \tag{2.3}$$

where h_n is the complete homogeneous symmetric function in the arguments $t_1,\ldots,t_N$ of degree n. One can write h_n in terms of the group element U as[11]

$$h_n = \frac{1}{2\pi i}\oint_C \frac{dz}{z^{n+1}}\,\frac{1}{\det(1-zU)}, \tag{2.4}$$

where C is a contour around the origin in the complex plane.

Let us consider the following power series expansion:

$$G(x,t) = \sum_{n=0}^{\infty} A_n(x)\,t^n, \tag{2.5}$$

where $A_n = 0$ when $n<0$, and $A_0 = 1$, and the series (2.5) is convergent for $|t| = 1$. Except for these restrictions the function $G(x,t)$ is completely *arbitrary*. Next, given N different t's:$t_1,t_2,\ldots,t_N$, we calculate the expression

$$\begin{aligned} \Delta\,(t_1,\ldots,t_N)\Big(\prod_{i=1}^{N} G(x,t_i)\Big) &= \Big[\prod_{i=1}^{N}\Big(\sum_{m=0}^{\infty} A_m(x)t_i^m\Big)\Big]\det\,(t_j^{N-k}) \\ &= \det|f_i(t_j)|, \end{aligned} \tag{2.6}$$

where the functions $f_i(t)$ are defined to be

$$f_i(t) \equiv \sum_{m=0}^{\infty} A_m(x)t^{m+N-i}. \tag{2.7}$$

Using the restriction stated above, one can rewrite Eq. (2.7) as

$$f_i(t) = \sum^{\infty} A_{k-N+i}(x)t^k. \tag{2.8}$$

Using Eq. (2.8) and by manipulating the properties of determinants, one can calculate[12]

$$\det|f_i(t_j)| = \sum_{l_1>l_2>\ldots>l_n\geqslant 0}\{\det\,[A_{n_j+i-j}(x)]\}\,[\det(t_i^{l_j})]. \tag{2.9}$$

Defining $n_i \equiv l_i + i - N$, the sum above can be rewritten as

$$\det|f_i(t_j)| = \sum_{n_1\geqslant n_2\geqslant\ldots\geqslant n_N\geqslant 0}\{\det\,[A_{n_j+i-j}(x)]\}\times[\det(t_i^{n_j+N-j})]. \tag{2.10}$$

Finally putting together Eqs. (2.6) and (2.10) one obtains

$$\Big(\prod_{i=1}^{N} G(x,t_i)\Big) = \sum_{n_1\geqslant n_2\geqslant\ldots\geqslant n_N\geqslant 0}[\det\,(A_{n_j+i-j})]\times\Big(\frac{\det(t_i^{n_j+N-j})}{\Delta\,(t_1,\ldots,t_N)}\Big). \tag{2.11}$$

Equation (2.11) is valid for *any* set of t_i's. Taking $t_i, i = 1,...,N$, to be the eigenvalues of the fundamental representation matrix U of U(N), and using Weyl's formula, Eq. (2.1), we get the following character expansion:

$$\left(\prod_{i=1}^{N} G(x,t_i)\right) = \sum_{n_1 \geq n_2 \geq ... \geq n_N \geq 0} \{\det [A_{n_j+i-j}(x)]\} \times \chi_{(n_1,...,n_N)}(U). \tag{2.12}$$

Note that the summation is over all irreducible representations of U(N).

III. EXAMPLES FOR U(N) CHARACTER EXPANSIONS

In the previous section, the generating function $G(x,t)$ was taken to be very general, only subject to certain restrictions. We not specify the form of $G(x,t)$ to obtain particular character expansions. Obviously the applicability of Eq. (2.12) is not limited to these examples. In this section r denotes any representation $(n_1,n_2,...,n_N)$ of U(N).

A. Example 1

Let $G(x,t) = \exp(xt)$. Then $A_n = x^n/n!$ for $n \geq 0$ and $A_n = 0$ for $n < 0$. Using the fact that the matrix U can always be diagonalized by a unitary transformation, Eq. (2.12) yields

$$\exp(x \operatorname{Tr} U) = \sum_r \alpha_r(x) \chi_r(U), \tag{3.1a}$$

where

$$\alpha_r(x) = \det\left(\frac{x^{n_j+i-j}}{(n_j+i-j)!}\right) = x^{n_1+n_2+...+n_N} \times \begin{vmatrix} \frac{1}{n_1!} & \frac{1}{(n_2-1)!} & \frac{1}{(n_3-2)!} & \cdots \\ \frac{1}{(n_1+1)!} & \frac{1}{n_2!} & \frac{1}{(n_3-1)!} & \cdots \\ \frac{1}{(n_1+2)!} & \frac{1}{(n_2+1)!} & \frac{1}{n_3!} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{vmatrix}. \tag{3.1b}$$

The same results have been obtained in Refs. 1, 2, and 6 by explicit integration over the group manifold. Our result follows from Eq. (2.12) immediately.

B. Example 2

Let $G(x,t)$ be the generating function for the Hermite polynomials[13]:

$$G(x,t) = \exp(2tx - t^2) = \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} t^n. \tag{3.2}$$

Then $A_n = [H_n(x)/n!]$ for $n \geq 0$ and $A_n = 0$ for $n < 0$. Equation (2.12) gives

$$\exp(2x \operatorname{Tr} U - \operatorname{Tr} U^2) = \sum_r \beta_r(x) \chi_r(U), \tag{3.3a}$$

where

$$\beta_r(x) = \begin{vmatrix} [H_{n_1}(x)/n_1!] & [H_{n_2-1}(x)/(n_2-1)!] & \cdots \\ [H_{n_1+1}(x)/(n_1+1)!] & [H_{n_2}(x)/n_2!] & \cdots \\ [H_{n_1+2}(x)/(n_1+2)!] & [H_{n_2+1}(x)/(n_2+1)!] & \cdots \\ \vdots & \vdots & \ddots \end{vmatrix} \tag{3.3b}$$

IV. U(N/M) CHARACTER EXPANSION

The representation theory of the unitary supergroup U(N/M) has recently been investigated by many authors.[11,14] In this paper we need only class I representations constructed from covariant bases. We denote a particular covariant class I representation of U(N/M) by $(n_1,n_2,...)$, where n_i is the number of boxes in the ith row of the corresponding supertableau.[11] The character of this representation is given by[11]

$$\chi_{(n_1,n_2,...)} = \det(\overline{H}_{n_i+i-j}), \tag{4.1}$$

where $\overline{H}_n$, the graded homogenous symmetric function of order n, can be written in terms of the group element $\mathscr{U}$ in the fundamental representation as[11]

$$\overline{H}_n(\mathscr{U}) = \frac{1}{2\pi i} \oint_C \frac{dz}{z^{n+1}} \frac{1}{\operatorname{Sdet}(1 - z\mathscr{U})}, \tag{4.2}$$

with C being a contour around the origin. Note that Eqs. (4.1) and (4.2) are formally identical to Eqs. (2.3) and (2.4), except that the determinants (and traces) are replaced by superdeterminants (and supertraces).

We now go back to the U(N) case. Let us, for example, consider the character expansion given in Eq. (3.1). Instead of expanding in terms of characters, we can first expand $\exp(x \operatorname{Tr} U)$ in terms of powers of Tr U, and then calculate $(\operatorname{Tr} U)^n$ in terms of characters, using the fact that $\chi_{(1,0,..)} = \operatorname{Tr} U$ and Eq. (2.3). Because the expansion (3.1) is *linear* in characters, the powers of Tr U should also be expressed as a linear combination of characters. Furthermore, since the irreducible characters of U(N) are linearly independent, such an expression is unique, and the final result is the same as Eq. (3.1). For illustration, a list of some lower powers of Tr U is given in Table I. As one can also see from this table, the term $(\operatorname{Tr} U)^n$ contains only the characters of the representations corresponding to Young tableaux with n boxes, and conversely those characters are the only ones included in the expansion of $\operatorname{Tr}(U)^n$.

Suppose we want to expand $\exp(x \operatorname{Str} \mathscr{U})$ in terms of the characters of the supergroup U(N/M), Str $\mathscr{U}$ being the supertrace of the fundamental representation matrix $\mathscr{U}$. We again first expand $\exp(x \operatorname{Str} \mathscr{U})$, in terms of powers of Str $\mathscr{U}$, and then calculate $(\operatorname{Str} \mathscr{U})^n$ in terms of characters of the supergroup, using the fact that $\chi_{(1,0,...)} = \operatorname{Str} \mathscr{U}$ and Eq. (4.1). The expressions we get for the powers of Str $\mathscr{U}$ are formally identical to those in Table I, except that the covariant characters of U(N/M)[11] are substituted for the characters of U(N). One can then write the expansion

$$\exp(x \operatorname{Str} \mathscr{U}) = \sum_s \alpha_s(x) \chi_s(\mathscr{U}). \tag{4.3a}$$

where $\alpha_s(x)$ is the same as Eq. (3.1b):

TABLE I. A list of some lower powers of Tr U as a linear combination of characters.

$$\mathrm{Tr}\,U = \chi_{(1,0,0,\ldots)}$$
$$\mathrm{Tr}\,U^2 = \chi_{(2,0,0,\ldots)} + \chi_{(1,1,0,\ldots)}$$
$$\mathrm{Tr}\,U^3 = \chi_{(3,0,0,\ldots)} + \chi_{(2,1,0,\ldots)} + \chi_{(1,1,1,0,\ldots)}$$
$$\mathrm{Tr}\,U^4 = \chi_{(4,0,0,\ldots)} + 3(\chi_{(3,1,0,\ldots)} + \chi_{(2,1,1,0,\ldots)}) + 2\chi_{(2,2,0,\ldots)} + \chi_{(1,1,1,1,0,\ldots)}$$

$$\alpha_s(x) = x^{n_1+n_2+\cdots} \det\left[1/(n_j+i-j)!\right]. \tag{4.3b}$$

In these formulas s denotes any representation $(n_1,n_2,\ldots)$ of $\mathrm{U}(N/M)$ constructed from covariant bases only and the sum is over *only* such representations: contravariant and mixed representations do *not* enter.

Starting from Eq. (3.3), similar arguments lead to the expansion

$$\exp(2x\,\mathrm{Str}\,\mathscr{U} - \mathrm{Str}\,\mathscr{U}^2) = \sum_s \beta_s(x)\chi_s(\mathscr{U}), \tag{4.4a}$$

where

$$\beta_s(x) = \det\left[H_{n_i+i-j}(x)/(n_i+i-j)!\right]. \tag{4.4b}$$

The procedure to obtain expansions over the supergroup $\mathrm{U}(N/M)$ characters is now straightforward. One first writes down the corresponding expression for the $\mathrm{U}(N)$ case using Eq. (2.12). Such an expression should be expressed in terms of the invariants (traces, determinants, characters) of the $\mathrm{U}(N)$ group. To obtain the associated expression for the supergroup, one then simply replaces them by the invariants (supertraces, superdeterminants, characters) of the $\mathrm{U}(N/M)$ and limits the summation only to the covariant characters. This result follows from the fact that Eq. (2.12) is essentially a combinatorial relation.

V. CONCLUSION

We have presented a formula to generate expansions over $\mathrm{U}(N)$ characters and described a procedure to obtain the corresponding $\mathrm{U}(N/M)$ supercharacter expansions starting from those of $\mathrm{U}(N)$. We have also worked out some examples for both cases. The coefficient of the character of a particular representation is the determinant of a given matrix, the dimension of which is equal to number of the rows in the corresponding (super)tableau. These coefficients cannot, in general, be put into a simpler form. Hence, in practice, our expansions might be difficult to use if one is interested in the coefficients of the characters corresponding to larger tableaux.

In studying systems with a particular symmetry one sometimes needs to calculate a specific quantity which transforms like a given representation of the symmetry group. However, in general, it is much easier first to compute the quantities which are invariant under this group, expand them in terms of the characters, and then project out the desired component. Our formulas are potentially very useful for such applications. In addition, with suitable choices of the functions $G(x,t)$ and the group element U, one can write down new generating functions for other invariants such as dimensions and eigenvalues of Casimir operators. Furthermore, character expansions can be used to obtain various combinatorial identities.

The expansions in terms of $\mathrm{U}(N/M)$ characters we considered in this paper include only covariant characters. It would be interesting to have expansions containing characters of contravariant and mixed representations as well. Work along this direction is in progress.

ACKNOWLEDGMENTS

The author thanks I. Bars and Y. Saint-Aubin for their comments on the manuscript.

This work is supported in part through funds provided by the U.S. Department of Energy (DOE) under Contract No. DE-AC02-76ERO3069.

[1] I. Bars, J. Math. Phys. **21**, 2678 (1980); S. Samuel, J. Math. Phys. **21**, 2695 (1980).
[2] I. Bars and F. Green, Phys. Rev. D **20**, 3311 (1979).
[3] B. Muller and J. Rafelski, Phys. Lett. B **116**, 274 (1982).
[4] K. Redlich and L. Turko, Z. Phys. C **5**, 201 (1980).
[5] M. I. Gorenstein, S. I. Lipskikh, V. K. Petrov, and G. M. Zinovjev, Phys. Lett. B **123**, 437 (1983).
[6] M. Gaudin and P. A. Mello, J. Phys. G **7**, 1085 (1981).
[7] F. J. Dyson, J. Math. Phys **3**, 140 (1962); M. L. Mehta, *Random Matrices* (Academic, New York, 1967).
[8] A. B. Balantekin and S. Levit (in preparation).
[9] R. Gilmore, *Lie Groups, Lie Algebras, and Some of Their Applications* (Wiley, New York, 1974); B. G. Wybourne, *Classical Groups for Physicists* (Wiley, New York 1974).
[10] H. Weyl, *The Classical Groups* (Princeton U. P., Princeton, NJ, 1948).
[11] A. B. Balantekin and I. Bars, J. Math. Phys. **22**, 1149 (1981); **22**, 1810 (1981); **23**, 1239 (1982).
[12] L. K. Hua, *Harmonic Analysis of Functions of Several Complex Variables in the Classical Domains* (AMS, Providence, RI, 1963), p. 22.
[13] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1972).
[14] V. G. Kac, Commun. Math. Phys. **53**, 31 (1977); Adv. Math. **26**, 8 (1977); "Representations of Classical Lie Superalgebras," in *Lecture Notes in Physics* No. 79, edited by P. Kramer and A. Rieckers (Springer-Verlag, Berlin, 1979); P. H. Dondi and P. D. Jarvis, J. Phys. A **14**, 547 (1981); Z. Phys. C **4**, 201 (1980); Y. Ne'eman and S. Sternberg, Proc. Natl. Acad. Sci. USA **77**, 3127 (1980); M. Scheunert, W. Nahm, and V. Rittenberg, J. Math. Phys. **18**, 155 (1977); I. Bars, *Proceeding of the School on Supersymmetry*, Mexico, 1981 (Plenum, New York, 1983); J-P. Hurni and B. Morel, J. Math. Phys. **24**, 157 (1983); J-Q. Chen, X-G. Chen, and M-J. Gao, J. Phys. A **16**, 1361 (1983); M. Scheunert, Bonn University preprints BONN-HE-82-24 and BONN-HE-82-26, (1982); I. Bars, B. Morel, and H. Ruegg, J. Math. Phys. **24**, 2253 (1983); F. Delduc and M. Gourdin, CERN Preprint TH-3568-CERN, 1983; A. B. Balantekin, J. Math. Phys. **23**, 486 (1982).

Supersymmetry and Clifford algebras

Z. Hasiewicz, A. K. Kwasniewski, and P. Morawiec
Institute of Theoretical Physics, University of Wroclaw, 50-205 Wroclaw, Poland

(Received 19 July 1983; accepted for publication 30 September 1983)

A canonical derivation of extended Lie superalgebras $a^N(p,q)$ right from the metric structure of the underlying "space-time" $E(p,q)$ is presented. The canonical method is that of Clifford algebras $C(p,q)$ and their representations theory. The results are tabularized.

PACS numbers: 11.30.Pb, 11.10.Qr

INTRODUCTION

1. Recently a development of the supersymmetric theories in dimensions higher than four has increased an interest of physicists in algebraic properties of spinors[1] and in supersymmetrization of quaternionic groups.[2] In all these considerations the Clifford algebras are looked upon from "Γ-matrices point of view." Such a treatment is based on Witt's decomposition, which gives a complex representation of Γ matrices only. The properties of these matrices as well as an eventual quaternionic structure are looked for among isomorphisms of real forms of the complex Lie algebras. This forces frequently separate considerations for each pseudoorthogonal metric.

2. One of the aims of the presented note is to derive, in a canonical way, the results of above papers and to compare them with Kac's classification of the Lie superalgebras.[3]

3. Most of the information on $\mathscr{C}(p,q)$ Clifford algebra representations is taken from Ref. 4, and this is the content of the first section.

4. In the second section the main goal of the paper is achieved: the extended Lie superalgebras $a^N(p,q)$ are constructed in a canonical way and natural conditions for the construction to be realized are found. The role played by spinors in extended superalgebras is made evident by the construction. Tabularized results are the form of comparing the method with what is already known about the subject. The construction is, at the same time, the way to understand the role of spinors in algebraic problems of supersymmetric theories in any dimension.

I. SOME PRELIMINARY INFORMATION ON $\mathscr{C}$(p,q) CLIFFORD ALGEBRAS REPRESENTATIONS[4-6]

1. Let $E(p,q)$ be $n=(p+q)$-dimensional, real vector space with the nondegenerate quadratic form:

$$E(p,q)\ni v\to Q(v):=v_1^2+\cdots+v_p^2-v_{p+1}^2-\cdots-v_{p+q}^2\in R,$$

and let $\mathscr{C}(p,q)$ denote its universal Clifford algebra with $\mathscr{C}^+(p,q)$ being an even subalgebra of $\mathscr{C}(p,q)$. Due to universality of $\mathscr{C}(p,q)$ an orthogonal transformation

$$E(p,q)\ni v\to -v\in E(p,q)$$

is uniquely extended to an automorphism inducing a Z_2 gradation of $\mathscr{C}(p,q)=\mathscr{C}^+(p,q)\oplus\mathscr{C}^-(p,q)$:

$$\mathscr{C}^\pm(p,q)\ni c\to\alpha(c)=\pm c\in\mathscr{C}^\pm(p,q)$$

and a map

$$\mathrm{Pin}(p,q)\ni s\to\pi_s\in\mathrm{O}(p,q),$$

where

$$E(p,q)\ni v\to\pi_s(v):=sv\alpha(s^{-1})\in E(p,q)$$

provides us with corresponding twofold coverings of O(p,q) and SO(p,q) groups, i.e., the sequences

$$1\to Z_2\to\mathrm{Pin}(p,q)\xrightarrow{\pi}\mathrm{O}(p,q)\to 1$$

$$1\to Z_2\to\mathrm{Spin}(p,q)\xrightarrow{\pi}\mathrm{SO}(p,q)\to 1$$

are exact.

The space of the fundamental representation of Pin(p,q) [or Spin (p,q)] group is then that of irreducible representation of $\mathscr{C}(p,q)$ [or $\mathscr{C}^+(p,q)$] algebra. Hence the construction of representation of twofold covering of O(p,q) [or SO(p,q)] group is achieved with the representation of $\mathscr{C}(p,q)$ [or $\mathscr{C}^+(p,q)$] algebra being given. For that purpose it is enough to consider $\mathscr{C}(p,q)$ Clifford algebras only as

$$\mathscr{C}^+(p,q)\sim\mathscr{C}(p,q-1),\quad q\geqslant 1,$$
$$\mathscr{C}(p,q)\sim\mathscr{C}^+(q,p),$$
$$\mathscr{C}^+(p,0)\sim\mathscr{C}^+(0,p),$$

and, of course, not all signatures need to be taken into account as

$$\mathscr{C}(p,q)\sim\mathscr{C}(p-4,q+4),\quad p\geqslant 4.$$

This isomorphism respects the Z_2 gradation of Clifford algebras. $\mathscr{C}(p,q)$ algebras are simple iff $p-q\neq 1$ (mod 4); otherwise, they are the direct sums of two two-sided ideals

$$\mathscr{C}(p,q)=\mathscr{C}(p,q)\tfrac{1}{2}(1+J)\oplus\mathscr{C}(p,q)\tfrac{1}{2}(1-J),$$

where $J=E_1\cdots E_n$ and $\{E_i\}_1^n$ is the canonical "(p,q)-orthonormal" basis of $E(p,q)\subset\mathscr{C}(p,q)$.

We know from Wedderburn's theorem[7] that any semisimple algebra with unity is isomorphic to the algebra of endomorphisms of an uniquely determined module over $F=R,C$, or H (quaternions). The representation module is an one-sided ideal of the algebra (we choose it to be left one). Clearly, an irreducible representation is obtained iff the ideal is minimal, i.e., iff it is generated by a primitive idempotent $e(p,q)$[8,4] [(A1)],

$$\mathscr{S}(p,q):=\mathscr{C}(p,q)e(p,q).$$

This left-sided ideal $\mathscr{C}(p,q)$ is a right module over the associative division ring $F(p,q):=e(p,q)\mathscr{C}(p,q)e(p,q)$ according to

$$\mathscr{S}(p,q)\times F(p,q)\ni(\psi,f)\to\psi f\in\mathscr{S}(p,q).$$

The modules $\mathscr{S}(p,q)$ and $\mathscr{S}^f(p,q)$ (defined below) are called

TABLE I. $C(p, q)$.

$p+q$ (mod 2) \ $p-q$ (mod 8)	0	1	2	3	4	5	6	7
0	$R(2n)$		$R(2n)$		$H(n)$		$H(n)$	
1		${}^2R(2n)$		$C(2n)$		${}^2H(n)$		$C(2n)$

with $n = 2^{[(p+q)/2]-1}$

pinor modules whence their elements—pinors.

The irreducible representation τ of $\mathscr{C}(p,q)$ algebra is now given by

$$\begin{aligned}&\mathscr{C}(p,q) \ni c \to \tau_c \in \mathrm{End}(\mathscr{S}(p,q)),\\ &\mathscr{S}(p,q) \ni \psi \to \tau_c(\psi) := c\psi \in \mathscr{S}(p,q),\end{aligned} \tag{I.1.1}$$

and τ is faithful iff $p-q \neq 1$ (mod 4) while for $p-q=1$ (mod 4) $\ker \tau = \mathscr{C}(p,q)\frac{1}{2}(1+J)$ or $\mathscr{C}(p,q)\frac{1}{2}(1-J)$. In this second case a faithful but reducible representation of $\mathscr{C}(p,q)$ is defined by endomorphisms of the direct sum

$$\begin{aligned}\mathscr{S}^f(p,q) &:= \mathscr{C}(p,q)e(p,q) \oplus \mathscr{C}(p,q)\alpha(e(p,q))\\ &\equiv \mathscr{S}(p,q) \oplus \alpha(\mathscr{S}(p,q))\end{aligned}$$

of α conjugate ideals.

2. A representation of Pin (p,q) is induced by that of $\mathscr{C}(p,q)$, namely,

$$\mathrm{Pin}(p,q) \ni s \to \tilde{\tau}_s \in \mathrm{Aut}(\mathscr{S}(p,q)),$$
$$\mathscr{S}(p,q) \ni \psi \to \tilde{\tau}_s(\psi) := s\psi \in \mathscr{S}(p,q).$$

One readily sees that

$$\ker \tilde{\tau} \sim \begin{cases} 1, & p-q \neq 1 \quad (\text{mod } 4),\\ Z_2, & p-q=1 \quad (\text{mod } 4).\end{cases}$$

In the case of $\mathscr{C}(p,q)$ being not simple ($\tilde{\tau}$ not faithful), one can introduce faithful (though reducible) representation induced on $\mathscr{S}^f(p,q)$.

3. We now proceed to construct a representation of Spin(p,q) on the space of representation of $\mathscr{C}(p,q-1)$ using an isomorphism

$$h: \mathscr{C}^+(p,q) \to \mathscr{C}(p,q-1),$$

which is explicitly given in (A2). The irreducible representation

$$\tau: \mathscr{C}(p,q-1) \to \mathrm{End}(\mathscr{S}(p,q-1))$$

induces a representation $\tilde{\tau}$ of Spin(p,q) group in Aut$(\mathscr{S}(p,q-1))$:

$$\mathrm{Spin}(p,q) \ni s \to \tilde{\tau}_s \in \mathrm{Aut}(\mathscr{S}(p,q-1)),$$
$$\mathscr{S}(p,q-1) \ni \psi \to \tilde{\tau}_s(\psi) := h(s)\psi \in \mathscr{S}(p,q-1),$$

$\tilde{\tau}$ being also irreducible.

Note: The case of $q=0$ could also be included due to the isomorphism Spin$(p,0) \sim$ Spin$(0,p)$.

With the irreducible representation $\tilde{\tau}$ of Spin(p,q) being now given, one finds for its kernel, depending on (p,q) signature, the following:

$$\ker \tilde{\tau} \sim \begin{cases} 1, & p-q \neq 0 \ (\text{mod } 4)\\ \mathrm{USp}(2), & (p-q) = (0,4),(4,0)\\ Sp(2,R), & (p,q) = (2,2)\\ Z_2, & \text{otherwise.}\end{cases}$$

Note: The kernels for $p+q=4$ are "that big" as a consequence of self-duality of $\mathscr{C}^{(2)}(p,q)$ subspace of bivectors which generate the Lie algebra of SO(p,q) (this only happens for $p+q=4$). The Lie algebras of ker $\tilde{\tau}$ are generated by either self-dual or anti-self-dual bivectors.

The faithful representation of Spin(p,q), $p-q=0$ (mod 4), is that induced by the faithful representation of $\mathscr{C}(p,q-1)$ with the representation module $\mathscr{S}^f(p,q-1)$. The modules of *irreducible* representations of spin groups are called spinor modules, whence their elements—spinors.

Note: In the important case of Spin(4,0), the two component quaternionic spinor decomposes into a sum of chiral and antichiral parts, both carrying the representation of USp(2) group. This is, however, not true for four-dimensional Euclidean pinors as the $S(4,0)$ pinor space does not decompose into two chiral subspaces because of noncommutativity of chiral projectors with reflections.

4. Matrix representations of $C(p,q)$ are provided by Table I,[4,5] which we include for completeness.

5. Let $\beta_\pm$ denote the antiautomorphisms of $\mathscr{C}(p,q)$ uniquely extended, due to universality, from $E(p,q)$, where they are defined by

$$E(p,q) \ni v \to \beta_\pm(v) := \pm v \in E(p,q).$$

In the following we are going to use these canonical antiautomorphisms $\beta (\equiv \beta_\pm)$ for a construction of $F(p,q)$ valued semilinear forms on $\mathscr{S}(p,q)$ as well as spinor conjugations, i.e., certain transformations of $\mathscr{S}(p,q)$ module onto $\mathscr{S}(p,q)_R$ module (right ideal).[4,5] The construction relies on the existence of an invertible element $\omega_\pm \in \mathscr{C}(p,q)$, satisfying

$$\omega_\pm \beta_\pm (e(p,q))\omega_\pm^{-1} = e(p,q). \tag{I.5.1}$$

If $\mathscr{C}(p,q)$ is simple, then $\omega_\pm$'s always exist and could be found among canonical basis elements. In this case we have then transformations

$$\mathscr{S}(p,q) \ni \psi \to \psi^{(\pm)} := \omega_\pm \beta_\pm(\psi) \in \mathscr{S}^{(\pm)}(p,q), \tag{I.5.2}$$

transforming the left ideal into the right one [(s)pinor conjugations]. The transformations under consideration yield involutive anti-automorphisms of $F(p,q)$:

$$F(p,q) \ni f \to \omega_\pm \beta_\pm(f)\omega_\pm^{-1} =: f^{(\pm)} \in F(p,q) \tag{I.5.3}$$

and the maps $\Theta_\pm$

$$\mathscr{S}(p,q) \times \mathscr{S}(p,q) \ni (\psi,\psi') \to \Theta_\pm(\psi,\psi') := \psi^{(\pm)}\psi' \in F(p,q),$$

define two (different in general) semilinear forms.

6. Consider now the case of $\mathscr{C}(p,q)$ being not simple. Let $P_\pm$ be central projectors onto respective two-sided ideals of $\mathscr{C}(p,q)$, i.e.,

$$P_\pm = \tfrac{1}{2}(1 \pm J);$$

there exists such β's (either β_+ or β_-) that

$$\beta(P_\pm) = P_\pm$$

and for the second antiautomorphism, which is $\alpha\circ\beta$, we have

$$(\alpha\circ\beta)(P_\pm) = P_\mp .$$

For the β anti-automorphism there exists $\omega\in\mathscr{C}(p,q)$ satisfying (I.5.1) so that one can extend the spinor conjugation onto $\mathscr{S}^f(p,q)$ according to

$$\begin{aligned}\mathscr{S}^f(p,q)\ni\psi&\to\omega\beta(P_+\psi + P_-\psi)\\ &= \omega\beta(P_+\psi)P_+ + \omega\beta(P_-\psi)P_-\in\mathscr{S}^f(p,q)_R .\end{aligned}$$

The above conjugation respects the direct sum structure of $\mathscr{S}^f(p,q)$, and the form on $\mathscr{S}^f$ is the sum of the corresponding forms on subspaces of $\mathscr{S}^f$. For $\alpha\circ\beta$, the spinor conjugation

$$\begin{aligned}\mathscr{S}^f(p,q)\ni\psi&\to\omega(\alpha\circ\beta)(P_-\psi)P_+\\ &+\omega(\alpha\circ\beta)(P_+\psi)P_-\in\mathscr{S}^f(p,q)_R\end{aligned}$$

is no more Z_2-graded transformation. Nevertheless, a semilinear form on $\mathscr{S}^f(p,q)$ may be also introduced.

7. Consider now the $\Theta_\pm$ form preserving groups

$$G_\pm(p,q) = \{g\in\mathscr{C}^*(p,q);\Theta_\pm(\psi,\psi') = \Theta_\pm(g\psi,g\psi')\},$$

where $\mathscr{C}^*(p,q)$ denotes the group of all invertible elements from $\mathscr{C}(p,q)$. These groups are described in Refs. 4 and 5. One can prove that $G_\pm(p,q)$ groups are the only form-preserving groups (that is to say, "metric") which contain the connected component of identity of Spin(p,q) group; hence $G_+(p,q)\cap G_-(p,q)$ group is the minimal "metric" one that does so.

Note, also, that if we consider the image of the connected component of identity of Spin(p,q) in the $\mathscr{C}(p,q-1)$ by the isomorphism h [(A2)], then it is contained only in $G_-(p,q-1)$ group.

8. The forms $\Theta_\pm$ have the definite symmetry with respect to the field conjugations (I.5.3)

$$\begin{aligned}(\Theta_\pm(\psi,\psi'))^{(\pm)} &= \omega_\pm\beta_\pm(\omega_\pm\beta_\pm(\psi)\psi')\omega_\pm^{-1}\\ &= \Theta_\pm(\psi',\psi)\beta_\pm(\omega_\pm)\omega_\pm^{-1}\\ &= :\Theta_\pm(\psi',\psi)\mathrm{sgn}(\omega_\pm).\end{aligned}$$

In the cases of nontrivial ($\neq$ identity) field conjugations it is always possible to redefine the (s)pinor conjugation (I.5.2) by composing it with the left multiplication by some nonzero $F(p,q)$ element [imaginary in a sense of (I.5.3)], thus obtaining a form with the opposite symmetry. The groups $G_\pm(p,q)$ remain unchanged. It is therefore natural and sufficient to consider only a class of conjugations yielding the $\Theta_\pm$ forms with fixed symmetry. We choose the class generating antisymmetric forms, and only this class will be used later on. This enables us to define the symbol

$$\epsilon_\pm := \begin{cases} -1, & \Theta_\pm \text{ is antisymmetric w.r.t. (I.5.3)},\\ +1, & \Theta_\pm \text{ is symmetric w.r.t. (I.5.3)},\end{cases}$$

which depend on the signature (p,q) only. Note that the value $+1$ is possible if field conjugation is trivial.

9. Exactness of the representation (I.1.1) leads to the following decomposition of Clifford algebra:

$$\mathscr{C}(p,q) = \mathscr{S}^f(p,q)\mathscr{S}^{f(\pm)}(p,q) = \mathscr{S}^f(p,q)F(p,q)\mathscr{S}^{f(\pm)}(p,q),$$

thus for arbitrary element we can write

$$\mathscr{C}(p,q)\ni c = u_\alpha c^{\alpha\beta}u_\beta^{(\pm)},\quad c^{\alpha\beta}\in F(p,q), \tag{I.9.1}$$

$\{u_\alpha\}$ and $\{u_\alpha^{(\pm)}\}$ are the bases of $\mathscr{S}^f(p,q)$ and $\mathscr{S}^{f(\pm)}(p,q)$, respectively; $1\leqslant\alpha,\beta\leqslant\dim_F\mathscr{S}^f(p,q)$. On the other hand,

$$\mathscr{C}(p,q)\ni c = z^J E_J,$$

with $z^J\in R$ and $\{E_J\}$ being the canonical basis of $\mathscr{C}(p,q)$. For z^J as a function of spin tensor (I.9.1), one easily obtains

$$z^J = \left(\frac{1}{m}\right)\mathrm{Re}(c^{\alpha\beta}E^{-1}_{J\beta\alpha})$$

with $m = \dim_F\mathscr{S}^f(p,q)$, and for spin tensor being simple, i.e., $c = \psi\phi^{(\pm)}$ one has

$$(\psi\phi^{(\pm)})^J = \left(\frac{1}{m}\right)\mathrm{Re}[\Theta_\pm(\phi,E_J^{-1}\psi)] = :z^J(\psi,\phi). \tag{I.9.2}$$

Note also that for simple spin tensor we have

$$\beta_\pm(\psi\phi^{(\pm)}) = \epsilon_\pm\phi\psi^{(\pm)} \tag{I.9.3}$$

in accordance with subsection I.8.

II. DERIVATION OF CANNONICAL LIE/SUPERALGEBRAS

Now we come over to the main subject of this paper; i.e., we are going to construct a class of Lie superalgebras canonically derived from $E(p,q)$ and its Clifford algebra $\mathscr{C}(p,q)$.

Since now on $\mathscr{S}(p,q)$ denotes the module of *faithful* representation.

1. Consider a naturally Z_2-graded vector space

$$V = \{F(p,q)\oplus\mathscr{C}(p,q)\}\oplus\{\mathscr{S}(p,q)\oplus\mathscr{S}^{(\pm)}(p,q)\} = :V_0\oplus V_1.$$

V is not a Z_2-graded algebra under Clifford algebra multiplication; however, it could be made to be so after vector space isomorphism identification of V with $\mathscr{S}\mathscr{C}(p,q)$, which is the algebra of 2×2 matrices with Clifford algebra entries and of the form

$$\begin{aligned}\mathscr{S}\mathscr{C}(p,q) :&= \{\hat f + \hat c + \hat\psi(\psi_1,\psi_2):\\ &= f\otimes I_0 + c\otimes I_2 + \psi_1\otimes I_1 + \psi_2^{(\pm)}\otimes I_3;\\ &f\in F(p,q),\ c\in\mathscr{C}(p,q),\ \psi_1\in\mathscr{S}(p,q),\\ &\psi_2^{(\pm)}\in\mathscr{S}^{(\pm)}(p,q)\},\end{aligned} \tag{II.1.1}$$

where $I_0 = \left(\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right)$, $I_1 = \left(\begin{smallmatrix}0&0\\1&0\end{smallmatrix}\right)$, $I_2 = \left(\begin{smallmatrix}0&0\\0&1\end{smallmatrix}\right)$, $I_3 = \left(\begin{smallmatrix}0&1\\0&0\end{smallmatrix}\right)$. The Z_2 grading is provided by

$$\begin{aligned}\mathscr{S}\mathscr{C}(p,q) &= \mathscr{S}\mathscr{C}_{(0)}(p,q)\oplus\mathscr{S}\mathscr{C}_{(1)}(p,q)\\ &:= \{(F(p,q)\otimes I_0)\oplus(\mathscr{C}(p,q)\otimes I_2)\}\\ &\quad\oplus\{(\mathscr{S}(p,q)\otimes I_1)\oplus(\mathscr{S}^{(\pm)}(p,q)\otimes I_3)\}\\ &:= \{\hat F(p,q)\oplus\hat{\mathscr{C}}(p,q)\}\oplus\{\hat{\mathscr{S}}(p,q)\oplus\hat{\mathscr{S}}^{(\pm)}(p,q)\}\\ &\equiv\begin{pmatrix}F(p,q) & \mathscr{S}^{(\pm)}(p,q)\\ \mathscr{S}(p,q) & \mathscr{C}(p,q)\end{pmatrix}.\end{aligned} \tag{II.1.2}$$

$\mathscr{S}\mathscr{C}(p,q)$ is an associative Z_2-graded algebra over *reals* under straightforward multiplication of direct sums and tensor product of algebras. Note that in addition $\hat{\mathscr{S}}\mathscr{C}_{(1)}(p,q)$ carries a linear structure over the division ring $\hat F(p,q)\sim F(p,q)$ with the multiplication "$*$" by scalars being defined according to

$$\mathscr{S}\mathscr{C}_{(1)}(p,q)\times\hat F(p,q)\ni(\hat\psi,\hat f)\to\hat\psi * f := [\hat\psi\hat f]\in\mathscr{S}\mathscr{C}_{(1)}(p,q). \tag{II.1.3}$$

This module structure cannot be extended to $\mathscr{S}\mathscr{C}(p,q)$ as $\hat F(p,q)\hat{\mathscr{C}}(p,q) = 0$. The Lie superalgebra associated with $\mathscr{S}\mathscr{C}(p,q)$, Z_2-graded algebra (i.e., algebra with

$$[a,b\} = ab - (-1)^{\deg(a)\deg(b)}ba)$$

$$\{\hat{\psi}(\psi_1,\psi_2),\hat{\Phi}(\phi_1,\phi_2)\} = (\psi_1\phi_2^{(\pm)} + \phi_1\psi_2^{(\pm)})\widehat{} + (\phi_2^{(\pm)}\psi_1 + \psi_2^{(\pm)}\phi_1)\widehat{} \in \widehat{\mathscr{C}}(p,q) \oplus \widehat{F}(p,q),$$

$$[\hat{c},\hat{\psi}(\psi_1,\psi_2)] = (c\psi_1) \,\hat{-}\, (\psi_2^{(\pm)}c)\widehat{}\in \mathscr{S}\mathscr{C}_{(1)}(p,q),$$

$$[\hat{f},\hat{\psi}(\psi_1,\psi_2)] = -(\psi f) \,\hat{+}\, (f\psi_2^{(\pm)})\widehat{}\in \mathscr{S}\mathscr{C}_{(1)}(p,q),$$

$$[\hat{c},\hat{c}'] = [c,c']\widehat{}\in\widehat{\mathscr{C}}(p,q), \tag{II.1.4}$$

$$[\hat{f},\hat{f}'] = [f,f']\widehat{}\in\widehat{F}(p,q),$$

$$[\hat{c},\hat{f}] = 0$$

is our starting point for the construction of Lie superalgebras whose even part contain SO(p,q) Lie algebras in a distinguished way. For that purpose note

$$\{\mathscr{S}\mathscr{C}_{(1)}(p,q),\mathscr{S}\mathscr{C}_{(1)}(p,q)\} \sim \mathscr{S}\mathscr{C}_{(0)}(p,q) \sim \begin{cases} \mathrm{gl}(1,\mathbb{F}) \oplus \mathrm{gl}(m;\mathbb{F}) & (p,q) \neq 1 \pmod 4 \\ {}^2\mathrm{gl}(1;\mathbb{F}) \oplus {}^2\mathrm{gl}(m/2;\mathbb{F}) & (p,q) = 1 \pmod 4, \end{cases} \tag{II.1.5}$$

where $m = \dim_F \mathscr{S}^f(p,q)$ (Table I), and this is due to the fact that simple spin tensors generate $\mathscr{C}(p,q)$ algebra. Thus $\mathscr{S}\mathscr{C}(p,q) \sim \mathrm{l}(1;m,F)$ or $\sim {}^2\mathrm{l}(1;m/2,F)$ in the notation of Ref. 3. In the canonical basis $\{E_J\}$ of $\mathscr{C}(p,q)$ the explicit form of (II.1.5) anticommutation can readily be obtained to be

$$\{\hat{\psi},\hat{\Phi}\} = Z^J(\hat{\psi},\hat{\Phi})\widehat{E}_J + \widehat{\Theta}_\pm(\hat{\psi},\hat{\Phi}), \tag{II.1.6}$$

where real coefficients Z^J of the right-hand side are bilinear functions of $\hat{\psi}$ and $\hat{\Phi}$ according to (I.9.2), and

$$\widehat{\Theta}_\pm(\hat{\psi},\hat{\Phi}) = \phi_2^{(\pm)}\psi_1 + \psi_2^{(\pm)}\phi_1.$$

Note that (II.1.4) induces adjoint representation of $\mathscr{C}^*(p,q)$ group on the module $\mathscr{S}\mathscr{C}_{(1)}(p,q)$

$$\mathscr{C}^{(*)}(p,q) \ni g \rightarrow \mathrm{Ad}\, g \in \mathrm{Aut}(\mathscr{S}\mathscr{C}_{(1)}(p,q)),$$

where (II.1.7)

$$\mathscr{S}\mathscr{C}_{(1)}(p,q) \ni \hat{\psi} \rightarrow \mathrm{Ad}\, \hat{g}\hat{\psi} := \hat{g}\hat{\psi}\hat{g}^{-1} \in \mathscr{S}\mathscr{C}_{(1)}(p,q).$$

2. We are now in a position to define canonical way of finding some important subalgebras of Lie superalgebras $\mathscr{S}\mathscr{C}(p,q)$. We are looking for superalgebras, the odd part of which is the minimal module of faithful representation of (S)Pin(p,q) group. We must then reduce the number of (s)pinorial generators as this minimal module has to be isomorphic to $\mathscr{S}(p,q)$. One can easily prove that the only admissible, and preserving exactness of the representation (I.2) projection, is generated by the involution

$$\mathscr{S}\mathscr{C}_{(1)}(p,q) \ni \hat{\psi}(\psi_1,\psi_2) \rightarrow j(\hat{\psi}(\psi_1,\psi_2)) := \hat{\psi}(\psi_2,\psi_1) \in \mathscr{S}\mathscr{C}_{(1)}(p,q), \tag{II.2.1}$$

which has only antiautomorphic extensions j^{ext} onto a whole $\mathscr{S}\mathscr{C}(p,q)$, namely, these generated by β_+ or β_- anti-automorphisms

$$\mathscr{S}\mathscr{C}(p,q) \ni \hat{\psi} + \hat{c} + \hat{f} \Rightarrow j^{\mathrm{ext}}(\hat{\psi} + \hat{c} + \hat{f}) = j(\hat{\psi}) + \beta_\pm(c) + \hat{f}^{(\pm)} \in \mathscr{S}\mathscr{C}(p,q), \tag{II.2.2}$$

according to whether $\mathscr{S}^{(+)}$ or $\mathscr{S}^{(-)}$ module is chosen in (II.1.1). One can now decompose $\mathscr{S}\mathscr{C}_{(1)}(p,q)$ onto a direct sum of two subspaces

$$\mathscr{S}^{(\frac{S}{A})}(p,q) = \tfrac{1}{2}(id \mp j)\mathscr{S}\mathscr{C}_{(1)}(p,q). \tag{II.2.3}$$

The condition for (s)pinorial generator to be (anti-) self-conjugate is the following:

$$j(\hat{\psi}(\psi_1,\psi_2)) = (-)\hat{\psi}(\psi_1,\psi_2), \tag{II.2.4}$$

which implies

$$\psi_1 = (-)\psi_2 \ \text{ for } \hat{\psi}(\psi_1,\psi_2) \in \mathscr{S}^{(A)S}(p,q).$$

One checks easily that neither $\mathscr{S}^{(S)}(p,q)$ nor $\mathscr{S}^{(A)}(p,q)$ carries the $\widehat{F}(p,q)$-linear structure in sense of (II.1.3).

The structure of anticommutators of self-conjugate and anti-self-conjugate generators is the same; more precisely, if

$$\mathscr{S}\mathscr{C}_{(1)} \ni \hat{\psi} = \hat{\psi}^S + \hat{\psi}^A$$

is the decomposition (II.2.3), then

$$\{\hat{\psi}^S,\hat{\Phi}^S\} = -\{\hat{\psi}^A,\hat{\Phi}^A\},$$

and both kinds may be equally used for contraction purposes. Note also that $\mathscr{S}^{(\frac{S}{A})}(p,q)$ subspaces are minimal modules of exact representations of (S)Pin(p,q) groups as they are isomorphic to $\mathscr{S}(p,q)$ due to (II.2.3).

The contraction procedure is now obvious; we rescale all the anti-self-conjugate generators by a parameter $\lambda \in R$:

$$\mathscr{S}\mathscr{C}_{(1)}(p,q) \ni \hat{\psi} \rightarrow \hat{\psi}_\lambda := \hat{\psi}^S + \lambda\hat{\psi}^A \in \mathscr{S}\mathscr{C}_{(1)}(p,q).$$

After anticommuting and taking the $\lambda \rightarrow 0$ limit

$$\{\hat{\psi}_\lambda,\hat{\Phi}_\lambda\} \xrightarrow{\lambda\rightarrow 0} \{\hat{\psi}^S,\hat{\Phi}^S\},$$

we obtain a resulting Z_2-graded vector space

$$a(p,q) = \mathscr{S}^S(p,q) \oplus \{\mathscr{S}^S(p,q),\mathscr{S}^S(p,q)\}, \tag{II.2.5}$$

which is the Lie superalgebra iff

$$j^{\mathrm{ext}}(\{\hat{\psi}^S,\hat{\Phi}^S\}) = -\{\hat{\psi}^S,\hat{\Phi}^S\},$$

due to the fact that the above property is the only one consistent with Lie algebra structure [defined by commutator on the even part of (II.2.5); see also I.8 and I.9)].

From (II.1.5) and (I.9.3) one easily deduces that

$$j^{\mathrm{ext}}(\{\hat{\psi}^S,\hat{\Phi}^S\}) = \epsilon_\pm\{\hat{\psi}^S,\hat{\Phi}^S\} \tag{II.2.6}$$

so the $a(p,q)$ space closes to the Lie superalgebra iff ϵ_+ or ϵ_- (or both) $= -1$. For this very case our ultimate goal is now achieved thanks to observation

$$\{\mathscr{S}^{(S)}(p,q),\mathscr{S}^{(S)}(p,q)\} = g_\pm(1;\mathbb{F}) \oplus g_\pm(p,q), \tag{II.2.7}$$

where $g_\pm(p,q)$ is the Lie algebra of $G_\pm(p,q)$ group (I.7) and $g_\pm(1,F)$ is the Lie algebra generated by purely imaginary (with respect to I.5.2) elements of the ring $F(p,q)$. More explicitly,

$$\{\hat{\psi}^S,\hat{\Phi}^S\} = (\psi\phi^{(\pm)} + \phi\psi^{(\pm)}) \,\hat{+}\, (\psi^{(\pm)}\phi + \phi^{(\pm)}\psi)$$

with properties

$$\beta_\pm(\{\hat{\psi}^S,\hat{\Phi}^S\}|_2) = \epsilon_\pm\{\hat{\psi}^S,\hat{\Phi}^S\}|_2 = -\{\hat{\psi}^S,\hat{\Phi}^S\}|_2, \tag{II.2.8}$$

$$\omega_\pm\beta_\pm(\{\hat{\psi}^S,\hat{\Phi}^S\}|_0)\omega_\pm^{-1} = -\{\hat{\psi}^S,\hat{\Phi}^S\}|_0.$$

$[\{\ ,\ \}|_i]$ means restricted to the block spanned by I_i; (II.1.1)]. For $\{\hat{\psi}^S,\hat{\Phi}^S\}$, in terms of canonical basis elements E_J, we then have

$$\{\hat{\psi}^S,\hat{\Phi}^S\} = Z^J(\phi,\psi)\widehat{E}_J + \widehat{\Theta}_\pm(\psi,\phi) \tag{II.2.8}$$

with only nonzero coefficients Z^J for E_J's satisfying

$$\beta_{\pm}(E_J) = -E_J$$

—the defining property of $g_{\pm}(p,q)$ Lie algebra generators. For the case of $\epsilon_{\pm}$ (II.2.5) equal to 1, the nontrivial reduction is impossible; if we want for $a_{(1)}(p,q)$ to be invariant with respect to $a_{(0)}(p,q)$, we have to contract $a_{(0)}(p,q)$ to the commutative algebra, which reduces it to zero in finite-dimensional representation we are working with. All the contracted Lie superalgebras are listed in Table II with $N = 1$.

For the contracted superalgebra $a(p,q)$ to be invariant under adjoint representation (II.1.8), it is necessary to restrict it to the subgroup $G_{\pm}(p,q)$, and then it has the form

$$\begin{aligned} a_{(1)}(p,q) \ni \hat{\psi} \to \mathrm{Ad}\, g(\hat{\psi}) &:= \hat{g}\hat{\psi}\hat{g}^{-1} \\ &= \hat{g}\hat{\psi}\beta_{\pm}(g) \in g_{(1)}(p,q), \\ & g \in G_{\pm}(p,q). \end{aligned} \tag{II.2.10}$$

Note that in general $G_{\pm}(p,q)$ groups do not contain whole Spin(p,q) group; however, they contain part of reflections, and this makes different the pinor and spinor structures on the superalgebra level. $a(p,q)$ algebras correspond to the so called *simple* ($N = 1$) supersymmetries for $E(p,q)$ space-times.

3. The superalgebras $a(p,q)$ can be extended by unique enlargement of the $g_{\pm}(1,F)$ [II.2.7] to $g_{\pm}(N,F)$ as we will see below. Let us consider N-dimensional right $F(p,q)$, module V, and let V^* be the left module, conjugate with respect to some semilinear form Ω. Introducing extended spinor modules

$$\mathscr{S}^N(p,q) := \mathscr{S}(p,q) \otimes V^*,$$
$$\mathscr{S}^{N(\pm)}(p,q) := V \otimes \mathscr{S}^{(\pm)}(p,q)$$

with obvious multiplications

$$(\psi \otimes v^*)(w \otimes \phi^{(\pm)}) := \psi(v^* w)\phi^{(\pm)} := \psi\Omega(v,w)\phi^{(\pm)},$$
$$(w \otimes \phi^{(\pm)})(\psi \otimes v^*) := w\Theta_{\pm}(\phi,\psi)v^* \in V \otimes V^* = \mathrm{End}(V),$$

one can construct $\mathscr{S}^N\mathscr{C}(p,q)$ algebras, which generalize (II.1.1), and associated superalgebras, ending up with

$$\mathscr{S}^N\mathscr{C}_{(0)}(p,q) \sim \begin{cases} \mathrm{gl}(N,\mathbb{F}) \oplus \mathrm{gl}(m,\mathbb{F}) & p - q \neq 1 \pmod 4 \\ {}^2\mathrm{gl}(N,\mathbb{F}) \oplus {}^2\mathrm{gl}(\tfrac{1}{2}m,\mathbb{F}) & p - q = 1 \pmod 4, \end{cases}$$

with arbitrary N [m as in (I.1.5)].

The involution (II.2.1) has obvious generalization, and one can introduce self-conjugate spinorial generators, analogous to that of (II.2.3). The consistency condition for self-conjugate generators built up from simple tensors

$$\beta_{\pm}(\{\hat{\psi}^S(\psi,v),\hat{\Phi}^S(\phi,w)\}|_2) = -\{\hat{\psi}^S(\psi,v),\hat{\Phi}^S(\phi,v)\}|_2$$

implies the "hermiticity" of Ω

$$\Omega^{(\pm)}(v,w) = \Omega(w,v)$$

with respect to (I.5.2) and this in turn gives $a^N_{(0)}(p,q)|_0$ being isomorphic to the Lie algebra of the group preserving the form Ω. Note that "hermiticity" condition fixes only the Cartan class of the above Lie algebra. The $a^N(p,q)$ superalgebras are listed in Table II.

TABLE II.[a]

	0	7	6	5	4	3	2	1	$d \pmod 8$
$q=0$ β_+	—	—	—	$U_\alpha U(N\|2n;C)$	$\mathrm{Usp}(N\|2n;H)^2$	$\mathrm{Usp}(N\|2n;H)$	$\mathrm{Usp}(N\|2n;H)$	$U_\alpha U(N\|2n;C)$	β_- $p=0$
$q=0$ β_-	—	—	$\mathrm{Usp}(N\|2n;R)$	$\mathrm{Usp}(N\|2n;C)$	$\mathrm{Usp}(N\|n,n;H)$	—	$U_\alpha U(N\|2n;H)$	—	β_+ $p=0$
$d \pmod 8$	0	1	2	3	4	5	6	7	
$d=p+q \pmod 8$: 0	—		—		$U_\alpha U(N\|2n;H)$		$U_\alpha U(N\|2n;H)$		
	—		—		$U_\alpha U(N\|2n;H)$		$U_\alpha U(N\|2n;H)$		
1		—		—		${}^2U_\alpha U(N\|2n;H)$		—	
		—		$U_\alpha U(N\|n,n;C)$		—		$U_\alpha U(N\|n,n;C)$	
2	—		$\mathrm{Usp}(N\|2n;R)$		$\mathrm{Usp}(N\|n,n;H)$		$U_\alpha U(N\|2n;H)$		
	$\mathrm{Usp}(N\|2n;R)$		$\mathrm{Usp}(N\|2n;R)$		$\mathrm{Usp}(N\|n,n;H)$		$\mathrm{Usp}(N\|n,n;H)$		
3		—		$U_\alpha U(N\|n,n;C)$		—		$U_\alpha U(N\|n,n;C)$	
		${}^2\mathrm{Osp}(N\|2n;R)$		$\mathrm{Osp}(N\|2n;C)$		${}^2\mathrm{Usp}(N\|n,n;H)$		$\mathrm{Osp}(N\|2n;C)$	
4	$\mathrm{Osp}(N\|2n;R)$		$\mathrm{Osp}(N\|2n;R)$		$\mathrm{Usp}(N\|n,n;H)$		$\mathrm{Usp}(N\|n,n;H)$		
	$\mathrm{Osp}(N\|2n;R)$		$\mathrm{Osp}(N\|2n;R)$		$\mathrm{Usp}(N\|n,n;H)$		$\mathrm{Usp}(N\|n,n;H)$		
5		${}^2\mathrm{Osp}(N\|2n;R)$		$\mathrm{Osp}(N\|2n;C)$		${}^2\mathrm{Usp}(N\|n,n;H)$		$\mathrm{Osp}(N\|2n;C)$	
		—		$U_\alpha U(N\|n,n;C)$		—		$U_\alpha U(N\|n,n;C)$	
6	$\mathrm{Osp}(N\|2n;R)$		$\mathrm{Osp}(N\|2n;R)$		$\mathrm{Usp}(N\|n,n;H)$		$\mathrm{Usp}(N\|n,n;H)$		
	—		—		$U_\alpha(N\|2n;H)$		$U_\alpha(N\|2n;H)$		
7		—		$U_\alpha U(N\|n,n;C)$		—		$U_\alpha U(N\|n,n;C)$	
		—		—		${}^2U_\alpha U(N\|2n;H)$		—	

$(p-q) \pmod 8$

[a] In Table II $n = \frac{1}{2}\dim_F S(p_1 q)$ and even parts are given by[9]

$$U_\alpha U(N|2n;F)|_0 = \begin{cases} \mathrm{U}(N) \oplus \mathrm{U}(2n), & F = C, \\ \mathrm{Sp}(N) \oplus \mathrm{SO}^*(4n), & F = H, \end{cases}$$

$$\mathrm{Osp}(N|2n;F)|_0 = \begin{cases} \mathrm{So}(N;C) \oplus \mathrm{Sp}(2n;C), & F = C, \\ \mathrm{SO}(N;R) \oplus \mathrm{Sp}(2n;R), & F = R, \end{cases}$$

$$U_\alpha(N|n,n;C)|_0 = \mathrm{U}(N) \oplus \mathrm{U}(n,n),$$

$$\mathrm{Usp}(N|n,n;H)|_0 = \mathrm{Sp}(N) \oplus \mathrm{Sp}(n,n).$$

III. FINAL REMARKS

1. An application of our approach to $a^N(6.2)$ superalgebra is now in preparation[10] and another interesting case, namely, that of $OSp(N,4)$ $[a^N(3.1)]$ is also being considered there.

2. Note that no particular choice of matrix representation of Clifford algebra generators was needed to obtaining the general commutation relations of the superalgebras considered above.

ACKNOWLEDGMENTS

We thank Professor J. Lukierski for a very stimulating discussion. We are also thankful to Professor H. Ruegg for his interest in this work and useful comments.

APPENDIX

1. Following Refs. 4–6 we write for the primitive idempotent $e(p,q)$

$$e(p,q) := \prod_J \tfrac{1}{2}(1 + E_J),$$

where the product is extended over the maximal set of pairwise commuting, unimodular elements from the canonical basis $\{E_J\}$ of $\mathscr{C}(p,q)$ algebra. Note that these elements which are products of others can be excluded from the product as they do not contribute nontrivially the projector $e(p,q)$.

2. The explicit formula for the isomorphism

$h: \mathscr{C}^+(p,q) \Rightarrow \mathscr{C}(p,q-1)$

is the following:

$$\begin{aligned}\mathscr{C}^+(p,q) \ni c^+ &\Rightarrow h(c^+)\\ &:= \tfrac{1}{2}(c^+ - E^n c^+ E^n) - \tfrac{1}{2}(c^+ + E^n c^+ E^n)E^n\\ &= \tfrac{1}{2}\{c^+, E^n\}E^n - \tfrac{1}{2}[c^+, E^n] \in \mathscr{C}(p,q-1),\end{aligned}$$

where the vector $E^n \in E(p,q)$ satisfies $E_n^2 = -1$ and $\mathscr{C}(p,q-1)$ is considered as the subalgebra of $\mathscr{C}(p,q)$.

[1]J. W. van Holten and A. Van Proeyen, J. Phys. **15**, 3763 (1982); R. Coquereaux, CERN Preprint TH 3279, April 1982; C. Wetterich, CERN Preprint TH 3310, May 1982; T. Kugo and P. K. Townsend, CERN Preprint TH 3459, November 1982.
[2]J. Lukierski and A. Nowicki, Fort Schr. Phys. **30**, 75 (1982); J. Lukierski and A. Nowicki, Preprint 34/82/E.P., International School of Advanced Studies (Trieste), June 1982.
[3]V. G. Kac, Adv. Math. **26**, 8 (1977).
[4]P. Lounesto, Report HTKK-MAT-A151, Helsinki University of Technology, 1979 (Ph.D.Thesis); see references therein, esp. Chevalley (1955), "The construction and study of certain important algebras."
[5]P. Lounesto, Found. Phys. **11**, 721 (1981).
[6]P. Lounesto, Ann. Inst. H. Poincaré **XXXIII**, 53, (1980).
[7]See, e.g., M. Marcus, *Finite Dimensional Multilinear Algebra* (Dekker, New York, 1975), Part II; W. Więsław, *Groups, Rings, Fields* (Wroclaw Univ. Press, Wrocław, 1977) (in Polish).
[8]P. Lounesto, Report HTKK-MAT-A113, Helsinki University of Technology, 1977.
[9]R. Gilmore, *Lie Groups, Lie Algebras and Some of Their Applications* (Wiley, New York, 1974).
[10]Z. Hasiewicz, J. Lukierski, and P. Morawiec Phys. Lett B **130**, 55 (1983).

Irreducible SO(2) extended superfields

Jaegu Kim[a)]
Department of Physics, University of California, Los Angeles, California 90024

(Received 7 September 1983; accepted for publication 23 December 1983)

All the irreducible SO(2) extended scalar superfields are obtained in SO(2)-invariant form. These superfields can be labelled by the eigenvalues of three Casimir operators of the enlarged algebra containing chiral transformation. They may be still more simply classified by introducing two invariant simple superspaces. Lagrangians and supersymmetry transformation rules are obtained from the explicit form of the irreducible superfields. The highest superspin multiplet is identified as the multiplet for the linearized SO(2) Weyl supergravity. The linearized SO(2) Poincaré supergravity is obtained by coupling this multiplet with the chiral superfields.

PACS numbers: 11.30.Pb

1. INTRODUCTION

Supersymmetry[1] permits a unified description of bosons and fermions, allows a nontrivial unification of space-time symmetry and internal symmetry, and is compatible with relativistic quantum field theory. Its extension to supergravity is less divergent in perturbation expansion than any other gravitational theory, and it is hoped, therefore, that this approach might lead us towards a renormalizable or finite theory which unifies all the fundamental forces in nature including gravity.

However, the structure of the supermultiplet is very complicated in general, and it is only completely known for $N=1$ and partly known for $N=2$. In addition to the physical fields, supermultiplets contain nonpropagating auxiliary fields which are needed to close the algebra off-shell. Closure of the algebra is essential to construct the tensor calculus, to get the Ward identities, and to quantize the theory in a manifest supersymmetric way.[2] The concept of superspace and superfield introduced by Salam and Strathdee[3] has turned out to be very useful to find the supermultiplet structures. Since the algebra is guaranteed to close in a superfield representation, one can automatically find the auxiliary fields which are needed to close the algebra. However, the superfields provide a basis for representation of an enlarged algebra which includes covariant derivatives. So, it is expected that the representation of the supersymmetry algebra in terms of general superfields is reducible. Besides, the number of independent field components in a real general SO(N) superfield is 2^{4N}, and this number increases very rapidly as N goes from 1 to 8. So, instead of working with general superfields, one must try to work with irreducible superfields.

In order to extract the irreducible superfields from a general superfield, one has to find a complete set of Casimir operators and their projection operators. Using these projection operators, one can decompose a general superfield into a sum of the irreducible superfields. This has been done for the $N=1$ case by Sokachev.[4] The method of obtaining the Lagrangian for the $N=1$ case is developed by Ogievetsky and Sokachev.[5] Taylor[6] extended their method to SO(2) and SO(4) extended supersymmetry and found all the projection operators. In this paper, we follow the same path but in an SO(2)-invariant way and find all the irreducible superfields in manifest SO(2)-invariant form. Lagrangians and supersymmetry transformation rules are obtained from these irreducible superfields.

In Sec. 2, we introduce the SO(N) extended supersymmetry algebra without central charges and its superspace representation. The Casimir operators are found, and the superspin decomposition is carried out in Sec. 3. In Sec. 4, we obtain the projection operators of the irreducible scalar superfields for $N=1$ and $N=2$. The SO(2) superspace is decomposed into two invariant simple superspaces, and simplified constraint equations for the irreducible superfields are found in Sec. 5. These superfields can be labelled by the eigenvalues of the three Casimir operators. In order to solve the constraint equations, we introduce a complete set of supertensors in Sec. 6. In Sec. 7, we obtain the explicit form of the irreducible superfields. It turns out that some of the superfields obtained from the constraint equations can be reduced further, and this introduces supersymmetric gauge invariance. The Lagrangian and the supersymmetry transformation rules for each multiplet are obtained. The highest superspin multiplet is identified as the linearized SO(2) Weyl supergravity multiplet.[7] By coupling this multiplet with the SO(2) tensor gauge multiplet and the SO(2) vector gauge multiplet, we obtained the SO(2) Poincaré supergravity[7] in Sec. 8. Notational conventions are given in Appendix A, and some supertensor reduction formulas are listed in Appendix B.

2. EXTENDED SUPERSYMMETRY ALGEBRA

The supersymmetry algebra is a graded extension of the Poincaré algebra. In addition to the generators of the Poincaré algebra (P_a, M_{ab}), the SO(N) extended supersymmetry algebra contains Majorana spinor charges $Q_{\alpha i}$, $\overline{Q}_{\dot{\alpha} i}$ and generators for SO(N) transformation T_l. They satisfy the following graded Lie algebra.[8]

$$(P_a, P_b) = 0,$$

$$(M_{ab}, P_c) = i(\eta_{bc} P_a - \eta_{ac} P_b),$$

[a)] Present address: Center for Relativity, Department of Physics, University of Texas, Austin, TX 78712.

 0022-2488/84/062037-15$02.50

$$
\begin{aligned}
&(M_{ab},M_{cd}) = i(\eta_{bc}M_{ad} - \eta_{ac}M_{bd} \\
&\qquad\qquad + \eta_{ad}M_{bc} - \eta_{bd}M_{ac}), \\
&(Q_{\alpha i},P_a) = (\overline{Q}_{\dot\alpha i},P_a) = 0, \\
&(T_l,P_a) = (T_l,M_{ab}) = 0, \\
&(Q_{\alpha i},\overline{Q}_{\dot\alpha j})_+ = -2\delta_{ij}\sigma^a_{\alpha\dot\alpha}P_a, \\
&(Q_{\alpha i},Q_{\beta j})_+ = 0, \\
&(M_{ab},Q_{\alpha i}) = -i(\sigma_{ab})_\alpha{}^\beta Q_{\beta i}, \\
&(M_{ab},\overline{Q}^{\dot\alpha i}) = -i(\bar\sigma_{ab})^{\dot\alpha}{}_{\dot\beta}\overline{Q}^{\dot\beta i}, \\
&(Q_{\alpha i},T_l) = (t_l)_{ij}Q_{\alpha j}, \\
&(\overline{Q}_{\dot\alpha i},T_l) = (t_l)_{ij}\overline{Q}_{\dot\alpha j}, \\
&(T_l,T_m) = iC_{lmn}T_n,
\end{aligned}
\tag{2.1}
$$

where $(t_l)_{ij}$ is a Hermitian representation of T_l and C_{lmn} are the structure constants of SO(N). This algebra can be realized in superspace as

$$
\begin{aligned}
&P_a = i\partial_a, \\
&M_{ab} = i(x_a\partial_b - x_b\partial_a) + \Sigma_{ab} \\
&\qquad + i(\sigma_{ab})_\alpha{}^\beta\theta^{\alpha i}\partial_{\beta i} + i(\bar\sigma_{ab})^{\dot\alpha}{}_{\dot\beta}\bar\theta_{\dot\alpha i}\bar\partial^{\dot\beta}, \\
&T_l = (t_l)_{ij}(\theta^{\alpha i}\partial_{\alpha j} + \bar\theta_{\dot\alpha i}\bar\partial^{\dot\alpha i}) + t_l, \\
&Q_{\alpha i} = \partial_{\alpha i} + i\sigma^a_{\alpha\dot\alpha}\bar\theta^{\dot\alpha i}\partial_a, \\
&\overline{Q}_{\dot\alpha i} = \bar\partial_{\dot\alpha i} - i\sigma^a_{\alpha\dot\alpha}\theta^{\alpha i}\partial_a.
\end{aligned}
\tag{2.2}
$$

In addition to the generators of the SO(N) extended supersymmetry algebra, superspace contains convariant derivatives which anticommute with the spinor charges. They satisfy a similar algebra among themselves:

$$
\begin{aligned}
&(D_{\alpha i},\overline{D}_{\dot\alpha j})_+ = 2\delta_{ij}\sigma^a_{\alpha\dot\alpha}P_a, \\
&(D_{\alpha i},D_{\beta j})_+ = 0.
\end{aligned}
\tag{2.3}
$$

The representation of these operators in superspace is

$$
\begin{aligned}
&D_{\alpha i} = \partial_\alpha - i\sigma^a_{\alpha\dot\alpha}\bar\theta^{\dot\alpha i}\partial_a, \\
&\overline{D}_{\dot\alpha i} = \bar\partial_{\dot\alpha} + i\sigma^a_{\alpha\dot\alpha}\theta^{\alpha i}\partial_a.
\end{aligned}
\tag{2.4}
$$

Due to these covariant derivatives, a general superfield satisfies an enlarged algebra which includes the covariant derivatives. So, it is expected that the representation of the supersymmetry algebra in terms of general superfields is reducible. In order to find the irreducible superfields, we will find Casimir operators of the algebra from which one can construct the projection operators for the irreducible superfields in the next section.

3. SUPERSPIN DECOMPOSITION

The supersymmetry algebra has two Casimir operators P^2 and C^2, where C^2 is square of the generalized Pauli–Lubanski vector[3]

$$C^2 = C^{ab}C_{ab}, \tag{3.1}$$

$$C_{ab} = P_aC_b - P_bC_a, \tag{3.2}$$

$$C_a = \tfrac{1}{2}\epsilon_{abcd}M^{bc}P^d - \tfrac{1}{4}Q^i\sigma_a\overline{Q}^i. \tag{3.3}$$

So, all the irreducible superfields can be labelled by the eigenvalues of P^2 and C^2.

In the rest frame, one has

$$
\begin{aligned}
&C_{ij} = 0, \quad i,j = 1,2,3, \\
&C_{0i} = mC_i, \\
&(C^i,C^j) = im\epsilon^{ijk}C^k.
\end{aligned}
\tag{3.4}
$$

Thus, C^i/m satisfies the algebra of angular momentum in the rest frame, and the eigenvalue of the Casimir operators C^2 is

$$C^2 = -2m^4Y(Y+1), \tag{3.5}$$

where Y is an integer or half-integer called superspin.[4]

Now, choose an irreducible superfield Φ_J with external spin J such that

$$
\begin{aligned}
&\overline{D}_{\dot\alpha j}\Phi_J = 0, \\
&C^2\Phi_J = -2m^2J(J+1)\Phi_J.
\end{aligned}
\tag{3.6}
$$

From (3.1), (3.2), and (3.3), one has

$$(C^2,D_{\alpha i}) = 4P^2P^b(\sigma_{ab})_\alpha{}^\beta D_{\beta i}C^a - \tfrac{3}{2}P^4D_{\alpha i}. \tag{3.7}$$

So

$$
\begin{aligned}
C^2D_{\alpha i}\Phi_J &= 2m^3D_{\beta i}(\mathbf{C}\cdot\boldsymbol{\sigma})_\alpha{}^\beta\Phi_J \\
&\quad - 2m^4\left[J(J+1) + \tfrac{3}{4}\right]D_{\alpha i}\Phi_J.
\end{aligned}
\tag{3.8}
$$

Since C^k/m satisfies the SU(2) algebra in the rest frame and since the eigenstates of $\boldsymbol{\sigma}\cdot\mathbf{C}/m$ are superspin $J\pm\frac{1}{2}$ states, $D_{\alpha i}$ raises or lowers the superspin of Φ_J by $\frac{1}{2}$. So, with proper Clebsch–Gordan coefficients, one can combine D_{1i} and D_{2i} such that it is either the superspin raising operator Y^i_+ or the superspin lowering operator Y^i_-. Since i goes from 1 to N, one can get superspin states $Y = J - N/2, J-(N-1)/2, \ldots, J+N/2$ by applying Y^i_+ and Y^i_- to Φ_J successively, if $J \geqslant N/2$. The multiplicity of the superspin $J-(N-k)/2$ state is $\binom{2N}{k}$.

By using the projection operator for C^2, one can separate all different superspin states but the states with the same superspin cannot be distinguished. In order to separate the degenerate superspin states, one has to introduce a new symmetry transformation Γ (chiral) into the algebra. It satisfies

$$
\begin{aligned}
&(Q_{\alpha i},\Gamma) = Q_{\alpha i}, \\
&(\overline{Q}_{\dot\alpha i},\Gamma) = -\overline{Q}_{\dot\alpha i}, \\
&(D_{\alpha i},\Gamma) = D_{\alpha i}, \\
&(\overline{D}_{\dot\alpha i},\Gamma) = -\overline{D}_{\dot\alpha i}, \\
&(\text{anything else}, \Gamma) = 0.
\end{aligned}
\tag{3.9}
$$

This enlarged algebra has a new Casimir operator for $N=1$:

$$G \equiv (P_a/2P^2)D\sigma^a\overline{D} - 1. \tag{3.10}$$

For $N=2$, one finds two more Casimir operators.

$$G \equiv (P_a/2P^2)D^i\sigma^a\overline{D}^i - 2, \tag{3.11}$$

$$E \equiv i\epsilon^{ij}(P_a/2P^2)D^i\sigma^a\overline{D}^j. \tag{3.12}$$

By using the projection operators of all the Casimir operators, we will obtain the projection operators for the irreducible superfields in the next section.

4. PROJECTION OPERATORS

A. $N=1$ supersymmetry

The generators M_{ab} for a scalar superfield are

$$M_{ab} = x_a P_b - x_b P_a + i(\sigma_{ab})_\alpha{}^\beta \theta^\alpha \partial_\beta + i(\bar\sigma_{ab})^{\dot\alpha}{}_{\dot\beta} \bar\theta_{\dot\alpha} \bar\partial^{\dot\beta}. \tag{4.1}$$

Substituting (4.1) into (3.2), one obtains

$$C_{ab} = -\tfrac{1}{4}(P_a D\sigma_b \bar D - P_b D\sigma_a \bar D). \tag{4.2}$$

Using (2.3), one finds

$$C^2 = -\tfrac{3}{2}P^4(1-S), \tag{4.3}$$

where

$$S = (1/16P^2)(D^2\bar D^2 + \bar D^2 D^2), \quad S^2 = S. \tag{4.4}$$

Since $S^2 = S$, C^2 satisfies the eigenvalue equation

$$C^2(C^2 + \tfrac{3}{2}P^4) = 0. \tag{4.5}$$

The eigenstates of C^2 correspond to the superspin states as follows:

$$Y = 0 \quad \text{for } C^2 = 0, \qquad Y = \tfrac{1}{2} \quad \text{for } C^2 = -\tfrac{3}{2}P^4. \tag{4.6}$$

The superspin projection operators for these states are

$$\Pi_0^Y = S, \qquad \Pi_{1/2}^Y = 1 - S. \tag{4.7}$$

In order to split two superspin-0 states, one needs the projection operator for G. The eigenvalue equation which G satisfies is

$$G(G+1)(G-1) = 0. \tag{4.8}$$

The projection operators are

$$\Pi_0^G = 1 - S, \qquad \Pi_{\pm 1}^G = \tfrac{1}{2}(S \pm G). \tag{4.9}$$

Combining (4.7) and (4.9), one obtains the complete projection operators for the irreducible scalar superfields labelled by (Y,G):

$$\Pi_{(1/2,0)} = 1 - S, \qquad \Pi_{(0,\pm 1)} = \tfrac{1}{2}(S \pm G). \tag{4.10}$$

B. $N = 2$ supersymmetry

In order to get explicit expressions for the projection operators, we introduce the following three projection operators:

$$\begin{aligned} A &\equiv (1/1024P^4)((D^iD^i)^2, (\bar D^j \bar D^j)^2)_+, \\ B &\equiv (1/1024P^4)((D^iD^i)^2, (\bar D^j \bar D^j)^2)_-, \\ K &\equiv (1/16P^2)(D^iD^j, \bar D^i \bar D^j)_+. \end{aligned} \tag{4.11}$$

All the projection operators can be expressed as linear combinations of $A, B, K, G, G^2, E, E^2, EG$, and EG^2. We present the multiplication tables of these operators in Table I.

Now consider an SO(2) scalar superfield. The generators M_{ab} are

$$M_{ab} = x_a P_b - x_b P_a + i(\sigma_{ab})_\alpha{}^\beta \theta^{\alpha i} \partial_{\beta i} + i(\bar\sigma_{ab})^{\dot\alpha}{}_{\dot\beta} \bar\theta_{\dot\alpha i} \bar\partial^{\dot\beta i}. \tag{4.12}$$

From (3.2), one has

$$C_{ab} = -\tfrac{1}{4}(P_a D^i\sigma_b \bar D^i - P_b D^i\sigma_a \bar D^i). \tag{4.13}$$

Using Table I, one obtains

$$C^2 = P^4(2K - \tfrac{1}{2}G^2 - 4). \tag{4.14}$$

It satisfies the eigenvalue equation

$$C^4(C^2 + \tfrac{3}{2}P^4)(C^2 + 4P^4) = 0. \tag{4.15}$$

The eigenstates of C^2 correspond to the superspin states as follows:

$$Y = 0 \quad \text{for } C^2 = 0, \qquad Y = \tfrac{1}{2} \quad \text{for } C^2 = -\tfrac{3}{2}P^4, \qquad Y = 1 \quad \text{for } C^2 = -4P^4. \tag{4.16}$$

The projection operators for these superspin states are

$$\begin{aligned} \Pi_0^Y &= \tfrac{1}{2}K - \tfrac{3}{4}G^2 + \tfrac{3}{2}A, \\ \Pi_{1/2}^Y &= G^2 - 4A, \\ \Pi_1^Y &= 1 - \tfrac{1}{2}K - \tfrac{1}{4}G^2 + \tfrac{3}{2}A. \end{aligned} \tag{4.17}$$

The Casimir operator G satisfies the eigenvalue equation

$$G(G^2 - 1)(G^2 - 4) = 0. \tag{4.18}$$

So, the projection operators are

$$\begin{aligned} \Pi_0^G &= 1 - G^2 + 3A, \\ \Pi_{\pm 1}^G &= \pm\tfrac{1}{2}G + \tfrac{1}{2}G^2 - 2A \mp B, \\ \Pi_{\pm 2}^G &= \tfrac{1}{2}(A \pm B). \end{aligned} \tag{4.19}$$

The Casimir operator E satisfies the same eigenvalue equation

$$E(E^2 - 1)(E^2 - 4) = 0, \tag{4.20}$$

and its projection operators are

$$\begin{aligned} \Pi_0^E &= 1 - \tfrac{1}{4}E^2 - \tfrac{3}{4}G^2 + 3A, \\ \Pi_{\pm 1}^E &= \tfrac{1}{2}G^2 - 2A \pm \tfrac{1}{2}EG^2, \\ \Pi_{\pm 2}^E &= \tfrac{1}{8}E^2 - \tfrac{1}{8}G^2 + \tfrac{1}{2}A \pm \tfrac{1}{4}E \mp \tfrac{1}{4}EG^2. \end{aligned} \tag{4.21}$$

Now, multiplying three projection operators for Y, G, and E, one can obtain the projection operator which singles

TABLE I. Multiplication table.

	A	B	K	G	G^2	E	E^2
A	A	B	$3A$	$2B$	$4A$	0	0
B		A	$3B$	$2A$	$4B$	0	0
K			$2K + 6A - \frac{3}{4}G^2$	$\frac{3}{2}G + 3B$	$\frac{3}{2}G^2 + 6A$	$2E - \frac{1}{2}EG^2$	$2E^2 - \frac{1}{2}G^2 + 2A$
G				G^2	$G + 6B$	EG	$G - 2B$
G^2					$G^2 + 12A$	EG^2	$G^2 - 4A$
E						E^2	$4E - 3EG^2$
E^2							$4E^2 - 3G^2 + 12A$

out the irreducible superfield labelled by (Y,G,E):

$$\Pi_{(Y,G,E)} = \Pi^{Y}\Pi^{G}\Pi^{E}. \tag{4.22}$$

Collecting all the nonvanishing products, one has

$$\Pi_{(0,0,0)} = \tfrac{1}{4}(2K - 2G^2 - E^2 + 2A), \tag{4.23}$$

$$\Pi_{(0,0,\pm 2)} = \tfrac{1}{8}(E^2 - G^2 + 4A \pm 2E \mp 2EG^2), \tag{4.24}$$

$$\Pi_{(0,\pm 2,0)} = \tfrac{1}{2}(A \pm B), \tag{4.25}$$

$$\Pi_{(1/2,1,\pm 1)} = \tfrac{1}{4}(G^2 - G - 4A - 2B \pm EG \pm EG^2), \tag{4.26}$$

$$\Pi_{(1/2,-1,\pm 1)} = \tfrac{1}{4}(G^2 - G - 4A + 2B \mp EG \pm EG^2), \tag{4.27}$$

$$\Pi_{(1,0,0)} = 1 - \tfrac{1}{2}K - \tfrac{1}{4}G^2 + \tfrac{3}{4}A. \tag{4.28}$$

There are ten irreducible representations. Taylor[6] has made a similar analysis and has also found that a scalar superfield can be split into ten irreducible representations.

5. CONSTRAINT EQUATIONS

Using the projection operators, one can obtain constraint equations for the irreducible superfields:

$$\Pi\Phi = \Phi. \tag{5.1}$$

However, the projection operators are very complicated in general, and it is not easy to solve (5.1). Instead of solving (5.1) directly, one can obtain simplified but equivalent equations due to the algebraic properties of the projection operators.

A. $N = 1$ supersymmetry

(I) $\Phi_{(0,1)}$:

$$\tfrac{1}{2}(S + G)\Phi_{(0,1)} = 0. \tag{5.2}$$

Since

$$D_\alpha(S + G) = 0, \tag{5.3}$$

one obtains from (5.2)

$$D_\alpha \Phi_{(0,1)} = 0. \tag{5.4}$$

Conversely, since

$$(D^2,\overline{D}^2) = 8P_a D\sigma^a\overline{D} - 16P^2, \tag{5.5}$$

one can obtain (5.2) from (5.4). Hence (5.2) is equivalent to (5.4).

(II) $\Phi_{(0,-1)}$: Since

$$\Phi_{(0,-1)} = \overline{\Phi}_{(0,1)}, \tag{5.6}$$

one obtains

$$\overline{D}_{\dot\alpha}\Phi_{(0,-1)} = 0. \tag{5.7}$$

(III) $\Phi_{(1/2,0)}$:

$$(D^2\overline{D}^2 + \overline{D}^2D^2)\Phi_{(1/2,0)} = 0. \tag{5.8}$$

Since

$$(D^2 \pm \overline{D}^2)(D^2\overline{D}^2 + \overline{D}^2D^2) = 16P^2(D^2 \pm \overline{D}^2), \tag{5.9}$$

one can obtain from (5.8)

$$(D^2 \pm \overline{D}^2)\Phi_{(1/2,0)} = 0. \tag{5.10}$$

Conversely, since

$$(D^2 \pm \overline{D}^2)^2 = \pm(D^2\overline{D}^2 + \overline{D}^2D^2), \tag{5.11}$$

one can obtain (5.8) from (5.10). Hence (5.10) is equivalent to (5.8)

B. $N = 2$ supersymmetry

In order to understand the meaning of the simplified constraint equations, we introduce two simple superspaces with $(\theta_+{}^\alpha,\overline{\theta}_{-\dot\alpha})$ and $(\theta_-{}^\alpha,\overline{\theta}_{+\dot\alpha})$. These θ's are defined by

$$\begin{aligned} \theta_\pm{}^\alpha &= (1/\sqrt{2})(\theta^{\alpha 1} \pm i\theta^{\alpha 2}), \\ \overline{\theta}_{\pm\dot\alpha} &= (1/\sqrt{2})(\overline{\theta}_{\dot\alpha 1} \pm i\overline{\theta}_{\dot\alpha 2}). \end{aligned} \tag{5.12}$$

The spinor charges and the covariant derivatives are

$$\begin{aligned} Q_{\pm\alpha} &= (1/\sqrt{2})(Q_{\alpha 1} \pm iQ_{\alpha 2}), \\ \overline{Q}_{\pm\dot\alpha} &= (1/\sqrt{2})(\overline{Q}_{\dot\alpha 1} \pm i\overline{Q}_{\dot\alpha 2}), \\ D_{\pm\alpha} &= (1/\sqrt{2})(D_{\alpha 1} \pm iD_{\alpha 2}), \\ \overline{D}_{\pm\dot\alpha} &= (1/\sqrt{2})(\overline{D}_{\dot\alpha 1} \pm i\overline{D}_{\dot\alpha 2}). \end{aligned} \tag{5.13}$$

They satisfy

$$\begin{aligned} (Q_{\pm\alpha},Q_{\pm\beta})_+ &= (Q_{\pm\alpha},Q_{\mp\beta})_+ = (Q_{\pm\alpha},\overline{Q}_{\pm\dot\alpha})_+ = 0, \\ (Q_{\pm\alpha},\overline{Q}_{\mp\dot\alpha})_+ &= -2\sigma^a_{\alpha\dot\alpha}P_a, \\ (D_{\pm\alpha},D_{\pm\beta})_+ &= (D_{\pm\alpha},D_{\mp\beta})_+ = (D_{\pm\alpha},\overline{D}_{\pm\dot\alpha})_+ = 0, \\ (D_{\pm\alpha},\overline{D}_{\mp\dot\alpha})_+ &= 2\sigma^a_{\alpha\dot\alpha}P_a. \end{aligned} \tag{5.14}$$

Under SO(2) transformation

$$\begin{aligned} &\theta_\pm{}^\alpha \rightarrow e^{\pm i\phi}\theta_\pm{}^\alpha, \\ &\overline{\theta}_{\pm\dot\alpha} \rightarrow e^{\pm i\phi}\overline{\theta}_{\pm\dot\alpha}. \end{aligned} \tag{5.15}$$

Since the decomposition of $(\theta^{\alpha i},\overline{\theta}_{\dot\alpha i})$ into $(\theta_+{}^\alpha,\overline{\theta}_{-\dot\alpha})$ and $(\theta_-{}^\alpha,\overline{\theta}_{+\dot\alpha})$ is SO(2)-invariant, the SO(2) superspace is a direct sum of the two invariant simple superspaces with the SO(2) transformation replaced by the chiral transformation. The Casimir operators G and E can be expressed in terms of the Casimir operators G_+ and G_- of the two simple superspaces:

$$\begin{aligned} G &= G_+ + G_-, \\ E &= G_+ - G_-, \end{aligned} \tag{5.16}$$

where

$$G_\pm = (P_a/2P^2)D_\mp\sigma_a\overline{D}_\pm - 1. \tag{5.17}$$

The irreducible superfields $\Phi_{(Y,G,E)}$ may now be described in the two simple superspaces.

(I) $\Phi_{(0,0,0)}$:

$$\Pi_{(0,0,0)}\Phi_{(0,0,0)} = \Phi_{(0,0,0)}. \tag{5.18}$$

From (4.23) and Table I, one has

$$(K - G^2 - E^2)\Pi_{(0,0,0)} = 2\Pi_{(0,0,0)}.$$

So

$$(1 - \tfrac{1}{2}K + \tfrac{1}{2}G^2 + \tfrac{1}{2}E^2)\Phi_{(0,0,0)} = 0. \tag{5.19}$$

Since

$$\begin{aligned} A(1 - \tfrac{1}{2}K + \tfrac{1}{2}G^2 + \tfrac{1}{2}E^2) &= \tfrac{3}{2}A, \\ B(1 - \tfrac{1}{2}K + \tfrac{1}{2}G^2 + \tfrac{1}{2}E^2) &= \tfrac{3}{2}B, \\ G(1 - \tfrac{1}{2}K + \tfrac{1}{2}G^2 + \tfrac{1}{2}E^2) &= \tfrac{5}{4}G + \tfrac{1}{2}B, \end{aligned}$$

$$E(1-\tfrac{1}{2}K+\tfrac{1}{2}G^2+\tfrac{1}{2}E^2)=2E-\tfrac{3}{4}EG^2,$$

one can obtain from (5.19)

$$A\Phi_{(0,0,0)}=B\Phi_{(0,0,0)}=G\Phi_{(0,0,0)}=E\Phi_{(0,0,0)}=0.$$

Hence, (5.19) is equivalent to (5.18).

(II) $\Phi_{(0,0,\pm 2)}$:

$$\Pi_{(0,0,\pm 2)}\Phi_{(0,0,\pm 2)}=\Phi_{(0,0,\pm 2)}. \tag{5.20}$$

Since

$$E\Pi_{(0,0,\pm 2)}=\pm 2\Pi_{(0,0,\pm 2)},$$

one has from (5.20)

$$E\Phi_{(0,0,\pm 2)}=\pm 2\Phi_{(0,0,\pm 2)}. \tag{5.21}$$

Using the relations

$$GE^2=G-2B,\quad AE=BE=0,$$

one can obtain (5.20) from (5.21). Hence, (5.21) is equivalent to (5.20). Now, multiplying (5.21) by EG, one obtains

$$G\Phi_{(0,0,\pm 2)}=0. \tag{5.22}$$

Rewriting (5.21) and (5.22) in terms of G_+ and G_-, one has

$$(G_++G_-)\Phi_{(0,0,\pm 2)}=0,$$
$$(G_+-G_-)\Phi_{(0,0,\pm 2)}=\pm 2\Phi_{(0,0,\pm 2)}.$$

Hence

$$G_+\Phi_{(0,0,\pm 2)}=\pm\Phi_{(0,0,\pm 2)},$$
$$G_-\Phi_{(0,0,\pm 2)}=\mp\Phi_{(0,0,\pm 2)}. \tag{5.23}$$

Using the results for the $N=1$ case, one can show that these constraints are equivalent to

$$D_{-\alpha}\Phi_{(0,0,2)}=\bar{D}_{-\dot\alpha}\Phi_{(0,0,2)}=0,$$
$$D_{+\alpha}\Phi_{(0,0,-2)}=\bar{D}_{+\dot\alpha}\Phi_{(0,0,-2)}=0. \tag{5.24}$$

So, $\Phi_{(0,0,\pm 2)}$ is antichiral (chiral) in the superspace with $(\theta_+{}^{\alpha}, \bar\theta_{-\dot\alpha})$ and chiral (antichiral) in the superspace with $(\theta_-{}^{\alpha}, \bar\theta_{+\dot\alpha})$. In a different context, Ogievetsky has called the constraint (5.24) the Grassmann analyticity condition.[9]

(III) $\Phi_{(0,\pm 2,0)}$:

$$\Pi_{(0,\pm 2,0)}\Phi_{(0,\pm 2,0)}=\Phi_{(0,\pm 2,0)}. \tag{5.25}$$

Since

$$G\Pi_{(0,\pm 2,0)}=\pm 2\Pi_{(0,\pm 2,0)},$$

one has

$$G\Pi_{(0,\pm 2,0)}=\pm 2\Phi_{(0,\pm 2,0)}. \tag{5.26}$$

Using the relations

$$G^3=G+6B,$$
$$G^4=G^2+12A,$$

one can obtain (5.25) from (5.26). Hence (5.26) is equivalent to (5.25). Since one can obtain from (5.25)

$$E\Phi_{(0,\pm 2,0)}=0, \tag{5.27}$$

the constraint equations can be written in terms of G_+ and G_- as

$$(G_++G_-)\Phi_{(0,\pm 2,0)}=\pm 2\Phi_{(0,\pm 2,0)},$$
$$(G_+-G_-)\Phi_{(0,\pm 2,0)}=0.$$

Hence

$$G_+\Phi_{(0,\pm 2,0)}=G_-\Phi_{(0,\pm 2,0)}=\pm\Phi_{(0,\pm 2,0)}. \tag{5.28}$$

It can be easily shown that (5.28) is equivalent to

$$D_{\pm\alpha}\Phi_{(0,2,0)}=0,$$
$$\bar{D}_{\pm\dot\alpha}\Phi_{(0,-2,0)}=0. \tag{5.29}$$

(IV) $\Phi_{(1/2,1,\pm 1)}, \Phi_{(1/2,-1,\pm 1)}$:

$$\Pi_{(1/2,g,e)}\Phi_{(1/2,g,e)}=\Phi_{(1/2,g,e)}. \tag{5.30}$$

Using the multiplication table, one can show that

$$G\Pi_{(1/2,g,e)}=g\Pi_{(1/2,g,e)},$$
$$E\Pi_{(1/2,g,e)}=e\Pi_{(1/2,g,e)}.$$

So, it follows from (5.30) that

$$G\Phi_{(1/2,g,e)}=g\Phi_{(1/2,g,e)},$$
$$E\Phi_{(1/2,g,e)}=e\Phi_{(1/2,g,e)}. \tag{5.31}$$

Since one can obtain from (5.31)

$$A\Phi_{(1/2,g,e)}=B\Phi_{(1/2,g,e)}=0,$$

(5.31) is equivalent to (5.30). In terms of the Casimir operators G_+ and G_-, (5.31) can be rewritten as

$$G_+\Phi_{(1/2,g,e)}=\tfrac{1}{2}(g+e)\Phi_{(1/2,g,e)},$$
$$G_-\Phi_{(1/2,g,e)}=\tfrac{1}{2}(g-e)\Phi_{(1/2,g,e)}. \tag{5.32}$$

Moreover, one can show that (5.32) is equivalent to

$$\begin{aligned}
&D_{-\alpha}\Phi_{(1/2,1,1)}=(D_+{}^2\pm\bar{D}_-{}^2)\Phi_{(1/2,1,1)}=0,\\
&D_{+\alpha}\Phi_{(1/2,1,-1)}=(D_-{}^2\pm\bar{D}_+{}^2)\Phi_{(1/2,1,-1)}=0,\\
&\bar{D}_{-\dot\alpha}\Phi_{(1/2,-1,1)}=(D_-{}^2\pm\bar{D}_+{}^2)\Phi_{(1/2,-1,1)}=0,\\
&\bar{D}_{+\dot\alpha}\Phi_{(1/2,-1,-1)}=(D_+{}^2\pm\bar{D}_-{}^2)\Phi_{(1/2,-1,-1)}=0.
\end{aligned} \tag{5.33}$$

Hence, $\Phi_{(1/2,g,e)}$ is chiral or antichiral in one of the two invariant simple superspaces and the superspin-1/2 representation in the other superspace.

(V) $\Phi_{(1,0,0)}$:

$$\Pi_{(1,0,0)}\Phi_{(1,0,0)}=\Phi_{(1,0,0)}. \tag{5.34}$$

Since

$$K\Pi_{(1,0,0)}=0,$$

one obtains from (5.34)

$$K\Phi_{(1,0,0)}=0. \tag{5.35}$$

Multiplying (5.35) by A, B, and G, one obtains

$$A\Phi_{(1,0,0)}=B\Phi_{(1,0,0)}=G\Phi_{(1,0,0)}=0.$$

So, one can obtain (5.34) from (5.33). Hence (5.35) is equivalent to (5.34).

We have seen that some of the irreducible superfields satisfy the constraints for the irreducible representation in the two invariant simple superspaces. In fact, all the superfields can be classified according to the irreducible representation of the two invariant simple superspaces. Both $\Phi_{(0,0,0)}$ and $\Phi_{(1,0,0)}$ belong to the superspin-1/2 representation of each of the two invariant superspaces. They can be distinguished by the constraints (5.19) and (5.35). The classification according to this scheme is given in Table II.

6. SO(2) SUPERTENSORS IN SUPERSPACE

In order to solve the constraint equations for the irreducible superfields, one has to apply the Casimir operators to a

TABLE II. Classification in invariant simple superspaces.[a]

(Y,G,E) $\pm$	$(0,0,0)$	$(0,0,2)$	$(0,0,-2)$	$(0,2,0)$	$(0,-2,0)$	$(\frac{1}{2},1,1)$	$(\frac{1}{2},1,-1)$	$(\frac{1}{2},-1,1)$	$(\frac{1}{2},-1,-1)$	$(1,0,0)$
$(\theta_+,\bar{\theta}_-)$	$\frac{1}{2}$	0^-	0^+	0^-	0^+	0^-	$\frac{1}{2}$	$\frac{1}{2}$	0^+	$\frac{1}{2}$
$(\theta_-,\bar{\theta}_+)$	$\frac{1}{2}$	0^+	0^-	0^-	0^+	$\frac{1}{2}$	0^-	0^+	$\frac{1}{2}$	$\frac{1}{2}$

[a] $\frac{1}{2}$, 0^+, and 0^- denote superspin $\frac{1}{2}$, chiral, and antichiral representations, respectively.

general superfield and solve simultaneous equations. So, before one tries to solve the constraint equations, one needs the most general superfield. In order to maintain SO(2) invariance, it should be a sum of contractions of irreducible component fields with irreducible SO(2) supertensors in superspace. Since the irreducible SO(2) tensors are either totally symmetric or totally antisymmetric in their SO(2) indices, one can obtain them by symmetrizing or antisymmetrizing the SO(2) indices. The list of all the irreducible SO(2) supertensors and their number of components are given in Table III. The total number of components turns out to be $2^8 = 256$, as required. The definition of some supertensors in the table are

$$\Theta^{(ij)} = \theta^i\theta^j - \tfrac{1}{2}\delta^{ij}(\theta^k\theta^k),$$
$$\theta^{(i}\sigma^a\bar{\theta}^{j)} = \tfrac{1}{2}(\theta^i\sigma^a\bar{\theta}^j + \theta^j\sigma^a\bar{\theta}^i - \delta^{ij}\theta^k\sigma^a\bar{\theta}^k),$$
$$\theta^{[i}\sigma^a\bar{\theta}^{j]} = \tfrac{1}{2}(\theta^i\sigma^a\bar{\theta}^j - \theta^j\sigma^a\bar{\theta}^i),$$
$$\Theta^{\alpha(ijk)} = \tfrac{1}{3}(\bar{\Theta}^{(jk)}\theta^{\alpha i} + \bar{\Theta}^{(ik)}\theta^{\alpha j} + \bar{\Theta}^{(ij)}\theta^{\alpha k}) - \text{trace},$$
$$\Theta^{(ijkl)} = \tfrac{1}{6}(\Theta^{(ij)}\bar{\Theta}^{(kl)} + \Theta^{(ik)}\bar{\Theta}^{(jl)} + \Theta^{(il)}\bar{\Theta}^{(jk)} + \Theta^{(jk)}\bar{\Theta}^{(il)} + \Theta^{(jl)}\bar{\Theta}^{(ik)} + \Theta^{(kl)}\bar{\Theta}^{(ij)}) - \text{trace},$$
$$\Theta^{[ij]} = \tfrac{1}{2}(\Theta^{(ik)}\bar{\Theta}^{(kj)} - \Theta^{(jk)}\bar{\Theta}^{(ki)}). \tag{6.1}$$

When one does explicit calculations, one frequently has to multiply two supertensors and decompose the product into a sum of irreducible supertensors. Also, in order to construct a tensor calculus, one needs these reduction formulas. Some of the reduction formulas which are frequently used are given in Appendix B.

7. IRREDUCIBLE SUPERFIELDS

In Sec. 5, we derived the simplified constraint equations. Although it is straightforward to solve those con-

TABLE III. Irreducible SO (2) supertensors.

Boson part		Fermion part	
Supertensors	# Comp.	Supertensors	# Comp.
1	1		
$\theta^i\theta^i$, $\bar{\theta}^i\bar{\theta}^i$	$1+1$	$\theta^{\alpha i}$, $\bar{\theta}_{\dot{\alpha}i}$	$4+4$
$\Theta^{(ij)}$, $\bar{\Theta}^{(ij)}$,	$2+2$		
$\theta^i\sigma^{ab}\theta^j$, $\bar{\theta}^i\bar{\sigma}^{ab}\bar{\theta}^j$,	$3+3$		
$\theta^i\sigma^a\bar{\theta}^i$	4		
$\theta^{(i}\sigma^a\bar{\theta}^{j)}$	8		
$\theta^{[i}\sigma^a\bar{\theta}^{j]}$	4	$(\theta^k\theta^k)\theta^{\alpha i}$, $(\bar{\theta}^k\bar{\theta}^k)\bar{\theta}_{\dot{\alpha}i}$	$4+4$
$(\theta^k\theta^k)^2$, $(\bar{\theta}^k\bar{\theta}^k)^2$	$1+1$	$\Theta^{\alpha(ijk)}$, $\bar{\Theta}_{\dot{\alpha}(ijk)}$	$4+4$
$(\theta^k\theta^k)(\theta^i\sigma^a\bar{\theta}^i)$, $(\bar{\theta}^k\bar{\theta}^k)(\theta^i\sigma^a\bar{\theta}^i)$	$4+4$	$(\theta^k\sigma^a\bar{\theta}^k)\theta^{\alpha i}$, $(\theta^k\sigma^a\bar{\theta}^k)\bar{\theta}_{\dot{\alpha}i}$	$16+16$
$(\theta^k\theta^k)\theta^{(i}\sigma^a\bar{\theta}^{j)}$, $(\bar{\theta}^k\bar{\theta}^k)\theta^{(i}\sigma^a\bar{\theta}^{j)}$	$8+8$	$(\bar{\theta}^k\bar{\theta}^k)\theta^{\alpha i}$, $(\theta^k\theta^k)\bar{\theta}_{\dot{\alpha}i}$	$4+4$
$(\theta^k\theta^k)\theta^{[i}\sigma^a\bar{\theta}^{j]}$, $(\bar{\theta}^k\bar{\theta}^k)\theta^{[i}\sigma^a\bar{\theta}^{j]}$,	$4+4$		
$(\theta^k\theta^k)\bar{\Theta}^{(ij)}$, $(\bar{\theta}^k\bar{\theta}^k)\Theta^{(ij)}$	$2+2$		
$(\bar{\theta}^k\bar{\theta}^k)(\theta^i\sigma^{ab}\theta^j)$, $(\theta^k\theta^k)(\bar{\theta}^i\bar{\sigma}^{ab}\bar{\theta}^j)$	$3+3$		
$\bar{\Theta}^{(ik)}(\theta^k\sigma^{ab}\theta^j)$, $\Theta^{(ik)}(\bar{\theta}^k\sigma^{ab}\bar{\theta}^j)$	$6+6$		
$\Theta^{(ijkl)}$	2	$(\bar{\theta}^k\bar{\theta}^k)^2\theta^{\alpha i}$, $(\theta^k\theta^k)^2\bar{\theta}_{\dot{\alpha}i}$	$4+4$
$(\theta^k\theta^k)(\bar{\theta}^l\bar{\theta}^l)$	1	$(\theta^l\theta^l)\Theta^{\alpha(ijk)}$, $(\bar{\theta}^l\bar{\theta}^l)\bar{\Theta}_{\dot{\alpha}(ijk)}$	$4+4$
$\Theta^{[ij]}$	1	$(\bar{\theta}^k\bar{\theta}^k)(\theta^l\sigma^a\bar{\theta}^l)\theta^{\alpha i}$, $(\theta^k\theta^k)(\theta^l\sigma^a\bar{\theta}^l)\bar{\theta}_{\dot{\alpha}i}$	$16+16$
$(\theta^k\sigma^a\bar{\theta}^k)(\theta^l\sigma^b\bar{\theta}^l)$	10	$(\theta^k\theta^k)(\bar{\theta}^l\bar{\theta}^l)\theta^{\alpha i}$, $(\theta^k\theta^k)(\bar{\theta}^l\bar{\theta}^l)\bar{\theta}_{\dot{\alpha}i}$	$4+4$
$(\theta^k\theta^k)^2(\bar{\theta}^l\bar{\theta}^l)$, $(\theta^k\theta^k)(\bar{\theta}^l\bar{\theta}^l)^2$	$1+1$		
$(\theta^k\theta^k)^2\bar{\Theta}^{(ij)}$, $(\bar{\theta}^k\bar{\theta}^k)^2\Theta^{(ij)}$	$2+2$		
$(\theta^k\theta^k)^2(\bar{\theta}^i\bar{\sigma}^{ab}\bar{\theta}^j)$, $(\bar{\theta}^k\bar{\theta}^k)^2(\theta^i\sigma^{ab}\theta^j)$	$3+3$		
$(\theta^k\theta^k)(\bar{\theta}^l\bar{\theta}^l)(\theta^i\sigma^a\bar{\theta}^i)$	4		
$(\theta^k\theta^k)(\bar{\theta}^l\bar{\theta}^l)\theta^{(i}\sigma^a\bar{\theta}^{j)}$	8	$(\theta^k\theta^k)(\bar{\theta}^l\bar{\theta}^l)^2\theta^{\alpha i}$, $(\theta^k\theta^k)^2(\bar{\theta}^l\bar{\theta}^l)\bar{\theta}_{\dot{\alpha}i}$	$4+4$
$(\theta^k\theta^k)(\bar{\theta}^l\bar{\theta}^l)\theta^{[i}\sigma^a\bar{\theta}^{j]}$	4		
$(\theta^k\theta^k)^2(\bar{\theta}^l\bar{\theta}^l)^2$	1		
Total	128	Total	128

straint equations, there are some ambiguities. Each term in the superfields can be multiplied by some powers of ∂^2, $\sigma\cdot\partial$, or $\bar{\sigma}\cdot\partial$ and still can satisfy the same constraint equations. So, in order to give correct dimension to the component fields in the irreducible superfields, one should divide each term in the superfields by proper powers of ∂^2, $\sigma\cdot\partial$, or $\bar{\sigma}\cdot\partial$ such that the solutions of the constraint equations are the same as the results obtained by applying the projection operators $\Pi_{(Y,G,E)}$ to the most general superfield.

It turns out that some of the solutions are still reducible, and this introduces supersymmetric gauge invariance. Since the Lagrangians and the supersymmetry transformation rules are much simpler in the Dirac four-component notation, they are given in terms of the four-component Majorana spinors.

(I) $\Phi_{(0,0,0)}$: This multiplet contains 8 + 8 field components. The irreducible superfield corresponding to this multiplet is

$$\begin{aligned}\partial^4\Phi_{(0,0,0)} &= \tfrac{1}{2}\overset{(2)}{\Theta}\partial^2(M+iN) + \tfrac{1}{2}\overset{(2)}{\bar{\Theta}}\partial^2(M-iN) + \overset{(2)}{\Theta}{}^{a[ij]}\partial^2 V_a^{[ij]} \\ &\quad - \overset{(0)}{\Theta}F + \overset{(2)}{\Theta}{}^{(ij)}G^{(ij)} + i\overset{(1)}{\Theta}{}^{i}\sigma\cdot\partial\bar{\chi}^i + i\overset{(1)}{\bar{\Theta}}{}^{i}\bar{\sigma}\cdot\partial\chi^i, \qquad (7.1)\end{aligned}$$

where the SO(2) supertensors are

$$\begin{aligned}\overset{(2)}{\Theta} &= \theta^k\theta^k + i(\theta^k\theta^k)(\theta^l\sigma^a\bar{\theta}^l)\partial_a - \tfrac{1}{8}(\theta^k\theta^k)^2(\bar{\theta}^l\bar{\theta}^l)\partial^2, \\ \overset{(2)}{\Theta}{}^{a[ij]} &= i\left[\theta^{[i}\sigma_b\bar{\theta}^{j]} - \tfrac{1}{4}(\theta^k\theta^k)(\bar{\theta}^l\bar{\theta}^l)\theta^{[i}\sigma_b\bar{\theta}^{j]}\partial^2\right]\Pi^{ab} \\ &\quad - \tfrac{1}{2}\left[(\bar{\theta}^k\bar{\theta}^k)(\theta^i\sigma^{ab}\theta^j) + (\theta^k\theta^k)(\bar{\theta}^i\bar{\sigma}^{ab}\bar{\theta}^j)\right]\partial_b, \\ \overset{(0)}{\Theta} &= 1 - \tfrac{1}{2}(\theta^k\theta^k)(\bar{\theta}^l\bar{\theta}^l)\partial^2 + \tfrac{1}{2}(\theta^k\sigma^a\bar{\theta}^k)(\theta^l\sigma^b\bar{\theta}^l)\partial_a\partial_b \\ &\quad + \tfrac{1}{64}(\theta^k\theta^k)^2(\bar{\theta}^l\bar{\theta}^l)^2\partial^4, \\ \overset{(2)}{\Theta}{}^{(ij)} &= \left[\theta^{(i}\sigma^a\bar{\theta}^{j)} - \tfrac{1}{4}(\theta^k\theta^k)(\bar{\theta}^l\bar{\theta}^l)\theta^{(i}\sigma^a\bar{\theta}^{j)}\partial^2\right]\partial_a \qquad (7.2) \\ &\quad - \tfrac{1}{4}i\left[(\theta^k\theta^k)\bar{\Theta}^{(ij)} - (\bar{\theta}^k\bar{\theta}^k)\Theta^{(ij)}\right]\partial^2, \\ \overset{(1)}{\Theta}{}^{\alpha i} &= \theta^{\alpha i} - i(\theta^k\theta^k)(\bar{\theta}^i\bar{\sigma}\cdot\partial)^\alpha + i(\theta^k\sigma^a\bar{\theta}^k)\theta^{\alpha i}\partial_a \\ &\quad - \tfrac{1}{4}(\theta^k\theta^k)(\bar{\theta}^l\bar{\theta}^l)\theta^{\alpha i}\partial^2 \\ &\quad + \tfrac{1}{2}(\theta^k\theta^k)(\theta^l\sigma^a\bar{\theta}^l)(\bar{\theta}^i\bar{\sigma}\cdot\partial)^\alpha\partial_a \\ &\quad + \tfrac{1}{16}i(\theta^k\theta^k)^2(\bar{\theta}^l\bar{\theta}^l)(\bar{\theta}^i\bar{\sigma}\cdot\partial)^\alpha\partial^2.\end{aligned}$$

Here Π^{ab} is the projection operator for the vector field:

$$\Pi^{ab} = \eta^{ab} - \frac{\partial^a\partial^b}{\partial^2}. \qquad (7.3)$$

The invariant Lagrangian of this multiplet is

$$\begin{aligned}L_{(0,0,0)} &= 2\int d^8\theta\, \Phi_{(0,0,0)}\,\partial^4\Phi_{(0,0,0)} \\ &= -\tfrac{1}{8}(F(V)_{ab}^{[ij]})^2 + \tfrac{1}{2}i\bar{\chi}^i\partial\!\!\!/\chi^i + \tfrac{1}{2}(\partial_a M)^2 + \tfrac{1}{2}(\partial_a N)^2 \\ &\quad + \tfrac{1}{2}F^2 + \tfrac{1}{4}(G^{(ij)})^2. \qquad (7.4)\end{aligned}$$

The fields of this multiplet transform under the supersymmetry transformation according to

$$\begin{aligned}\delta V_a^{[ij]} &= -2i\bar{\epsilon}^{[i}\gamma_a\chi^{j]}, \\ \delta\chi^i &= F(V)_{ab}^{[ij]}\sigma^{ab}\epsilon^j - i\partial\!\!\!/(M + i\gamma^5 N)\epsilon^i \\ &\quad + F\epsilon^i + iG^{(ij)}\gamma^5\epsilon^j, \\ \delta M &= \bar{\epsilon}^i\chi^i, \qquad\qquad (7.5) \\ \delta N &= i\bar{\epsilon}^i\gamma^5\chi^i, \\ \delta F &= -i\bar{\epsilon}^i\partial\!\!\!/\chi^i, \\ \delta G^{(ij)} &= 2\bar{\epsilon}^{(i}\gamma^5\partial\!\!\!/\chi^{j)}.\end{aligned}$$

(II) $\Phi_{(0,0,\pm 2)}$: All the derivative terms of these superfields are factorizable, and the rest of the superfields can be written in terms of $\theta_-{}^\alpha$ and $\bar{\theta}_{-\dot{\alpha}}$ ($\theta_+{}^\alpha$ and $\bar{\theta}_{+\dot{\alpha}}$). The irreducible superfields are

$$\begin{aligned}\Phi_{(0,0,\pm 2)} &= \exp[\pm\epsilon^{mn}\theta^m\sigma^b\bar{\theta}^n\partial_b] \\ &\quad\times(A \pm \tfrac{1}{2}i\epsilon^{ij}B^{[ij]} + \tfrac{1}{2}\Theta^{(ij)}(1 \pm T)(M^{(ij)} + iN^{(ij)}) \\ &\quad + \tfrac{1}{2}\bar{\Theta}^{(ij)}(1 \pm T)(M^{(ij)} - iN^{(ij)}) \\ &\quad + \theta^{(i}\sigma^a\bar{\theta}^{j)}(1 \pm T)A_a^{(ij)} \\ &\quad + \tfrac{1}{4}\Theta^{(ijkl)}(1 \pm T)C^{(ijkl)} \\ &\quad + \theta^{\alpha i}(1 \pm T)\chi_{\alpha i} + \bar{\theta}_{\dot{\alpha} i}(1 \pm T)\bar{\chi}^{\dot{\alpha} i} \\ &\quad + \Theta^{\alpha(ijk)}(1 \pm T)\lambda_{\alpha(ijk)} + \bar{\Theta}_{\dot{\alpha}(ijk)}(1 \pm T)\bar{\lambda}^{\dot{\alpha}(ijk)}), \qquad (7.6)\end{aligned}$$

where the SO(2) supertensors are defined in (6.1) and

$$\begin{aligned}T\chi^{\alpha i} &= i\epsilon^{ij}\chi^{\alpha j}, \\ TM^{(ij)} &= \tfrac{1}{2}i(\epsilon^{ik}M^{(kj)} + \epsilon^{jk}M^{(ik)}), \\ T\lambda^{\alpha(ijk)} &= \tfrac{1}{3}i(\epsilon^{il}\lambda^{\alpha(ljk)} + \epsilon^{jl}\lambda^{\alpha(ilk)} + \epsilon^{kl}\lambda^{\alpha(ijl)}), \quad \text{etc.} \qquad (7.7)\end{aligned}$$

Since all the component fields in these superfields have different SO(2) tensor properties, they cannot be reduced any further. However, these superfields cannot give second-order gauge-invariant Lagrangians without central charges. So, we will consider the superfields with two symmetric SO(2) indices instead of $\Phi_{(0,0,\pm 2)}$. For notational convenience, we will drop the label $(0,0,\pm 2)$ from now on:

$$\begin{aligned}\Phi^{(ij)} &= \exp[\pm\epsilon^{mn}\theta^m\sigma^b\bar{\theta}^n\partial_b] \\ &\quad\times(C^{(ij)} + \tfrac{1}{2}\Theta^{(ij)}[M + iN \pm \tfrac{1}{2}i\epsilon^{kl}(M^{[kl]} + iN^{[kl]})] \\ &\quad + \tfrac{1}{2}\bar{\Theta}^{(ij)}[M - iN \pm \tfrac{1}{2}i\epsilon^{kl}(M^{[kl]} - iN^{[kl]})] \\ &\quad + \theta^{(i}\sigma^a\bar{\theta}^{j)}(A_a \pm \tfrac{1}{2}i\epsilon^{kl}A_a^{[kl]}) \\ &\quad + \tfrac{1}{4}\Theta^{(ijkl)}(1 \pm T)F^{(kl)} + 2\theta^{(i}\chi^{j)} \\ &\quad + 2\bar{\theta}^{(i}\bar{\chi}^{j)} + \Theta^{\alpha(ijk)}(1 \pm T)\psi_{\alpha k} \\ &\quad + \bar{\Theta}_{\dot{\alpha}(ijk)}(1 \pm T)\bar{\psi}^{\dot{\alpha} k}). \qquad (7.8)\end{aligned}$$

Rewriting these superfields in terms of $\theta^{\pm\alpha}$ and $\bar{\theta}_{\pm\dot{\alpha}}$, one has

$$
\begin{aligned}
\Phi^{\mp\mp} = \exp[\pm i(\theta_+\sigma^a\bar\theta_- - \theta_-\sigma^a\bar\theta_+)\partial_a]
&\times (C^{\mp\mp} + \tfrac{1}{2}\theta_\mp{}^2[M + iN \\
&\pm \tfrac{1}{2}i\epsilon^{kl}(M^{[kl]} + iN^{[kl]})] \\
&+ \tfrac{1}{2}\bar\theta_\mp{}^2[M - iN \pm \tfrac{1}{2}i\epsilon^{kl}(M^{[kl]} - iN^{[kl]})] \\
&+ \theta_\mp\sigma^a\bar\theta_\mp(A_a \pm \tfrac{1}{2}i\epsilon^{kl}A_a^{[kl]}) \\
&+ \theta_\mp{}^2\bar\theta_\mp{}^2F^{\pm\pm} + 2\theta_\mp\chi_\mp \\
&+ 2\bar\theta_\mp\bar\chi_\mp + 2\bar\theta_\mp{}^2(\theta_\mp\psi_\pm) + 2\theta_\mp{}^2(\bar\theta_\mp\bar\psi_\pm)),
\end{aligned}
\tag{7.9}
$$

where

$$
\begin{aligned}
&\chi_\pm = (1/\sqrt{2})(\chi^1 \pm i\chi^2), \\
&C^{\pm\pm} = C^{(11)} \pm iC^{(12)}.
\end{aligned}
\tag{7.10}
$$

These superfields can be reduced further by imposing the constraint

$$
D_+{}^2\bar D_+{}^2\Phi_\pm^{--} = \pm 16\partial^2\Phi_\pm^{++}.
\tag{7.11}
$$

Choosing the plus sign and solving the constraint, one obtains

$$
\begin{aligned}
\Phi_+^{\mp\mp} = \exp[\pm i(\theta_+\sigma^a\bar\theta_- - \theta_-\sigma^a\bar\theta_+)\partial_a]
&\times (C^{\mp\mp} \pm \theta_\mp{}^2\tfrac{1}{2}i\epsilon^{ij}(M^{[ij]} + iN^{[ij]}) \\
&\pm \bar\theta_\mp{}^2\tfrac{1}{2}i\epsilon^{ij}(M^{[ij]} - iN^{[ij]}) \\
&+ \theta_\mp\sigma_a\bar\theta_\mp(\Pi^{ab}A_b \pm \tfrac{1}{2}i\epsilon^{ij}\partial^a A^{[ij]}) \\
&+ \theta_\mp{}^2\bar\theta_\mp{}^2\partial^2 C^{\pm\pm} + 2\theta_\mp\chi_\mp \\
&+ 2\bar\theta_\mp\bar\chi_\mp + 2i\bar\theta_\mp{}^2(\theta_\mp\sigma\cdot\partial\bar\chi_\pm) \\
&+ 2i\theta_\mp{}^2(\bar\theta_\mp\bar\sigma\cdot\partial\chi_\pm)).
\end{aligned}
\tag{7.12}
$$

With minus sign, one obtains

$$
\begin{aligned}
\Phi_-^{\mp\mp} = \exp[\pm i(\theta_+\sigma^a\bar\theta_- - \theta_-\sigma^a\bar\theta_+)\partial_a]
&\times (C^{\mp\mp} + \tfrac{1}{2}\theta_\mp{}^2(M + iN) \\
&+ \tfrac{1}{2}\bar\theta_\mp{}^2(M - iN) - \theta_\mp{}^2\bar\theta_\mp{}^2\partial^2 C^{\pm\pm} \\
&+ \theta_\mp\sigma_a\bar\theta_\mp(\partial^a A \pm \tfrac{1}{2}i\epsilon^{ij}\Pi^{ab}A_b^{[ij]}) \\
&+ 2\theta_\mp\chi_\mp + 2\bar\theta_\mp\bar\chi_\mp - 2i\bar\theta_\mp{}^2(\theta_\mp\sigma\cdot\partial\bar\chi_\pm) \\
&- 2i\theta_\mp{}^2(\bar\theta_\mp\bar\sigma\cdot\partial\chi_\pm)).
\end{aligned}
\tag{7.13}
$$

One can easily see that these superfields $\Phi_\pm^{\mp\mp}$ cannot form physical supermultiplets by dimensional argument. Instead, they appear as supersymmetric gauge degrees of freedom.

We now construct invariant Lagrangians. Since the superfields have two axial vector fields, they contain two supermultiplets. The multiplet with A_a has the following Lagrangian:

$$
L_+ = 2\int d^8\theta\left(\Phi^{++}\Phi^{--} + \frac{1}{32}\left(\Phi^{--}\frac{1}{\partial^2}D_+{}^2\bar D_+{}^2\Phi^{--} + \Phi^{++}\frac{1}{\partial^2}D_-{}^2\bar D_-{}^2\Phi^{++}\right)\right).
\tag{7.14}
$$

L_+ has supersymmetric gauge invariance. It is invariant under

$$
\Phi^{\mp\mp} \to \Phi^{\mp\mp} + \Phi_-^{\mp\mp}.
\tag{7.15}
$$

There are 8 + 8 field components in this multiplet. Redefining some of the fields as

$$
\begin{aligned}
&F^{(ij)} \to F^{(ij)} - \partial^2 C^{(ij)}, \\
&\psi^i \to \psi^i - i\not\partial\chi^i,
\end{aligned}
\tag{7.16}
$$

L_+ takes the form

$$
\begin{aligned}
L_+ = &-\tfrac{1}{4}(F(A)_{ab})^2 + \tfrac{1}{2}i\bar\psi^i\not\partial\psi^i \\
&+ \tfrac{1}{4}(\partial_a M^{[ij]})^2 + \tfrac{1}{4}(\partial_a N^{[ij]})^2 \\
&+ \tfrac{1}{4}(F^{(ij)})^2 + \tfrac{1}{4}(G^{[ij]})^2,
\end{aligned}
\tag{7.17}
$$

where

$$
G^{[ij]} = \partial\cdot A^{[ij]}.
\tag{7.18}
$$

In the Wess–Zumino gauge, this multiplet transforms under supersymmetry transformation according to

$$
\begin{aligned}
&\delta A_a = \bar\epsilon^i\gamma_5\gamma_a\psi^i, \\
&\delta\psi^i = -iF(A)_{ab}\gamma^5\sigma^{ab}\epsilon^i + i\not\partial(M^{[ij]} + iN^{[ij]}\gamma^5)\epsilon^j \\
&\qquad + (F^{(ij)} - iG^{[ij]}\gamma^5)\epsilon^j, \\
&\delta M^{[ij]} = 2\bar\epsilon^{[i}\psi^{j]}, \\
&\delta N^{[ij]} = 2i\,\bar\epsilon^{[i}\gamma^5\psi^{j]}, \\
&\delta F^{(ij)} = -2i\,\bar\epsilon^{(i}\not\partial\psi^{j)}, \\
&\delta G^{[ij]} = 2\bar\epsilon^{[i}\gamma^5\not\partial\psi^{j]}.
\end{aligned}
\tag{7.19}
$$

The second multiplet containing $A_a^{[ij]}$ has the following Lagrangian:

$$
L_- = 2\int d^8\theta\left(\Phi^{++}\Phi^{--} - \frac{1}{32}\left(\Phi^{--}\frac{1}{\partial^2}D_+{}^2\bar D_+{}^2\Phi^{--} + \Phi^{++}\frac{1}{\partial^2}D_-{}^2\bar D_-{}^2\Phi^{++}\right)\right).
\tag{7.20}
$$

L_- is invariant under the following supersymmetric gauge transformation:

$$
\Phi^{\mp\mp} \to \Phi^{\mp\mp} + \Phi_+^{\mp\mp}.
\tag{7.21}
$$

This multiplet also contains 8 + 8 field components. Redefining $F^{(ij)}$ and ψ^i as

$$
\begin{aligned}
&F^{(ij)} \to F^{(ij)} + \partial^2 C^{(ij)}, \\
&\psi^i \to \psi^i + i\not\partial\chi^i,
\end{aligned}
\tag{7.22}
$$

L_- becomes

$$
\begin{aligned}
L_- = &-\tfrac{1}{8}(F(A)_{ab}^{[ij]})^2 + \tfrac{1}{2}i\bar\psi^i\not\partial\psi^i + \tfrac{1}{2}(\partial_a M)^2 + \tfrac{1}{2}(\partial_a N)^2 \\
&+ \tfrac{1}{4}(F^{(ij)})^2 + \tfrac{1}{2}G^2,
\end{aligned}
\tag{7.23}
$$

where

$$
G = \partial\cdot A.
\tag{7.24}
$$

In the Wess–Zumino gauge, this multiplet transforms according to

$$\begin{aligned}
&\delta A_a^{[ij]} = 2\bar\epsilon^{[i}\gamma_5\gamma_a\psi^{j]},\\
&\delta\psi^i = iF(A)_{ab}^{[ij]}\gamma^5\sigma^{ab}\epsilon^j - i\not\partial(M+iN\gamma^5)\epsilon^i + F^{(ij)}\epsilon^j + iG\gamma^5\epsilon^i,\\
&\delta M = \bar\epsilon^i\psi^i,\\
&\delta N = i\bar\epsilon^i\gamma^5\psi^i,\\
&\delta F^{(ij)} = -2i\bar\epsilon^{(i}\not\partial\psi^{j)},\\
&\delta G = \bar\epsilon^i\gamma^5\not\partial\psi^i.
\end{aligned} \tag{7.25}$$

(III) $\Phi_{(0,\pm 2,0)}$: These superfields like the preceding ones are factorizable, and the rest of them can be written in terms of $\theta^{\alpha i}$ or $\bar\theta_{\alpha i}$:

$$\begin{aligned}
\Phi_{(0,-2,0)} &= \exp[-i(\theta^l\sigma^c\bar\theta^l)\partial_c](A+iB+\tfrac{1}{2}(\theta^k\theta^k)(M+iN)\\
&\quad+\tfrac{1}{2}\Theta^{(ij)}(M^{(ij)}+iN^{(ij)})+\tfrac{1}{2}(\theta^i\sigma^{ab}\theta^j)T_{ab}^{[ij]}\\
&\quad+\tfrac{1}{2}(\theta^k\theta^k)^2(F+iG)+2\theta^i\chi^i+2(\theta^k\theta^k)(\theta^i\psi^i)),
\end{aligned} \tag{7.26}$$

$$\Phi_{(0,2,0)} = \bar\Phi_{(0,-2,0)}. \tag{7.27}$$

We will drop the label for the chiral superfield for notational simplicity from now on. The chiral superfield can be reduced further by imposing the constraint[10]

$$(D^kD^k)^2\Phi_\pm = \pm 32\partial^2\bar\Phi_\pm. \tag{7.28}$$

Solving the constraint with plus sign, one obtains

$$\begin{aligned}
\Phi_+ &= \exp[-i(\theta^l\sigma^c\bar\theta^l)\partial_c](A-iB+i(\theta^k\theta^k)G+\Theta^{(ij)}F^{(ij)}\\
&\quad+2i(\theta^i\sigma^{ab}\theta^j)\partial_a A_b^{[ij]}+\tfrac{1}{2}(\theta^k\theta^k)^2\partial^2(A+iB)+2\theta^i\chi^i\\
&\quad+2i(\theta^k\theta^k)(\theta^i\sigma\cdot\partial\bar\chi^i)).
\end{aligned} \tag{7.29}$$

There are 8 + 8 field components in this multiplet. The invariant Lagrangian of this multiplet is

$$\begin{aligned}
L_+ &= -\frac{1}{2}\int d^4\theta\,\Phi_+\Phi_+\\
&= -\tfrac{1}{8}(F(A)_{ab}^{[ij]})^2+\tfrac{1}{2}i\bar\chi^i\not\partial\chi^i+\tfrac{1}{2}(\partial_aA)^2+\tfrac{1}{2}(\partial_aB)^2\\
&\quad+\tfrac{1}{4}(F^{(ij)})^2+\tfrac{1}{2}G^2.
\end{aligned} \tag{7.30}$$

The supersymmetry transformation rules for this multiplet are

$$\begin{aligned}
&\delta A_a^{[ij]} = 2\bar\epsilon^{[i}\gamma_5\gamma_a\chi^{j]},\\
&\delta\chi^i = iF(A)_{ab}^{[ij]}\gamma^5\sigma^{ab}\epsilon^j - i\not\partial(A+iB\gamma^5)\epsilon^i + F^{(ij)}\epsilon^j + iG\gamma^5\epsilon^i,\\
&\delta A = \bar\epsilon^i\chi^i,\\
&\delta B = i\bar\epsilon^i\gamma^5\gamma^i,\\
&\delta F^{(ij)} = -2i\bar\epsilon^{(i}\not\partial\chi^{j)},\\
&\delta G = \bar\epsilon^i\gamma^5\not\partial\chi^i.
\end{aligned} \tag{7.31}$$

It is interesting to note that this multiplet has the same Lagrangian and supersymmetry transformation rules as the one in (7.23)–(7.25) but that they have totally different superfields.

One can change the SO(2) tensor property of the component fields by attaching antisymmetric SO(2) indices on Φ_+:

$$\begin{aligned}
\Phi_+^{[ij]} &= \exp[-i(\theta^l\sigma^c\bar\theta^l)\partial_c](A^{[ij]}-iB^{[ij]}+i(\theta^k\theta^k)G^{[ij]}\\
&\quad+(\Theta^{(ik)}F^{(jk)}-\Theta^{(jk)}F^{(ik)})-4i(\theta^i\sigma^{ab}\theta^j)\partial_aA_b\\
&\quad+\tfrac{1}{2}(\theta^k\theta^k)^2\partial^2(A^{[ij]}+iB^{[ij]})+4\theta^{[i}\chi^{j]}\\
&\quad+4i(\theta^k\theta^k)\theta^{[i}\sigma\cdot\partial\bar\chi^{j]}).
\end{aligned} \tag{7.32}$$

The Lagrangian for this multiplet with 8 + 8 field components is

$$\begin{aligned}
L'_+ &= -\frac{1}{4}\int d^4\theta\,\Phi_+^{[ij]}\Phi_+^{[ij]}\\
&= -\tfrac{1}{4}(F(A)_{ab})^2+\tfrac{1}{2}i\bar\chi^i\not\partial\chi^i+\tfrac{1}{4}(\partial_aA^{[ij]})^2+\tfrac{1}{4}(\partial_aB^{[ij]})^2\\
&\quad+\tfrac{1}{4}(F^{(ij)})^2+\tfrac{1}{4}(G^{[ij]})^2.
\end{aligned} \tag{7.33}$$

This multiplet transforms under supersymmetry as follows:

$$\begin{aligned}
&\delta A_a = \bar\epsilon^i\gamma_5\gamma_a\chi^i,\\
&\delta\chi^i = -iF(A)_{ab}\gamma^5\sigma^{ab}\epsilon^i + i\not\partial(A^{[ij]}-iB^{[ij]}\gamma^5)\epsilon^j\\
&\qquad+(F^{(ij)}-iG^{[ij]}\gamma^5)\epsilon^j,\\
&\delta A^{[ij]} = 2\bar\epsilon^{[i}\chi^{j]},\\
&\delta B^{[ij]} = -2i\bar\epsilon^{[i}\gamma^5\chi^{j]},\\
&\delta F^{(ij)} = -2i\bar\epsilon^{(i}\not\partial\chi^{j)},\\
&\delta G^{[ij]} = 2\bar\epsilon^{[i}\gamma^5\not\partial\chi^{j]}.
\end{aligned} \tag{7.34}$$

This multiplet has the same Lagrangian and supersymmetry transformation rules as the one in (7.17)–(7.19).

The constraint (7.28) with minus sign leads to

$$\begin{aligned}
\Phi_- &= \exp[-i(\theta^l\sigma^c\bar\theta^l)\partial_c](A-iB+(\theta^k\theta^k)F+i\Theta^{(ij)}G^{(ij)}\\
&\quad+2(\theta^i\sigma^{ab}\theta^j)\partial_aV_b^{[ij]}-\tfrac{1}{2}(\theta^k\theta^k)^2\partial^2(A+iB)+2\theta^i\chi^i\\
&\quad-2i(\theta^k\theta^k)(\theta^i\sigma\cdot\partial\bar\chi^i)).
\end{aligned} \tag{7.35}$$

This 8 + 8 multiplet has the following Lagrangian:

$$\begin{aligned}
L_- &= \frac{1}{2}\int d^4\theta\,\Phi_-\Phi_-\\
&= -\tfrac{1}{8}(F(V)_{ab}^{[ij]})^2+\tfrac{1}{2}i\bar\chi^i\not\partial\chi^i+\tfrac{1}{2}(\partial_aA)^2+\tfrac{1}{2}(\partial_aB)^2\\
&\quad+\tfrac{1}{2}F^2+\tfrac{1}{4}(G^{(ij)})^2.
\end{aligned} \tag{7.36}$$

The supersymmetry transformation rules are

$$\begin{aligned}
&\delta V_a^{[ij]} = -2i\bar\epsilon^{[i}\gamma_a\chi^{j]},\\
&\delta\chi^i = F(V)_{ab}^{[ij]}\sigma^{ab}\epsilon^j - i\not\partial(A+iB\gamma^5)\epsilon^i + F\epsilon^i + iG^{(ij)}\gamma^5\epsilon^j,\\
&\delta A = \bar\epsilon^i\chi^i,\\
&\delta B = i\bar\epsilon^i\gamma^5\chi^i,\\
&\delta F = -i\bar\epsilon^i\not\partial\chi^i,\\
&\delta G^{(ij)} = 2i\bar\epsilon^{(i}\gamma^5\not\partial\chi^{j)}.
\end{aligned} \tag{7.37}$$

This multiplet has the same Lagrangian and supersymmetry transformation rules as the (0,0,0) multiplet.

Adding antisymmetric SO(2) indices to Φ_-, one has

$$\begin{aligned}
\Phi_-^{[ij]} &= \exp[-i(\theta^l\sigma^c\bar\theta^l)\partial_c](A^{[ij]}-iB^{[ij]}+(\theta^k\theta^k)F^{[ij]}\\
&\quad+i(\Theta^{(ik)}G^{(jk)}-\Theta^{(jk)}G^{(ik)})+4(\theta^i\sigma^{ab}\theta^j)\partial_aV_b\\
&\quad-\tfrac{1}{2}(\theta^k\theta^k)^2\partial^2(A^{[ij]}+iB^{[ij]})+4\theta^{[i}\chi^{j]}\\
&\quad-4i(\theta^k\theta^k)\theta^{[i}\sigma\cdot\partial\bar\chi^{j]}).
\end{aligned} \tag{7.38}$$

De Wit has called this $8 + 8$ multiplet the SO(2) vector gauge multiplet.[7] The Lagrangian for this vector gauge multiplet is

$$
\begin{aligned}
L'_- &= \frac{1}{4}\int d^4\theta\, \Phi_-^{[ij]}\Phi_-^{[ij]} \\
&= -\tfrac{1}{4}(F(V)_{ab})^2 + \tfrac{1}{2}i\bar{\chi}^i\partial\!\!\!/\chi^i + \tfrac{1}{4}(\partial_a A^{[ij]})^2 + \tfrac{1}{4}(\partial_a B^{[ij]})^2 \\
&\quad + \tfrac{1}{4}(F^{[ij]})^2 + \tfrac{1}{4}(G^{(ij)})^2.
\end{aligned} \tag{7.39}
$$

The supersymmetry transformation rules are

$$
\begin{aligned}
\delta V_a &= i\bar{\epsilon}^i\gamma_a\chi^i, \\
\delta\chi^i &= F(V)_{ab}\sigma^{ab}\epsilon^i + i\partial\!\!\!/(A^{[ij]} - iB^{[ij]}\gamma^5)\epsilon^j \\
&\quad - (F^{[ij]} - iG^{(ij)}\gamma^5)\epsilon^j, \\
\delta A^{[ij]} &= 2\bar{\epsilon}^{[i}\chi^{j]}, \\
\delta B^{[ij]} &= 2i\,\bar{\epsilon}^{[i}\gamma^5\chi^{j]}, \\
\delta F^{[ij]} &= -2i\,\bar{\epsilon}^{[i}\partial\!\!\!/\chi^{j]}, \\
\delta G^{(ij)} &= 2\bar{\epsilon}^{(i}\gamma^5\partial\!\!\!/\chi^{j)}.
\end{aligned} \tag{7.40}
$$

We now consider the chiral superfield (7.26). There are two ways to construct the invariant Lagrangian from this superfield. The first one is

$$
L_{\rm I} = 2\int d^8\theta\,\bar{\Phi}\Phi - \frac{1}{2}\Big(\Big(\int d^4\theta\,\Phi\,\partial^2\Phi + \text{h.c.}\Big). \tag{7.41}
$$

This Lagrangian has supersymmetric gauge invariance. It is invariant under

$$
\Phi \to \Phi + \Phi_+. \tag{7.42}
$$

Redefining some of the component fields as

$$
\begin{aligned}
T_{ab}^{[ij]} &\to T_{ab}^{[ij]} + i(\partial_a A_b^{[ij]} - \partial_b A_a^{[ij]}), \\
\psi &\to \psi + i\partial\!\!\!/\chi^i, \\
F &\to F + \partial^2 A, \\
G &\to G - \partial^2 B,
\end{aligned} \tag{7.43}
$$

$L_{\rm I}$ becomes

$$
\begin{aligned}
L_{\rm I} &= -\tfrac{1}{4}(\partial^b T_{ab}^{[ij]})^2 + \tfrac{1}{2}i\bar{\psi}^i\partial\!\!\!/\psi^i + \tfrac{1}{2}(\partial_a M)^2 + \tfrac{1}{4}(\partial_a N^{(ij)})^2 \\
&\quad + \tfrac{1}{2}(F^2 + G^2).
\end{aligned} \tag{7.44}
$$

In the Wess–Zumino gauge, the supersymmetry transformation rules are

$$
\begin{aligned}
\delta T_{ab}^{[ij]} &= -4\bar{\epsilon}^{[i}\sigma_{ab}\psi^{j]}, \\
\delta\psi^i &= i\partial^b T_{ab}^{[ij]}\gamma^a\epsilon^j - i\partial\!\!\!/(M\epsilon^i + iN^{(ij)}\gamma^5\epsilon^j) + (F + iG\gamma^5)\epsilon^i, \\
\delta M &= \bar{\epsilon}^i\psi^i, \\
\delta N^{(ij)} &= 2i\,\bar{\epsilon}^{(i}\gamma^5\psi^{j)}, \\
\delta F &= -i\,\bar{\epsilon}^i\partial\!\!\!/\psi^i, \\
\delta G &= \bar{\epsilon}^i\gamma^5\partial\!\!\!/\psi^i.
\end{aligned} \tag{7.45}
$$

De Wit has called this the SO(2) tensor gauge multiplet.[7]

The second invariant Lagrangian can be obtained by

$$
L_{\rm II} = 2\int d^8\theta\,\bar{\Phi}\Phi + \frac{1}{2}\Big(\Big(\int d^4\theta\,\Phi\partial^2\Phi + \text{h.c.}\Big). \tag{7.46}
$$

It is invariant under the supersymmetric gauge transformation

$$
\Phi \to \Phi + \Phi_-. \tag{7.47}
$$

With the field redefinition

$$
\begin{aligned}
T_{ab}^{[ij]} &\to T_{ab}^{[ij]} + \partial_a V_b^{[ij]} - \partial_b V_a^{[ij]}, \\
\psi &\to \psi - i\partial\!\!\!/\chi^i, \\
F &\to F - \partial^2 A, \\
G &\to G + \partial^2 B,
\end{aligned} \tag{7.48}
$$

$L_{\rm II}$ takes the following form:

$$
\begin{aligned}
L_{\rm II} &= -\tfrac{1}{4}(\partial^b \tilde{T}_{ab}^{[ij]})^2 + \tfrac{1}{2}i\bar{\psi}^i\partial\!\!\!/\psi^i + \tfrac{1}{4}(\partial_a M^{(ij)})^2 + \tfrac{1}{2}(\partial_a N)^2 \\
&\quad + \tfrac{1}{2}(F^2 + G^2).
\end{aligned} \tag{7.49}
$$

In the Wess–Zumino gauge, this multiplet transforms according to

$$
\begin{aligned}
\delta T_{ab}^{[ij]} &= -4\bar{\epsilon}^{[i}\sigma_{ab}\psi^{j]}, \\
\delta\psi^i &= -i\partial^b\tilde{T}_{ab}^{[ij]}\gamma^5\gamma^a\epsilon^j + i\partial\!\!\!/(M^{(ij)}\epsilon^j - iN\gamma^5\epsilon^i) \\
&\quad + (F + iG\gamma^5)\epsilon^i, \\
\delta M^{(ij)} &= -2\bar{\epsilon}^{(i}\psi^{j)}, \\
\delta N &= -i\,\bar{\epsilon}^i\gamma^5\psi^i, \\
\delta F &= -i\,\bar{\epsilon}^i\partial\!\!\!/\psi^i, \\
\delta G &= \bar{\epsilon}^i\gamma^5\partial\!\!\!/\psi^i.
\end{aligned} \tag{7.50}
$$

(IV) $\Phi_{(1/2,1,\pm 1)}$, $\Phi_{(1/2,-1,\pm 1)}$: The derivative terms of these superfields are partially factorizable, and the rest of them can be written in terms of three of $\theta_+{}^\alpha$, $\bar{\theta}_{-\dot{\alpha}}$, $\theta_-{}^\alpha$, and $\bar{\theta}_{+\dot{\alpha}}$. The final result is written in terms of $\theta^{\alpha i}$ and $\bar{\theta}_{\dot{\alpha}i}$ for manifest SO(2) invariance. The irreducible superfields are

$$
\begin{aligned}
\partial^2\Phi_{(1/2,-1,\pm 1)} &= \exp\big[-\tfrac{1}{2}i(\theta^k\sigma^c\bar{\theta}^k \pm i\epsilon^{kl}\theta^k\sigma^c\bar{\theta}^l)\partial_c\big]\Big(\overset{(A)}{\Theta}_\pm\,[A + iB \pm \tfrac{1}{2}i\epsilon^{ij}(A^{[ij]} + iB^{[ij]})] \\
&\quad + \tfrac{1}{2}\overset{(M)}{\Theta}_\pm\,[M + iN \pm \tfrac{1}{2}i\epsilon^{ij}(M^{[ij]} + iN^{[ij]})] + \tfrac{1}{2}\overset{(M)}{\Theta}{}^{(ij)}(1 \pm T)(M^{(ij)} + iN^{(ij)}) \\
&\quad + \tfrac{1}{2}i\overset{(A)}{\Theta}_\pm\,[A_a + iB_a \pm \tfrac{1}{2}i\epsilon^{ij}(A_a^{[ij]} + iB_a^{[ij]})] + \tfrac{1}{2}\overset{(V)}{\Theta}_\pm\,\partial^2[V_a + iW_a \pm \tfrac{1}{2}i\epsilon^{ij}(V_a^{[ij]} + iW_a^{[ij]})] \\
&\quad + i\overset{(C)}{\Theta}{}^{(ij)}(1 \pm T)(C^{(ij)} + iD^{(ij)}) + i\overset{(V)}{\Theta}{}^{a(ij)}(1 \pm T)(V_a^{(ij)} + iW_a^{(ij)}) + \tfrac{1}{4}\overset{(A)}{\Theta}{}^{a(ij)}(1 \pm T)\partial^2(A_a^{(ij)} + iB_a^{(ij)}) \\
&\quad + i\overset{(\rho)}{\Theta}{}^{\alpha i}(1 \pm T)(\sigma\cdot\partial\bar{\rho}^i)_\alpha + i\overset{(\psi)}{\Theta}{}^{\alpha i}(1 \mp T)(\sigma\cdot\partial\bar{\psi}^i)_\alpha + i\overset{(\phi)}{\Theta}{}_{\dot{\alpha}i}(1 \pm T)(\bar{\sigma}\cdot\partial\phi^i)^{\dot{\alpha}} + \overset{(\lambda)}{\Theta}{}^{\alpha i}(1 \pm T)\lambda_{\alpha i} \\
&\quad + i\overset{(\chi)}{\Theta}{}^{\alpha i}(1 \pm T)(\sigma\cdot\partial\bar{\chi}^i)_\alpha + i\overset{(\psi)}{\Theta}{}^{a\alpha i}(1 \pm T)\partial^2\psi_{a\alpha i} + i\overset{(\lambda)}{\Theta}{}_{\dot{\alpha}(ijk)}(1 \pm T)(\bar{\sigma}\cdot\partial\lambda^{(ijk)})^{\dot{\alpha}}\Big),
\end{aligned} \tag{7.51}
$$

where the SO(2) supertensors are

$$
\begin{aligned}
\overset{(A)}{\Theta}_{\pm} &= 1 - \tfrac{1}{16}(\theta^k\theta^k)(\bar\theta^l\bar\theta^l)\partial^2 \\
&\qquad + \tfrac{1}{16}(\theta^k\sigma^a\bar\theta^k)(\theta^l\sigma_a\bar\theta^l)\partial^2 \mp \tfrac{1}{8}i\epsilon^{kl}\Theta^{[ij]}\partial^2, \\
\overset{(M)}{\Theta}_{\pm} &= \theta^k\theta^k + \tfrac{1}{2}i(\theta^k\theta^k)(\theta^i\sigma^a\bar\theta^i \mp i\epsilon^{ij}\theta^i\sigma^a\bar\theta^j)\partial_a, \\
\overset{(M)}{\Theta}{}^{(ij)} &= \Theta^{(ij)} - \tfrac{1}{8}(\theta^k\theta^k)^2\bar\Theta^{(ij)}\partial^2, \\
\overset{(A)}{\Theta}_{\pm}{}^{a} &= \pm\epsilon^{ij}(\theta^i\sigma^{ab}\theta^j)\partial_b \\
&\qquad - \tfrac{1}{4}(\theta^k\theta^k)(\theta^i\sigma_b\bar\theta^i \mp i\epsilon^{ij}\theta^i\sigma_b\bar\theta^j)\partial^2\Pi^{ab}, \\
\overset{(V)}{\Theta}_{\pm}{}^{a} &= (\theta^i\sigma_b\bar\theta^i \mp i\epsilon^{ij}\theta^i\sigma_b\bar\theta^j)\Pi^{ab}, \\
\overset{(C)}{\Theta}{}^{(ij)} &= \theta^{(i}\sigma^a\bar\theta^{j)}\partial_a - \tfrac{1}{4}i(\theta^k\theta^k)\bar\Theta^{(ij)}\partial^2, \\
\overset{(V)}{\Theta}{}^{a(ij)} &= \theta^{(i}\sigma_b\bar\theta^{j)}\Pi^{ab} + i\bar\Theta^{(ik)}(\theta^k\sigma^{ab}\theta^j)\partial_b, \\
\overset{(A)}{\Theta}{}^{a(ij)} &= (\theta^k\theta^k)\theta^{(i}\sigma_b\bar\theta^{j)}\Pi^{ab}, \\
\overset{(\rho)}{\Theta}{}^{\alpha i} &= \theta^{\alpha i} - \tfrac{1}{8}(\theta^k\theta^k)(\bar\theta^l\bar\theta^l)\theta^{\alpha i}\partial^2 \qquad\qquad (7.52)\\
&\qquad + \tfrac{1}{8}(\theta^k\theta^k)(\theta^l\sigma^a\bar\theta^l)(\bar\theta^i\bar\sigma_a)^\alpha\partial^2, \\
\overset{(\psi)}{\Theta}{}^{\alpha i} &= \theta^{\alpha i} - \tfrac{1}{4}i(\theta^k\theta^k)(\bar\theta^i\bar\sigma\cdot\partial)^\alpha \\
&\qquad + \tfrac{1}{4}i(\theta^k\sigma^a\bar\theta^k)(\theta^i\sigma_a\bar\sigma\cdot\partial)^\alpha, \\
\overset{(\phi)}{\bar\Theta}_{\dot\alpha i} &= \bar\theta_{\dot\alpha i} - \tfrac{1}{4}i(\bar\theta^k\bar\theta^k)(\theta^i\sigma\cdot\partial)_{\dot\alpha} \\
&\qquad - \tfrac{1}{4}i(\theta^k\sigma^a\bar\theta^k)(\bar\theta^i\bar\sigma_a\sigma\cdot\partial)_{\dot\alpha}, \\
\overset{(\lambda)}{\Theta}{}^{\alpha i} &= \tfrac{1}{4}i(\theta^k\theta^k)(\bar\theta^i\bar\sigma\cdot\partial)^\alpha + \tfrac{1}{12}i(\theta^k\sigma^a\bar\theta^k)(\theta^i\sigma_a\bar\sigma\cdot\partial)^\alpha \\
&\qquad - \tfrac{1}{3}i(\theta^k\sigma^a\bar\theta^k)\theta^{\alpha i}\partial_a, \\
\overset{(\chi)}{\Theta}{}^{\alpha i} &= (\theta^k\theta^k)\theta^{\alpha i} + \tfrac{1}{4}i(\theta^k\theta^k)^2(\bar\theta^i\bar\sigma\cdot\partial)^\alpha, \\
\overset{(\lambda)}{\bar\Theta}_{\dot\alpha(ijk)} &= \bar\Theta_{\dot\alpha(ijk)} - \tfrac{1}{2}i(\theta^l\theta^l)(\Theta^{(ijk)}\sigma\cdot\partial)_{\dot\alpha}, \\
\overset{(\psi)}{\Theta}{}^{a\alpha i} &= (\theta^k\sigma_b\bar\theta^k)(\theta^i\Lambda^{ba})^\alpha.
\end{aligned}
$$

Here Λ^{ab} is the projection operator which extracts the spin-3/2 part from the vector spinor[11]:

$$
\begin{aligned}
\Lambda^{ab} &= \eta^{ab} - \frac{2}{3}\frac{\partial^a\partial^b}{\partial^2} - \frac{1}{3}\gamma^a\gamma^b \\
&\quad + \frac{1}{3}\frac{\partial\!\!\!/\gamma^b\partial^a}{\partial^2} - \frac{1}{3}\frac{\partial\!\!\!/\gamma^a\partial^b}{\partial^2},
\end{aligned}
\qquad (7.53)
$$

$$
\Phi_{(1/2,1,\pm 1)} = \bar\Phi_{(1/2,-1,\mp 1)}. \qquad (7.54)
$$

All these four superfields form a single multiplet with $64 + 64$ field components. Each supertensor in these superfields should be contracted with SO(2) irreducible component fields with at least four degrees of freedom for consistent supersymmetry transformation rules. All the Majorana spinors in these superfields are decomposed into four parts in an SO(2) invariant way, and each superfield contains one of them, i.e., $(1+T)\phi_{\alpha i}$, $(1-T)\phi_{\alpha i}$, $(1+T)\bar\phi_{\dot\alpha i}$, or $(1-T)\bar\phi_{\dot\alpha i}$. Since the vector spinor is accompanied by the spin-3/2 projection operator, one cannot obtain the Pauli–Fierz Lagrangian for the vector spinor with this multiplet alone. So, this multiplet should be coupled to other multiplets to yield the correct Lagrangian. Since this multiplet does not have direct physical interest, we will not consider it any further.

$(V)\,\Phi_{(1,0,0)}$: This multiplet has turned out to be the multiplet for the linearized SO(2) Weyl supergravity with $24 + 24$ field components. The irreducible superfield is

$$
\begin{aligned}
\partial^4\Phi_{(1,0,0)} &= 6\Theta D + (\Theta^{ab[ij]} + \bar\Theta^{ab[ij]})T^{[ij]}_{ab} + 4\Theta^a\partial^2 A_a \\
&\quad + 2\Theta^{a(ij)}\partial^2 A^{(ij)}_a + 2\Theta^{a[ij]}\partial^2 V^{[ij]}_a + \Theta^{ab}\partial^4 h_{ab} \\
&\quad + 6i\bar\Theta^i\bar\sigma\cdot\partial\chi^i + 6i\Theta^i\sigma\cdot\partial\bar\chi^i + 4\bar\Theta^{ai}\bar\sigma\cdot\partial\partial^2\psi_a{}^i \\
&\quad - 4\Theta^{ai}\sigma\cdot\partial\partial^2\bar\psi_a{}^i,
\end{aligned}
\qquad (7.55)
$$

where the SO(2) supertensors are

$$
\begin{aligned}
\Theta &= 1 + \tfrac{1}{6}(\theta^k\sigma^a\bar\theta^k)(\theta^l\sigma^b\bar\theta^l)\partial^2\Pi_{ab} + \tfrac{1}{64}(\theta^k\theta^k)^2(\bar\theta^l\bar\theta^l)^2\partial^4, \\
\Theta^{ab[ij]} &= (\eta^{ac}\eta^{bd} - \eta^{ad}\eta^{bc} - i\epsilon^{abcd}) \\
&\quad \times[(\bar\theta^i\bar\sigma_{de}\bar\theta^j)\partial_c\partial^e - \tfrac{1}{2}i(\bar\theta^k\bar\theta^k)\theta^{[i}\sigma_d\bar\theta^{j]}\partial^2\partial_c] \\
&\quad + \tfrac{1}{8}(\bar\theta^k\bar\theta^k)^2(\theta^i\sigma^{ab}\theta^j)\partial^4, \\
\Theta^a &= [\theta^k\sigma_b\bar\theta^k + \tfrac{1}{4}(\theta^k\theta^k)(\bar\theta^l\bar\theta^l)(\theta^i\sigma_b\bar\theta^i)\partial^2]\Pi^{ab}, \\
\Theta^{a(ij)} &= [\theta^{(i}\sigma_b\bar\theta^{j)} + \tfrac{1}{4}(\theta^k\theta^k)(\bar\theta^l\bar\theta^l)\theta^{(i}\sigma_b\bar\theta^{j)}\partial^2]\Pi^{ab} \\
&\quad + i[\bar\Theta^{(ik)}(\theta^k\sigma^{ab}\theta^j) - \Theta^{(ik)}(\bar\theta^k\bar\sigma^{ab}\bar\theta^j)]\partial_b, \qquad (7.56)\\
\Theta^{a[ij]} &= i[\theta^{[i}\sigma_b\bar\theta^{j]} - \tfrac{1}{4}(\theta^k\theta^k)(\bar\theta^l\bar\theta^l)\theta^{[i}\sigma_b\bar\theta^{j]}\partial^2]\Pi^{ab} \\
&\quad + \tfrac{1}{2}[(\bar\theta^k\bar\theta^k)(\theta^i\sigma^{ab}\theta^j) + (\theta^k\theta^k)(\bar\theta^i\bar\sigma^{ab}\bar\theta^j)]\partial_b, \\
\Theta^{ab} &= (\theta^k\sigma_c\bar\theta^k)(\theta^l\sigma_d\bar\theta^l)\Pi^{ab,cd}, \\
\Theta^{\alpha i} &= \theta^{\alpha i} - \tfrac{1}{3}i(\theta^k\sigma^a\bar\theta^k)\theta^{\alpha i}\partial_a + \tfrac{1}{3}i(\theta^k\sigma^a\bar\theta^k)(\theta^i\sigma_a\bar\sigma\cdot\partial)^\alpha \\
&\quad + \tfrac{1}{4}(\theta^k\theta^k)(\bar\theta^l\bar\theta^l)\theta^{\alpha i}\partial^2 - \tfrac{1}{6}(\theta^k\theta^k)(\theta^l\sigma^a\bar\theta^l)(\bar\theta^i\bar\sigma\cdot\partial)^\alpha\partial_a \\
&\quad + \tfrac{1}{6}(\theta^k\theta^k)(\theta^l\sigma^a\bar\theta^l)(\bar\theta^i\bar\sigma_a)^\alpha\partial^2 \\
&\quad + \tfrac{1}{16}i(\theta^k\theta^k)^2(\bar\theta^l\bar\theta^l)(\bar\theta^i\bar\sigma\cdot\partial)^\alpha\partial^2, \\
\Theta^{a\alpha i} &= (\theta^k\sigma_b\bar\theta^k)(\theta^i\Lambda^{ba})^\alpha + \tfrac{1}{2}i(\theta^k\theta^k)(\theta^l\sigma_b\bar\theta^l)(\bar\theta^i\bar\sigma\cdot\partial\Lambda^{ba})^\alpha.
\end{aligned}
$$

Here $\Pi^{ab,cd}$ is the projection operator for spin 2 [11]:

$$
\Pi^{ab,cd} = \tfrac{1}{2}(\Pi^{ac}\Pi^{bd} + \Pi^{ad}\Pi^{bc}) - \tfrac{1}{3}\Pi^{ab}\Pi^{cd}. \qquad (7.57)
$$

The Lagrangian corresponding to this multiplet is

$$
\begin{aligned}
L_{(1,0,0)} &= \int d^8\theta\, \Phi_{(1,0,0)}\partial^4\Phi_{(1,0,0)} \\
&= \tfrac{1}{4}h_{ab}\partial^4\Pi^{ab,cd}h_{cd} + i\bar\psi_a{}^i\partial^2\partial\!\!\!/\Lambda^{ab}\psi_b{}^i - (F(A)_{ab})^2 \\
&\quad - \tfrac{1}{4}(F(A)^{(ij)}_{ab})^2 - \tfrac{1}{4}(F(V)^{[ij]}_{ab})^2 - \tfrac{1}{2}[(\partial^b T^{[ij]}_{ab})^2 \\
&\quad + (\partial^b\bar T^{[ij]}_{ab})^2] + 3i\bar\chi^i\partial\!\!\!/\chi^i + 3D^2.
\end{aligned}
\qquad (7.58)
$$

This Lagrangian has high degrees of invariance. It is invariant under

$$\delta h_{ab} = \eta_{ab}\Psi + \partial_a\xi_b + \partial_b\xi_a,$$
$$\delta\psi_a{}^i = \partial_a\epsilon^i + \gamma_a\lambda^i,$$
$$\delta A_a = \partial_a\Lambda, \tag{7.59}$$
$$\delta A_a^{(ij)} = \partial_a\Lambda^{(ij)},$$
$$\delta V_a^{[ij]} = \partial_a\Lambda^{[ij]}.$$

In order to identify the Lagrangian (7.58) as the linearized SO(2) Weyl supergravity Lagrangian, we expand the connection and the Ricci tensor up to order κ:

$$\omega_{abc} = -\tfrac{1}{2}\kappa(\partial_b h_{ac} - \partial_c h_{ab}), \tag{7.60}$$
$$R_{ab} = \tfrac{1}{2}\kappa(\partial_a\partial^c h_{bc} + \partial_b\partial^c h_{ac} - \partial^2 h_{ab} - \partial_a\partial_b h). \tag{7.61}$$

For the Rarita–Schwinger fields, we define the analog of field strength tensor as

$$R^{ai} = \epsilon^{abcd}\gamma_5\gamma_b\partial_c\psi_d{}^i. \tag{7.62}$$

The Lagrangian can now be rewritten up to order κ^2 as

$$L_W = (1/\kappa^2)(R_{mn}{}^2 - \tfrac{1}{3}R^2) + \epsilon^{mnpq}(\overline{R_m{}^i - \tfrac{1}{3}\gamma_m\gamma\cdot R^i})\gamma_5\gamma_n D_p(R_q{}^i - \tfrac{1}{3}\gamma_q\gamma\cdot R^i) - (F(A)_{mn})^2 - \tfrac{1}{4}(F(A)_{mn}^{(ij)})^2 - \tfrac{1}{4}(F(V)_{mn}^{[ij]})^2 - \tfrac{1}{2}[(\partial^n T_{mn}^{[ij]})^2 + (\partial^n \widetilde{T}_{mn}^{[ij]})^2] + 3i\bar{\chi}^i\not{D}\chi^i + 3D^2, \tag{7.63}$$

where D_m is the covariant derivative

$$D_m = \partial_m + \tfrac{1}{2}\omega_{mab}\sigma^{ab}. \tag{7.64}$$

Combining the spinor gauge transformation $\delta\psi_m{}^i = \partial_m\epsilon^i$ with the global supersymmetry transformation, one obtains the following supersymmetry transformation rules:

$$\begin{aligned}
\delta e^a{}_m &= -i\kappa\bar{\epsilon}^i\gamma^a\psi_m{}^i,\\
\delta\psi_m{}^i &= (2/\kappa)D_m\epsilon^i - \tfrac{1}{2}iT_{pq}^{[ij]}\sigma^{pq}\gamma_m\epsilon^j\\
&\quad - iA_m\gamma_5\epsilon^i + (V_m^{[ij]} + iA_m^{(ij)}\gamma_5)\epsilon^j,\\
\delta A_m &= -\tfrac{1}{2}i\,\bar{\epsilon}^i\gamma_5(R_m{}^i - \tfrac{1}{3}\gamma_m\gamma\cdot R^i) - \bar{\epsilon}^i\gamma_5\gamma_m\chi^i,\\
\delta A_m^{(ij)} &= 2i\,\bar{\epsilon}^{(i}\gamma_5(R_m{}^{j)} - \tfrac{1}{3}\gamma_m\gamma\cdot R^{j)}) - 2\bar{\epsilon}^{(i}\gamma_5\gamma_m\chi^{j)},\\
\delta V_m^{[ij]} &= -2\bar{\epsilon}^{[i}(R_m{}^{j]} - \tfrac{1}{3}\gamma_m\gamma\cdot R^{j]}) - 2i\,\bar{\epsilon}^{[i}\gamma_m\chi^{j]},\\
\delta T_{mn}^{[ij]} &= 4\bar{\epsilon}^{[i}\sigma^{pq}\sigma_{mn}\partial_p\psi_q{}^{j]}\\
&\quad + \tfrac{1}{3}i\,\bar{\epsilon}^{[i}\sigma_{mn}\gamma\cdot R^{j]} + 4\bar{\epsilon}^{[i}\sigma_{mn}\chi^{j]},\\
\delta\chi^i &= -\not{D}\epsilon^i - \tfrac{1}{3}i\partial_p T_{mn}^{[ij]}\sigma^{mn}\gamma^p\epsilon^j + \tfrac{2}{3}iF(A)_{mn}\gamma^5\sigma^{mn}\epsilon^i\\
&\quad + \tfrac{1}{3}(F(V)_{mn}^{[ij]} + iF(A)_{mn}^{(ij)}\gamma^5)\sigma^{mn}\epsilon^j,\\
\delta D &= i\,\bar{\epsilon}^i\not{\partial}\chi^i.
\end{aligned} \tag{7.65}$$

8. LINEARIZED SO(2) SUPERGRAVITY

We have seen that the superspin 1 multiplet is the linearized SO(2) Weyl supergravity multiplet in Sec. 7. Its Lagrangian contains a fourth-order derivative for the spin-2 field and a third-order derivative for the spin-3/2 field. In order to get the Pauli–Fierz Lagrangians for these fields and to make the auxiliary fields nonpropagating, one should insert ∂^2 in (7.58) instead of ∂^4. This would produce some nonlocal terms which can be cancelled out by introducing compensating superfields.[12] Due to those compensating superfields, part of the invariance for the physical fields and all for the auxiliary fields are lost. Before breaking the invariances that will be lost eventually by the compensating superfields, one should add gauge terms in the supersymmetry transformation for consistency.

The full 40 + 40 multiplet consists of three irreducible superfields which are the SO(2) Weyl supergravity multiplet, the SO(2) tensor gauge multiplet, and the SO(2) vector gauge multiplet.[7] With proper field redefinitions for cancellation of nonlocal terms, the irreducible superfields become

$$\begin{aligned}
\partial^4\Phi_W &= 6\Theta(\partial\cdot V - (1/3\kappa)R)\\
&\quad + (\Theta^{ab[ij]} + \overline{\Theta}^{ab[ij]})(t_{ab}^{[ij]} - F(B)_{ab}^{[ij]})\\
&\quad + 2\Theta^a\partial^2 A_a + 2\Theta^{a(ij)}\partial^2 A_a^{(ij)}\\
&\quad + 2\Theta^{a[ij]}\partial^2 V_a^{[ij]} + \Theta^{ab}\partial^4 h_{ab}\\
&\quad + 3i\overline{\Theta}^i\bar{\sigma}\cdot\partial(\chi^i + i\sigma\cdot\partial\bar{\lambda}^i - \tfrac{2}{3}i\sigma\cdot\overline{R}^i)\\
&\quad + 3i\Theta^i\sigma\cdot\partial(\bar{\chi}^i + i\bar{\sigma}\cdot\partial\lambda^i - \tfrac{2}{3}i\bar{\sigma}\cdot R^i)\\
&\quad + 4\overline{\Theta}^{ai}\bar{\sigma}\cdot\partial\partial^2\Lambda_a{}^b\psi_b{}^i - 4\Theta^{ai}\sigma\cdot\partial\partial^2\overline{\Lambda}_a{}^b\bar{\psi}_b{}^i,
\end{aligned} \tag{8.1}$$

$$\begin{aligned}
\Phi_T &= \exp[-i(\theta^k\sigma^a\bar{\theta}^k)\partial_a][A + iB + (\theta^k\theta^k)(S + iP)\\
&\quad + \Theta^{(ij)}(S^{(ij)} + iP^{(ij)}) - (\theta^i\sigma^{ab}\theta^j)(t_{ab}^{[ij]} - \tfrac{1}{2}F(B)_{ab}^{[ij]})\\
&\quad + \tfrac{1}{2}(\theta^k\theta^k)^2[-2\partial\cdot V + (1/\kappa)R\\
&\quad - i\partial\cdot A + \partial^2(A - iB)] + 2\theta^i\phi^i\\
&\quad + 2(\theta^k\theta^k)\theta^i(\chi^i + i\sigma\cdot\partial\lambda^i - i\sigma\cdot\overline{R}^i + i\sigma\cdot\partial\bar{\phi}^i)],
\end{aligned} \tag{8.2}$$

$$\begin{aligned}
\Phi_V^{[ij]} &= \exp[-i(\theta^k\sigma^a\bar{\theta}^k)\partial_a](2(M^{[ij]} - iN^{[ij]})\\
&\quad + 2(\theta^k\theta^k)\partial\cdot V^{[ij]} - 2i(\Theta^{(ik)}\partial\cdot A^{(jk)} - \Theta^{(jk)}\partial\cdot A^{(ik)})\\
&\quad + 8(\theta^i\sigma^{ab}\theta^j)\partial_a V_b\\
&\quad - (\theta^k\theta^k)^2\partial^2(M^{[ij]} + iN^{[ij]}) + 4\theta^{[i}(\chi^{j]} - i\sigma\cdot\bar{\lambda}^{j]})\\
&\quad - 4i(\theta^k\theta^k)\theta^{[i}\sigma\cdot\partial(\bar{\chi}^{j]} - i\sigma\cdot\bar{\lambda}^{j]})).
\end{aligned} \tag{8.3}$$

The linearized SO(2) Poincaré supergravity Lagrangian is given by

$$\begin{aligned}
L_p &= -\int d^8\theta\,\Phi_W\partial^2\Phi_W + 2\Big(\int d^8\theta\,\overline{\Phi}_T\frac{1}{\partial^2}\Phi_T\\
&\quad - \frac{1}{4}\int d^4\theta\,\Phi_T\Phi_T - \frac{1}{4}\int d^4\bar{\theta}\,\overline{\Phi}_T\overline{\Phi}_T\Big)\\
&\quad + \frac{1}{8}\int d^4\theta\,\Phi_V^{[ij]}\frac{1}{\partial^2}\Phi_V^{[ij]}\\
&= \tfrac{1}{4}(\partial_c h_{ab})^2 - \tfrac{1}{4}(\partial_a h)^2 - \tfrac{1}{2}(\partial^b h_{ab})^2 + \tfrac{1}{2}(\partial^b h_{ab})(\partial^a h)\\
&\quad + \epsilon^{abcd}\bar{\psi}_a{}^i\gamma_5\gamma_b\partial_c\psi_d{}^i - \tfrac{1}{8}(F(B)_{ab}^{[ij]})^2 + \tfrac{1}{4}(t_{ab}^{[ij]})^2\\
&\quad - \tfrac{1}{2}(A_a)^2 - \tfrac{1}{2}(A_a^{(ij)})^2 - \tfrac{1}{2}(V_a^{[ij]})^2 + (V_a)^2 + \bar{\chi}^i\lambda^i\\
&\quad - \tfrac{1}{2}S^2 - \tfrac{1}{2}(P^{(ij)})^2 - \tfrac{1}{2}(M^{[ij]})^2 - \tfrac{1}{2}(N^{[ij]})^2.
\end{aligned} \tag{8.4}$$

Ignoring perfect divergences, L_p can be written in terms of the Ricci scalar as follows:

$$L_p = -(1/\kappa^2)eR + \bar\psi_m{}^i R^{mi} - \tfrac18(F(B)^{[ij]}_{mn})^2$$
$$+ \tfrac14(t^{[ij]}_{mn})^2 - \tfrac12(A_m)^2$$
$$- \tfrac12(A^{(ij)}_m)^2 - \tfrac12(V^{[ij]}_m)^2$$
$$+ (V_m)^2 + \bar\chi^i\lambda^i - \tfrac12 S^2 - \tfrac14(P^{(ij)})^2$$
$$- \tfrac12(M^{[ij]})^2 - \tfrac12(N^{[ij]})^2. \tag{8.5}$$

The supersymmetry transformation rules for linearized SO(2) supergravity are

$$\begin{aligned}
\delta e^a{}_m &= -i\kappa\bar\epsilon^i\gamma^a\psi_m{}^i,\\
\delta\psi_m{}^i &= (2/\kappa)D_m\epsilon^i - \tfrac12 i(t^{[ij]}_{pq} - F(B)^{[ij]}_{pq})\sigma^{pq}\gamma_m\epsilon^j\\
&\quad - \tfrac12 iA_m\gamma_5\epsilon^i + (V^{[ij]}_m + iA^{(ij)}_m\gamma_5)\epsilon^j\\
&\quad + \tfrac12 i\gamma_m(S\epsilon^i + iP^{(ij)}\gamma^5\epsilon^j + \not{A}\gamma^5\epsilon^i - t^{[ij]}_{pq}\sigma^{pq}\epsilon^j),\\
\delta B^{[ij]}_m &= 4\bar\epsilon^{[i}\psi_m{}^{j]},\\
\delta t^{[ij]}_{mn} &= -2\epsilon_{mnpq}\bar\epsilon^{[i}\gamma^5\gamma^p R^{qj]} + 2\epsilon^{[i}\sigma_{mn}(\chi^{j]} + i\not\partial\lambda^{j]}),\\
\delta A_m &= -i\bar\epsilon^i\gamma_5 R_m{}^i - \bar\epsilon^i\gamma_5\gamma_m(\chi^i + i\not\partial\lambda^i - i\gamma\cdot R^i),\\
\delta A^{(ij)}_m &= 2i\bar\epsilon^{(i}\gamma_5 R_m{}^{j)} - \bar\epsilon^{(i}\gamma_5(\gamma_m\chi^{j)} - i\not\partial\gamma_m\lambda^{j)}),\\
\delta V^{[ij]}_m &= -2\bar\epsilon^{[i}R_m{}^{j]} - i\bar\epsilon^{[i}(\gamma_m\chi^{j]} - i\not\partial\gamma_m\lambda^{j]}),\\
\delta V_m &= \tfrac12 i\bar\epsilon^i(\gamma_m\chi^i + i\not\partial\gamma_m\lambda^i),\\
\delta\chi^i &= i\not\partial(S\epsilon^i + iP^{(ij)}\gamma^5\epsilon^j + \not{A}\gamma^5\epsilon^i - t^{[ij]}_{mn}\sigma^{mn}\epsilon^j + M^{[ij]}\epsilon^j\\
&\quad + iN^{[ij]}\gamma^5\epsilon^j) - \partial_m\not{V}\gamma^m\epsilon^i\\
&\quad - \partial_m(\not{V}^{[ij]} - i\not{A}^{(ij)}\gamma^5)\gamma^m\epsilon^j,\\
\delta\lambda^i &= -t^{[ij]}_{mn}\sigma^{mn}\epsilon^j + i\not{V}\epsilon^i + \not{A}\gamma^5\epsilon^i - i(\not{V}^{[ij]} + i\not{A}^{(ij)}\gamma^5)\epsilon^j\\
&\quad - (M^{[ij]} + iN^{[ij]}\gamma^5)\epsilon^j + S\epsilon^i + iP^{(ij)}\gamma^5\epsilon^j,\\
\delta S &= \bar\epsilon^i(\chi^i + i\not\partial\lambda^i - i\gamma\cdot R^i),\\
\delta P^{(ij)} &= 2i\bar\epsilon^{(i}\gamma^5(\chi^{j)} + i\not\partial\lambda^{j)} - i\gamma\cdot R^{j)}),\\
\delta M^{[ij]} &= \bar\epsilon^{[i}(\chi^{j]} - i\not\partial\lambda^{j]}),\\
\delta N^{[ij]} &= i\bar\epsilon^{[i}\gamma^5(\chi^{j]} - i\not\partial\lambda^{j]}).
\end{aligned}\tag{8.6}$$

One might expect that this method can be extended to higher N. However, Taylor[13] has proved a no-go theorem that prevents one from constructing supergravity with higher N without off-shell central charges. So, it is essential to introduce off-shell central charges to construct supergravity with higher N.

ACKNOWLEDGMENT

I would like to thank Professor Robert J. Finkelstein for reading the manuscript, numerous discussions, and continuous encouragement.

APPENDIX A: NOTATIONS

Latin letters $a,b,c,\ldots$ are used for Lorentz indices, $m,n,p,\ldots$ for curved space indices, $i,j,k,\ldots$ for SO(2) indices, and Greek letters for spinor indices. Bars denote Hermitian conjugation.

$$\eta_{ab} = \mathrm{diag}(1,-1,-1,-1), \quad \epsilon_{0123} = 1; \tag{A1}$$

$$\epsilon_{\alpha\beta} = \epsilon^{\alpha\beta} = \epsilon_{\dot\alpha\dot\beta} = \epsilon^{\dot\alpha\dot\beta} = \begin{pmatrix} 0 & 1\\ -1 & 0\end{pmatrix}; \tag{A2}$$

$$\epsilon^{ij} = \begin{pmatrix} 0 & 1\\ -1 & 0\end{pmatrix}; \tag{A3}$$

$$\sigma^a_{\alpha\dot\alpha} = (1, -\boldsymbol\sigma), \quad \bar\sigma^{a\dot\alpha\alpha} = (1, \boldsymbol\sigma); \tag{A4}$$

$$\begin{aligned}
\theta_{\alpha i} &= \theta^{\beta i}\epsilon_{\beta\alpha}, \quad \theta^{\alpha i} = \epsilon^{\alpha\beta}\theta_{\beta i};\\
\bar\theta_{\dot\alpha i} &= \bar\theta^{\dot\beta i}\epsilon_{\dot\beta\dot\alpha}, \quad \bar\theta^{\dot\alpha i} = \epsilon^{\dot\alpha\dot\beta}\bar\theta_{\dot\beta i},
\end{aligned}\tag{A5}$$

$$\theta^k\theta^k = \theta^{\alpha k}\theta_{\alpha k}, \quad \bar\theta^k\bar\theta^k = \bar\theta_{\dot\alpha k}\bar\theta^{\dot\alpha k}; \tag{A6}$$

$$\begin{aligned}
(\sigma_{ab})_\alpha{}^\beta &= \tfrac14(\sigma_{a\alpha\dot\alpha}\bar\sigma_b^{\dot\alpha\beta} - \sigma_{b\alpha\dot\alpha}\bar\sigma_a^{\dot\alpha\beta}),\\
(\bar\sigma_{ab})^{\dot\alpha}{}_{\dot\beta} &= \tfrac14(\bar\sigma_a^{\dot\alpha\alpha}\sigma_{b\alpha\dot\beta} - \bar\sigma_b^{\dot\alpha\alpha}\sigma_{a\alpha\dot\beta});
\end{aligned}\tag{A7}$$

$$\gamma_a = \begin{pmatrix} 0 & \sigma_a\\ \bar\sigma_a & 0\end{pmatrix}, \quad \gamma^5 = \begin{pmatrix} 1 & 0\\ 0 & -1\end{pmatrix},$$
$$\sigma_{ab} = \begin{pmatrix} \sigma_{ab} & 0\\ 0 & \bar\sigma_{ab}\end{pmatrix}; \tag{A8}$$

$$\text{Majorana spinors:} \quad \chi^i = \begin{pmatrix} \chi_{\alpha i}\\ \bar\chi^{\dot\alpha i}\end{pmatrix}, \quad \bar\chi^i(\chi^{\alpha i}\bar\chi_{\dot\alpha i}); \tag{A9}$$

$$\begin{aligned}
X^{(i}Y^{j)} &= \tfrac12(X^iY^j + X^jY^i - \tfrac12\delta^{ij}X^kY^k),\\
X^{[i}Y^{j]} &= \tfrac12(X^iY^j - X^jY^i);
\end{aligned}\tag{A10}$$

$$\partial_{\alpha i}\theta^{\beta i} = \delta_\alpha{}^\beta\delta^{ij}, \quad \bar\partial^{\dot\alpha i}\bar\theta_{\dot\beta j} = \delta_{\dot\beta}{}^{\dot\alpha}\delta^{ij}; \tag{A11}$$

$$\int d^4\theta\,(\theta^k\theta^k)^2 = \int d^4\bar\theta\,(\bar\theta^k\bar\theta^k)^2 = 1. \tag{A12}$$

APPENDIX B: SUPERTENSOR REDUCTION FORMULAS

Here we list some of the reduction formulas which are frequently used in explicit calculations. Most of the formulas not listed here can be derived by taking Hermitian conjugation of the formulas listed here or applying some of them successively.

$$\theta^{\alpha i}\theta_{\beta j} = \tfrac12\delta_\beta{}^\alpha\Theta^{(ij)} + \tfrac14\delta_\beta{}^\alpha\delta^{ij}(\theta^k\theta^k) - \tfrac12(\sigma_{ab})_\beta{}^\alpha(\theta^i\sigma^{ab}\theta^j); \tag{B1}$$

$$\theta^{\alpha i}\bar\theta^{\dot\alpha j} = \tfrac12\bar\sigma_a^{\dot\alpha\alpha}(\theta^{(i}\sigma^a\bar\theta^{j)} + \theta^{[i}\sigma^a\theta^{j]} + \tfrac12\delta^{ij}(\theta^k\sigma^a\bar\theta^k)); \tag{B2}$$

$$\Theta^{(ij)}\theta^{\alpha k} = \tfrac12(\delta^{ij}\delta^{kl} - \delta^{ik}\delta^{jl} - \delta^{il}\delta^{jk})(\theta^m\theta^m)\theta^{\alpha l}; \tag{B3}$$

$$\begin{aligned}
\bar\Theta^{(ij)}\theta^{\alpha k} &= \theta^{\alpha(ijk)} + \tfrac14(\delta^{ij}\delta^{kl} - \delta^{ik}\delta^{jl} - \delta^{il}\delta^{jk})(\bar\theta^m\bar\theta^m)\theta^{\alpha l}\\
&\quad + \tfrac14\bar\sigma_a^{\dot\alpha\alpha}(\delta^{ij}\delta^{kl} - \delta^{ik}\delta^{jl} - \delta^{il}\delta^{jk})(\theta^m\sigma^a\bar\theta^m)\bar\theta_{\dot\alpha l};
\end{aligned}\tag{B4}$$

$$(\theta^i\sigma^{ab}\theta^j)\theta^{\alpha k} = \tfrac12(\sigma_{ab})_\beta{}^\alpha(\delta^{ik}\delta^{jl} - \delta^{il}\delta^{jk})(\theta^k\theta^k)\theta^{\beta l}; \tag{B5}$$

$$(\theta^i\sigma^{ab}\theta^j)\bar\theta_{\dot\alpha k} = -\tfrac12(\sigma^{ab}\sigma^c)_{\alpha\dot\alpha}(\delta^{ik}\delta^{jl} - \delta^{il}\delta^{jk})(\theta^m\sigma_c\bar\theta^m)\theta^{\alpha l}; \tag{B6}$$

$$\begin{aligned}
(\theta^i\sigma^a\bar\theta^j)\bar\theta_{\dot\alpha k} &= -\tfrac12\sigma^a_{\alpha\dot\alpha}\Theta^{\alpha(ijk)}\\
&\quad + \tfrac18\sigma^a_{\alpha\dot\alpha}(\delta^{ij}\delta^{kl} + \delta^{ik}\delta^{jl} - 3\delta^{il}\delta^{jk})(\bar\theta^m\bar\theta^m)\theta^{\alpha l}\\
&\quad + (\delta^{ij}\delta^{kl} - \delta^{ik}\delta^{jl})(\theta^m\sigma^a\bar\theta^m)\bar\theta_{\dot\alpha l}\\
&\quad - \tfrac18(\bar\sigma^b\sigma^a)^{\dot\beta}{}_{\dot\alpha}(\delta^{ij}\delta^{kl} - 3\delta^{ik}\delta^{jl}\\
&\quad + \delta^{il}\delta^{jk})(\theta^m\sigma_b\bar\theta^m)\bar\theta_{\dot\beta l};
\end{aligned}\tag{B7}$$

$$\Theta^{\alpha(ijk)}\theta_{\beta l} = \tfrac{1}{2}\delta_\beta{}^\alpha \Theta^{(ijkl)} - \tfrac{1}{12}(\sigma_{ab})_\beta{}^\alpha[\delta^{ij}\delta^{km}\delta^{ln} - 2\delta^{im}\delta^{jl}\delta^{kn} + \mathrm{sym}(ijk)]\bar\Theta^{(mp)}(\theta^p\sigma^{ab}\theta^n) - \tfrac{1}{24}\delta_\beta{}^\alpha[\delta^{ij}\delta^{km}\delta^{ln} - 2\delta^{il}\delta^{jm}\delta^{kn} + \mathrm{sym}(ijk)](\theta^p\theta^p)\bar\Theta^{(mn)}; \qquad \text{(B8)}$$

$$\Theta^{\alpha(ijk)}\bar\theta^{\dot\alpha l} = \tfrac{1}{12}\bar\sigma_a^{\dot\alpha\alpha}[\delta^{ij}\delta^{km}\delta^{ln} - 2\delta^{il}\delta^{jm}\delta^{kn} + \mathrm{sym}(ijk)] \times(\bar\theta^p\bar\theta^p)\theta^{(m}\sigma^a\bar\theta^{n)}; \qquad \text{(B9)}$$

$$\Theta^{(ijkl)}\theta^{\alpha m} = -\tfrac{1}{4}(\theta^n\theta^n)[\delta^{im}\Theta^{\alpha(jkl)} + \mathrm{sym}(ijkl) - \mathrm{tr}(ijkl)]; \qquad \text{(B10)}$$

$$\Theta^{[ij]}\theta^{\alpha k} = \tfrac{1}{4}(\delta^{ik}\delta^{jl} - \delta^{il}\delta^{jk})[(\theta^m\theta^m)(\bar\theta^n\bar\theta^n)\theta^{\alpha l} + \bar\sigma_a^{\dot\alpha\alpha}(\theta^m\theta^m)(\theta^n\sigma^a\bar\theta^n)\bar\theta_{\dot\alpha l}]; \qquad \text{(B11)}$$

$$\begin{aligned}(\theta^k\sigma^a\bar\theta^k)(\theta^l\sigma^b\bar\theta^l)\theta^{\alpha i} &= \tfrac{1}{2}(\eta^{ac}\eta^{bd} + \eta^{ad}\eta^{bc})\bar\sigma_d^{\dot\alpha\alpha}(\theta^k\theta^k)(\theta^l\sigma_c\bar\theta^l)\bar\theta_{\dot\alpha i}\\ &\quad + \tfrac{1}{2}\eta^{ab}(\theta^k\theta^k)(\bar\theta^l\bar\theta^l)\theta^{\alpha i}; \qquad \text{(B12)}\end{aligned}$$

$$\Theta^{(ij)}\Theta^{(kl)} = -\tfrac{1}{4}(\delta^{ik}\delta^{jl} + \delta^{il}\delta^{jk} - \delta^{ij}\delta^{kl})(\theta^m\theta^m)^2; \qquad \text{(B13)}$$

$$\begin{aligned}\Theta^{(ij)}\bar\Theta^{(kl)} &= \Theta^{(ijkl)} - \tfrac{1}{8}(\delta^{ik}\delta^{jl} + \delta^{il}\delta^{jk} - \delta^{ij}\delta^{kl})\\ &\quad\times[(\theta^m\theta^m)(\bar\theta^n\bar\theta^n) - (\theta^m\sigma^a\bar\theta^m)(\theta^n\sigma_a\bar\theta^n)]\\ &\quad+\tfrac{1}{4}[\delta^{ik}\delta^{jm}\delta^{ln} + \mathrm{sym}(ij)(kl)]\Theta^{[mn]}; \qquad \text{(B14)}\end{aligned}$$

$$\Theta^{(ij)}(\bar\theta^k\bar\sigma^{ab}\bar\theta^l) = (\delta^{ik}\delta^{jm}\delta^{ln} - \delta^{il}\delta^{jm}\delta^{kn})\theta^{(mp)}(\bar\theta^p\bar\sigma^{ab}\bar\theta^l); \qquad \text{(B15)}$$

$$\Theta^{(ij)}(\theta^k\sigma^a\bar\theta^l) = \tfrac{1}{2}(\delta^{ij}\delta^{km} - \delta^{ik}\delta^{jm} - \delta^{im}\delta^{jk})(\theta^n\theta^n)\theta^{(m}\sigma^a\bar\theta^{l)}; \qquad \text{(B16)}$$

$$\begin{aligned}&\Theta^{(ij)}\theta^{\alpha(klm)}\\ &= \tfrac{1}{8}[\delta^{ik}(\theta^n\theta^n)\Theta^{\alpha(jlm)} + \mathrm{sym}(ij)(klm) - \mathrm{tr}(ij)(klm)]\\ &\quad + \tfrac{1}{12}\{\delta^{ik}\delta^{jl}[(\theta^n\theta^n)(\bar\theta^p\bar\theta^p)\theta^{\alpha m}\\ &\quad + \bar\sigma_a^{\dot\alpha\alpha}(\theta^n\theta^n)(\theta^p\sigma^a\bar\theta^p)\bar\theta_{\dot\alpha m}]\\ &\quad + \mathrm{sym}(ij)(klm) - \mathrm{tr}(ij)(klm)\}; \qquad \text{(B17)}\end{aligned}$$

$$\begin{aligned}&\Theta^{(mn)}\Theta^{(ijkl)}\\ &= \tfrac{1}{72}(\theta^p\theta^p)^2[(\delta^{kl}\delta^{mn} - 3\delta^{km}\delta^{ln} - 3\delta^{kn}\delta^{lm})\bar\Theta^{(ij)}\\ &\quad + \mathrm{sym}(ijkl)] + [(\delta^{jk}\delta^{ln} + \delta^{jl}\delta^{km} + \delta^{il}\delta^{jk})\bar\Theta^{(im)}\\ &\quad + \mathrm{sym}(mn)(ijkl)] - (\delta^{ij}\delta^{kl} + \delta^{ik}\delta^{jl} + \delta^{il}\delta^{jk})\bar\Theta^{(mn)}; \qquad \text{(B18)}\end{aligned}$$

$$\Theta^{(ij)}(\theta^k\sigma^a\bar\theta^k)(\theta^l\sigma^b\bar\theta^l) = -\tfrac{1}{4}\eta^{ab}(\theta^m\theta^m)^2\bar\Theta^{(ij)}; \qquad \text{(B19)}$$

$$\begin{aligned}&\Theta^{(ij)}\Theta^{[kl]}\\ &= -\tfrac{1}{8}(\delta^{ik}\delta^{jm}\delta^{ln} + \delta^{im}\delta^{jk}\delta^{ln} - \delta^{il}\delta^{jm}\delta^{kn} - \delta^{im}\delta^{jl}\delta^{kn})\\ &\quad\times(\theta^p\theta^p)^2\bar\Theta^{(mn)}; \qquad \text{(B20)}\end{aligned}$$

$$(\theta^i\sigma^{ab}\theta^j)(\theta^k\sigma^{cd}\theta^l) = -\tfrac{1}{16}(\eta^{ac}\eta^{bd} - \eta^{ad}\eta^{bc} - i\epsilon^{abcd}) \times(\delta^{ik}\delta^{jl} - \delta^{il}\delta^{jk})(\theta^m\theta^m)^2; \qquad \text{(B21)}$$

$$(\theta^i\sigma^{ab}\theta^j)(\bar\theta^k\bar\sigma^{cd}\bar\theta^l) = -\tfrac{1}{8}\,\mathrm{tr}(\sigma^{ab}\sigma^e\bar\sigma^{cd}\bar\sigma^f)(\delta^{ik}\delta^{jl} - \delta^{il}\delta^{jk}) \times(\theta^m\sigma_e\bar\theta^m)(\theta^n\sigma_f\bar\theta^n); \qquad \text{(B22)}$$

$$(\theta^i\sigma^{ab}\theta^j)(\theta^k\sigma^c\bar\theta^l) = \tfrac{1}{4}(\eta^{ad}\eta^{bc} - \eta^{ac}\eta^{bd} + i\epsilon^{abcd})(\delta^{ik}\delta^{jm} - \delta^{im}\delta^{jk})(\theta^n\theta^n)(\theta^m\sigma_d\bar\theta^l); \qquad \text{(B23)}$$

$$(\theta^i\sigma^{ab}\theta^j)\Theta^{\alpha(klm)} = \tfrac{1}{6}(\sigma^{ab})_\beta{}^\alpha[\delta^{ik}(\theta^n\theta^n)\Theta^{\beta(jlm)} + \mathrm{sym}(klm) + \mathrm{antisym}[ij]]; \qquad \text{(B24)}$$

$$\begin{aligned}(\theta^i\sigma^a\bar\theta^j)(\theta^k\sigma^b\bar\theta^l) &= \tfrac{1}{2}(\delta^{ij}\delta^{kl} - \delta^{il}\delta^{jk})(\theta^m\sigma^a\bar\theta^m)(\theta^n\sigma^b\bar\theta^n)\\ &\quad + \tfrac{1}{16}\eta^{ab}(3\delta^{il}\delta^{jk} - \delta^{ij}\delta^{kl} - \delta^{ik}\delta^{jl})\\ &\quad\times(\theta^m\sigma^c\bar\theta^m)(\theta^n\sigma_c\bar\theta^n)\\ &\quad + \tfrac{1}{2}\eta^{ab}\Theta^{(ijkl)} + \tfrac{1}{8}(\delta^{ij}\delta^{km}\delta^{ln}\\ &\quad + \delta^{il}\delta^{km}\delta^{jn} + \delta^{im}\delta^{jk}\delta^{ln}\\ &\quad + \delta^{im}\delta^{jn}\delta^{kl})\Theta^{[mn]} + \tfrac{1}{4}\eta^{ab}\\ &\quad\times[\delta^{ik}(\theta^m\theta^m)\bar\Theta^{(jl)} + \delta^{jl}(\bar\theta^m\bar\theta^m)\Theta^{(ik)}]\\ &\quad + (\delta^{ij}\delta^{km}\delta^{ln} - \delta^{il}\delta^{jm}\delta^{kn})\Theta^{(mp)}\\ &\quad\times(\bar\theta^p\bar\sigma^{ab}\bar\theta^n) + (\delta^{ij}\delta^{km} - \delta^{im}\delta^{jk})\\ &\quad\times\bar\Theta^{(mp)}(\theta^p\sigma^{ab}\theta^l) + \tfrac{1}{2}\delta^{ik}(\theta^m\theta^m)\\ &\quad\times(\bar\theta^j\bar\sigma^{ab}\bar\theta^l) + \tfrac{1}{2}\delta^{jl}(\bar\theta^m\bar\theta^m)(\theta^i\sigma^{ab}\theta^k)\\ &\quad + \tfrac{1}{16}\eta^{ab}(3\delta^{ik}\delta^{jl} - \delta^{ij}\delta^{kl} - \delta^{il}\delta^{jk})\\ &\quad\times(\theta^m\theta^m)(\bar\theta^n\bar\theta^n); \qquad \text{(B25)}\end{aligned}$$

$$\begin{aligned}&(\theta^i\sigma^a\bar\theta^j)\Theta^{\alpha(klm)}\\ &= -\tfrac{1}{24}\bar\sigma^{a\dot\alpha\alpha}(\bar\theta^n\bar\theta^n)\{3\delta_{ij}\bar\Theta_{\dot\alpha(\mathrm{klm})} + [3\delta_{jk}\bar\Theta_{\dot\alpha(ilm)}\\ &\quad - \delta_{ik}\bar\Theta_{\dot\alpha(jlm)} + \mathrm{sym}(klm) - \mathrm{tr}(klm)]\}\\ &\quad - \tfrac{1}{6}[\eta^{ab}\delta_\beta{}^\alpha - \tfrac{1}{4}(\sigma^b\bar\sigma^a)_\beta{}^\alpha](\bar\theta^n\bar\theta^n)(\theta^p\sigma_b\bar\theta^p)\\ &\quad\times[(\delta^{il}\delta^{jm} + \delta^{im}\delta^{jl})\theta^{\beta k} + \mathrm{sym}(klm) - \mathrm{tr}(klm)]\\ &\quad - \tfrac{1}{24}\bar\sigma^{a\dot\alpha\alpha}(\theta^n\theta^n)(\bar\theta^p\bar\theta^p)\\ &\quad\times[(\delta_{il}\delta_{jm} + \delta_{im}\delta_{jl})\bar\theta_{\dot\alpha k} + \mathrm{sym}(klm) - \mathrm{tr}(klm)]; \qquad \text{(B26)}\end{aligned}$$

$$\begin{aligned}&(\theta^i\sigma^a\bar\theta^j)\Theta^{(klmn)}\\ &= \tfrac{1}{36}(\theta^p\theta^p)(\bar\theta^q\bar\theta^q)\{(\delta^{kl}\delta^{mn} + \delta^{km}\delta^{ln} + \delta^{kn}\delta^{lm})\\ &\quad\times\theta^{(i}\sigma^a\bar\theta^{j)} - [(\delta^{in}\delta^{lm} + \delta^{im}\delta^{ln}\\ &\quad + \delta^{il}\delta^{mn})\theta^{(j}\sigma^a\bar\theta^{k)} + \mathrm{sym}(klmn)]\\ &\quad + [(3\delta^{im}\delta^{jn} + 3\delta^{in}\delta^{jm} - \delta^{ij}\delta^{mn})\\ &\quad\times\theta^{(k}\sigma^a\bar\theta^{l)} + \mathrm{sym}(klmn)]\}; \qquad \text{(B27)}\end{aligned}$$

$$\begin{aligned}&(\theta^i\sigma^a\bar\theta^j)\Theta^{[kl]}\\ &= \tfrac{1}{2}(\delta^{ik}\delta^{lm} - \delta^{il}\delta^{km})(\theta^n\theta^n)(\bar\theta^p\bar\theta^p)\theta^{[j}\sigma^a\theta^{m]}\\ &\quad + \tfrac{1}{4}(\delta^{ik}\delta^{jl} - \delta^{il}\delta^{jk})(\theta^m\theta^m)(\bar\theta^n\bar\theta^n)(\theta^p\sigma^a\bar\theta^p); \qquad \text{(B28)}\end{aligned}$$

$$(\theta^k\theta^k)(\theta^l\sigma^a\bar\theta^l)(\theta^m\sigma^b\bar\theta^m) = \tfrac{1}{4}\eta^{ab}(\theta^k\theta^k)^2(\bar\theta^l\bar\theta^l); \qquad \text{(B29)}$$

$$\begin{aligned}&(\theta^k\sigma^a\bar\theta^k)(\theta^l\sigma^b\bar\theta^l)(\theta^m\sigma^c\bar\theta^m)\\ &= \tfrac{1}{2}(\eta^{ab}\eta^{cd} + \eta^{ac}\eta^{bd} + \eta^{ad}\eta^{bc})\\ &\quad\times(\theta^k\theta^k)(\bar\theta^l\bar\theta^l)(\theta^m\sigma_d\bar\theta^m). \qquad \text{(B30)}\end{aligned}$$

[1] J. Wess and B. Zumino, Nucl. Phys. B **70**, 39 (1974).

[2] W. Siegel and M. Roček, Phys. Lett. B **105**, 275 (1981).

[3]Salam and J. Strathdee, Nucl. Phys. B **76**, 477 (1974).
[4]E. Sokachev, Nucl. Phys. B **99**, 96 (1975).
[5]V. Ogievetsky and E. Sokachev, J. Phys. A **10**, 2021 (1977).
[6]J. G. Taylor. Nucl. Phys. B **169**, 484 (1980); C. Pickup and J. G. Taylor, Nucl. Phys. B **188**, 577 (1981).
[7]B. de Wit and J. W. van Holten, Nucl. Phys. B **155**, 530 (1979).
[8]R. Haag, J. T. Łopuszański, and M. Sohnius, Nucl. Phys. B **88**, 257 (1975).
[9]A. Gal'perin, E. Ivanov, and V. Ogievetsky, Pis'ma Zh. Eksp. Teor. Fiz. **33**, 176 (1981) [JETP Lett. **33**, 168 (1981)].
[10]R. J. Firth and J. D. Jenkins, Nucl. Phys. B **85**, 525 (1975).
[11]P. Carruthers, *Spin and Isospin in Particle Physics* (Gordon and Breach, New York, 1971).
[12]W. Siegel, Nucl. Phys. B **142**, 301 (1978); W. Siegel and S. J. Gates, Jr., Nucl. Phys. B **147**, 77 (1979).
[13]V. O. Rivelles and J. G. Taylor, Phys. Lett. B **104**, 131 (1981); J. G. Taylor, J. Phys. A **15**, 867 (1982); V. O. Rivelles and J. G. Taylor, Phys. Lett. B **121**, 37 (1983).

On the construction of amplitudes with Mandelstam analyticity from observable quantities

I. Sabba Stefanescu
Institut für Theoretische Kernphysik der Universität Karlsruhe, 75 Karlsruhe 1, West Germany

(Received 23 September 1982; accepted for publication 14 October 1983)

It is shown that the problem of the construction of scattering amplitudes with Mandelstam analyticity from knowledge of their modulus in the three physical channels can be reduced, within a rather large class of functions, to the second Cousin problem of the theory of functions of two complex variables. As a consequence, it can be solved completely and explicitly. We derive conditions on the modulus function, under which at least one solution exists, as well as criteria for the correct resolution of the discrete ambiguity at fixed energy.

PACS numbers: 11.50.Nk, 11.80.Gw, 11.20.Fm, 03.80. + r

I. INTRODUCTION

The problem of the determination of the phase of the scattering amplitude from observable quantities (i.e., $d\sigma/d\Omega$ for scattering of spinless particles, $d\sigma/d\Omega$ and polarization for spin-0-spin-$\frac{1}{2}$ scattering, etc.) has an obvious physical interest and has led, in the course of time, to a set of very elegant studies in mathematical physics.[1–8] These studies (see Ref. 9 for a review) have succeeded in establishing with precision the extent of the ambiguity that is left in the phase if one takes into account, at a fixed energy, data over the whole angular region and uses the unitarity property of the amplitude.[1–6,8]

It is profitable to recall right now in more detail the problem of phase shift analysis at fixed energy, for the case of a reaction between spinless particles. The modulus (squared) of the amplitude $A(z=\cos\theta)$ (θ = c.m. scattering angle) is supposed to be known on the physical region $-1<\cos\theta<1$, from measurements of the differential cross section

$$d\sigma/d\Omega\,(z) = A(z)A^*(z), \quad -1<z<1, \tag{1.1}$$

and the question is to understand the way the phase of $A(z)$ is restricted by this information. The amplitude $A(z)$ is supposed to be holomorphic, e.g., in the z plane except for the half-lines ("cuts") $(-\infty, -z_-)$, (z_+, ∞), $z_-, z_+>0$. With this hypothesis on $A(z)$, one can divide the ambiguity of the phase in two parts: the discrete and the continuum ambiguity.

For their definition, one requires first an extension of Eq. (1.1) to the whole complex plane (minus the two half lines). This is done by simply writing the right-hand side of Eq. (1.1) as $A(z)A^*(z^*)$; this expression is manifestly holomorphic in z in the (cut) complex plane. We call $\mathscr{M}(z)$ the function obtained by this extension,

$$\mathscr{M}(z) \equiv A(z)A^*(z^*). \tag{1.2}$$

$\mathscr{M}(z)$ can, in principle, be found everywhere by analytic continuation from knowledge of $d\sigma/d\Omega$ on $-1<z<1$. It is real analytic and coincides with $|A(z)|^2$ for all z on $(-z_-, z_+)$ (but not elsewhere). We can thus assume $|A(z)|^2$ is known not only on $(-1, 1)$ but on $(-z_-, z_+)$.

In general, $\mathscr{M}(z)$ vanishes at several pairs of complex conjugate points (z_i, z_i^*) in the complex plane. By (1.2), the amplitude $A(z)$ will vanish at one of these points, but we cannot *a priori* decide at which. There exists thus a twofold ambiguity concerning the location of the zeros of $A(z)$, corresponding to each pair (z_i, z_i^*). It is easy to show that if N pairs of zeros are present, we can choose at will any one of the zeros in each pair and construct an amplitude with the correct modulus along (z_-, z_+), analytic in the cut z plane and vanishing precisely at those zeros. If we define

$$\mathscr{M}_1(z) = \frac{\mathscr{M}(z)}{\Pi_{i=1}^{N}(z-z_i)(z-z_i^*)}, \tag{1.3}$$

then a possible $A(z)$ is given by

$$A(z) = \prod_{j=1}^{N}(z-z_j)\sqrt{\mathscr{M}_1(z)}, \tag{1.4}$$

where the product extends over the given choice of N zeros. There exists thus at least a 2^N ambiguity in the reconstruction of the amplitude, for N distinct pairs of simple zeros. This is the discrete ambiguity "of the zeros."

If the amplitude were a polynomial of degree N, this would exhaust the ambiguity in the determination of the phase. However, the amplitude has cuts along $(-\infty, -z_-)$, (z_+, ∞) and it is easy to construct functions $\Omega(z)$, holomorphic in the z plane minus the cuts and having unit modulus on $(-z_-, z_+)$. For instance,

$$\Omega(z) = \exp\left[i\left(\frac{1}{\pi}\int_{-\infty}^{-z_-}\frac{f(z')}{z'-z}dz' + \frac{1}{\pi}\int_{z_+}^{\infty}\frac{f(z')}{z'-z}dz'\right)\right] \tag{1.5}$$

with $f(z)$ any real function giving meaning to the integrals. Clearly, if $A(z)$ is a possible amplitude, $A_1(z) = A(z)\Omega(z)$ is also an amplitude. This gives the continuum ambiguity of the phase.

If the unitarity property of $A(z)$ is taken into account, a drastic reduction of these ambiguities occurs. In the elastic region, at most a discrete type of ambiguity (Crichton[10]) is allowed (see Refs. 1, 5, and 8), whereas, above the inelastic threshold, apart from possible discrete ambiguities of zeros, the continuum ambiguity gradually develops with increasing energy.[2–4,6]

The question which naturally arises is then to describe the indeterminacy of the phase allowed by knowledge of the observables on larger intervals of energy and by the analytic structure of the amplitude. For πN scattering, use of a number of specific experimental facts and of the linear relation between the amplitudes of measurable processes can even restore uniqueness within a certain class of amplitudes, in some interval of t values around $t=0$, if analyticity in the energy is taken into account, together with data at all energies.[11]

An "extreme" and interesting question was raised in Ref. 7: to determine the extent to which analyticity of the amplitude in two variables can fix its phase, e.g., in $\pi\pi$ scattering, if one assumes the differential cross section is measured at all energies and angles in the three channels. In this connection, it is shown in Ref. 7 that if two amplitudes $A_{1,2}(s,t,u)$ have the same modulus in the three physical channels and have the same (axiomatic) analyticity domain, then they differ by a factor

$$R(s,t,u) = \frac{A_1(s,t,u)}{A_2(s,t,u)} = \frac{[(4m^2-s)(4m^2-t)(4m^2-u)]^{1/2} - f(s,t,u)}{[(4m^2-s)(4m^2-t)(4m^2-u)]^{1/2} + f(s,t,u)}, \tag{1.6}$$

with $f(s,t,u)$ the ratio of two entire functions, $h(s,t,u)$ and $g(s,t,u)$ of s,t,u (s,t,u are the Mandelstam variables, $s+t+u=4m^2$, $m=$ mass of the pion). In Ref. 12, it was shown that, in a special class of amplitudes, holomorphic in the Mandelstam domain, continuous on the cuts and such that $h(s,t,u)$ and $g(s,t,u)$ are polynomials of s,t,u, $f(s,t,u)$ cannot in fact be freely chosen; rather, there exists only a discrete[13] set of such functions: if $f(s,t,u)$ does not belong to this set, $A_2(s,t,u)R(s,t,u)$ is not analytic in the Mandelstam domain. Multiplying the amplitude $A_2(s,t,u)$ by an admissible factor $R(s,t,u)$ amounts to the global reflection across the physical region of a certain complex of zero trajectories (see Sec. II) of $A_2(s,t,u)$. The restriction to polynomial $h(s,t,u)$, $g(s,t,u)$ in (1.6) is not as drastic as it might seem: it is achieved in a class of amplitudes $A(s,t,u)$ for which there exists an interval of t values, $-4m^2<t<4m^2$, so that $A(s,t,u)$ has a finite number of zeros in the complex s plane and an interval of real s (or u) values, $-4m^2<s<4m^2$, where $A(s,t,u)$ has a finite number of zeros in the corresponding t plane.

The result of Ref. 12 is qualitative: If all the trajectory functions are known at all energies in all three physical channels, the amplitude is uniquely determined, from analyticity requirements, up to a trivial overall sign. A first open problem is thus the actual construction of the amplitude from the available modulus. A second question is, given a certain resolution of the discrete ambiguity at each energy in the three channels, i.e., a certain choice for the positions of the zero trajectories, to find a computational means to decide whether it is one of the "correct" possibilities, consistent with analyticity in two variables. Indeed, the reasoning at fixed energy made above would allow us in principle to reverse the sign of the imaginary part of one whole trajectory function, without affecting the modulus of the amplitude. It is not clear, however, that such a reflection would not destroy the property of analyticity in two variables of the amplitude. In fact, it turns out it does so in most cases, and we call "correct" resolutions of the discrete ambiguity those for which the resulting amplitude function does not violate this requirement. In Refs. 14 and 15 a method has been proposed to pick out "incorrect" possibilities, on the basis of local two variable analyticity, under the assumption that the amplitude is also holomorphic in some domain of a second sheet.

The present paper wishes to show that if the amplitude $A(s,t,u)$ for the scattering of scalar particles belongs to a certain class of functions, even more restricted than the one of Ref. 12, but still rather large and physically acceptable, and if its modulus *and* the position of the zero trajectories are known at all energies in (the physical region of) all channels, then $A(s,t,u)$ can indeed be constructed and, in the opinion of the author, in a surprisingly simple manner. In fact, we show that the construction of $A(s,t,u)$ can be reduced to the solution of the second Cousin problem of the theory of functions of several (in this case, two) complex variables; the latter problem is completely and explicitly solvable for the domains we are considering. The reduction is possible only if the zero trajectories are appropriately situated; this yields certain conditions (sum rules) that have to be fulfilled, if the discrete ambiguity of the zeros has been "correctly" solved. Thus, we can give an answer to the second problem above.

It is important to realize from the outset that the functions that describe the modulus in the three channels cannot be prescribed independently. Indeed, consider the special situation of an amplitude which has no zeros in the whole Mandelstam domain. Then, knowledge of its modulus in two channels only (e.g., s and u) is sufficient for the determination of the amplitude. This is done by writing at fixed t, $|t|<4m^2$, a dispersion relation for the function $\ln A(s,t,u)/[(4m^2-s)(4m^2-u)]^{1/2}$, whose imaginary part is known and whose real part is related to the phase of the amplitude. A similar dispersion relation can certainly be written at fixed s, $|s|<4m^2$, using the modulus in the t and u channels, and it must lead to the same amplitude (in particular in the domain $|s|$, $|t|$, $|u|<4m^2$). This is impossible unless the modulus in the t channel is appropriately constrained. The presence of zeros in the analyticity domain will not modify this situation.

In this paper, we shall assume that we are given in the physical region of the three channels the correct modulus function of an amplitude. The constraints that this function must obey will appear in the end as the condition that the amplitude constructed by two different routes is the same.

In the following, we discuss an example in which the construction can proceed in a simple manner; this will clarify the difficulties arising in the general case. Consider the artificial case of an amplitude $f(\sigma,\tau)$, holomorphic in σ and τ in a neighborhood of the direct product $\overline{\mathscr{D}}$ of the two unit disks $|\sigma|\leqslant 1$ and $|\tau|\leqslant 1$. We assume the modulus is available in the set $\{(\operatorname{Im}\sigma=0, |\sigma|<1)\otimes|\tau|=1\}\cup\{|\sigma|=1\otimes(\operatorname{Im}\tau=0, |\tau|<1)\}$ (this is the natural analog of the "physical" region if the amplitude has only two cuts, see Sec. II).

We place ourselves in the following very particular situation: for each σ on $|\sigma|=1$, $f(\sigma,\tau)$ has precisely N zeros in $|\tau|<1$, is nonvanishing for $|\tau|=1$, and for each τ on $|\tau|=1$,

$f(\sigma,\tau)$ has no zeros in $|\sigma|<1$. For this special case, we can easily determine a Cousin function $C_0(\sigma,\tau)$, holomorphic in $\mathscr{B}\equiv(|\sigma|<1)\otimes(|\tau|<1)$ and vanishing precisely where $f(\sigma,\tau)$ vanishes and nowhere else (in $\mathscr{B}$).

To this end, we first notice that the number of zeros in $|\tau|<1$ of $f(\sigma,\tau)$ for any σ inside $|\sigma|<1$ is always equal to N. This follows from the relation

$$N=\frac{1}{2\pi i}\oint_{|\tau|=1}\frac{\partial f/\partial\tau(\sigma,\tau)}{f(\sigma,\tau)}d\tau \tag{1.7}$$

evaluated for σ on $|\sigma|=1$. Since for $|\tau|=1$, $f(\sigma,\tau)$ is nonvanishing in $|\sigma|<1$ and holomorphic in a neighborhood of $|\sigma|\leqslant 1$, we can extend the right-hand side to $|\sigma|<1$; it represents an integer-valued, holomorphic function of σ in $|\sigma|<1$ and must therefore be the constant N itself. Let the N zeros, at a given σ, $|\sigma|<1$, be $\tau_1(\sigma)$, $\tau_2(\sigma)$, ... , $\tau_N(\sigma)$. They are locally holomorphic functions of σ, with at most branch points of finite order, by Weierstrass' preparation theorem (see Appendix A). At a branch point σ_0, a subset of the $\tau_i(\sigma)$, $i=1,\dots,N$, are connected to each other by analytic continuation around σ_0. As a consequence (see also Appendix A), the symmetric combinations

$$\begin{aligned}\beta_1(\sigma)&=-\sum_{i=1}^{N}\tau_i(\sigma),\\ \beta_2(\sigma)&=\sum_{i<j}\tau_i(\sigma)\tau_j(\sigma),\\ &\ \vdots\\ \beta_N(\sigma)&=\prod_{i=1}^{N}\tau_i(\sigma)*(-1)^N\end{aligned} \tag{1.8}$$

are holomorphic functions of σ in $|\sigma|<1$. Their values are known on $|\sigma|=1$, by assumption. We can determine them inside $|\sigma|<1$, by Cauchy's theorem:

$$\beta_k(\sigma)=\frac{1}{2\pi i}\oint\frac{\beta_k(\sigma')}{\sigma'-\sigma}d\sigma'\quad(k=1,\dots,N). \tag{1.9}$$

We conclude that the function $[\beta_0(\sigma)\equiv 1]$

$$\begin{aligned}C_0(\sigma,\tau)&=\sum_{k=0}^{N}\tau^k\beta_{N-k}(\sigma)\\ &=\prod_{i=1}^{N}(\tau-\tau_i(\sigma))\end{aligned} \tag{1.10}$$

has the required properties: it is holomorphic in $\mathscr{B}$ and vanishes only there where $f(\sigma,\tau)$ vanishes. The function

$$E(\sigma,\tau)=f(\sigma,\tau)/C_0(\sigma,\tau) \tag{1.11}$$

is then free of zeros and has a known modulus in the "physical region." Its construction is straightforward. It can be done by writing Poisson integrals either at fixed τ on $\operatorname{Im}\tau=0$ or at fixed σ on $\operatorname{Im}\sigma=0$. This freedom shows that the values of the modulus in the "two channels" cannot be prescribed independently, as remarked before. Further, a condition for the discrete ambiguity to be solved correctly (in a manner compatible with two variable analyticity), at each σ on $|\sigma|=1$ is that the functions $\beta_i(\sigma)$, known on $|\sigma|=1$, are limiting values of functions holomorphic in $|\sigma|<1$; this is a nontrivial restriction (cf. Refs. 14 and 15).

This example is very special; in general, zeros of $f(\sigma,\tau)$ "enter" or "get out" of the unit disk $|\tau|<1$, as we move along $|\sigma|=1$, and the same happens for $|\tau|=1$. At first sight, a complete solution to the problem of constructing the Cousin function might seem out of reach. Nevertheless, it turns out it is possible to perform the construction completely, for functions holomorphic in a neighborhood of $\overline{\mathscr{B}}$ and arbitrary distributions of zeros. Further, as announced, the treatment can be generalized to functions holomorphic in the Mandelstam domain, provided they are suitably restricted, but still sufficiently free to be of physical interest.

The paper is organized as follows.

Sections II and III deal with the construction of $f(\sigma,\tau)$, holomorphic in a neighborhood of $\overline{\mathscr{B}}$, from the values of its modulus in the "physical" region. In Sec. II we show that knowledge of the zeros of $f(\sigma,\tau)$ on the boundary $\partial\mathscr{B}$ of $\overline{\mathscr{B}}$: $\partial\mathscr{B}=\{|\sigma|=1\otimes|\tau|\leqslant 1\}\cup\{|\sigma|\leqslant 1\otimes|\tau|=1\}$ determines the whole manifold $\mathscr{Z}_f$ of the zeros of $f(\sigma,\tau)$ in all of $\mathscr{B}$. This is, in fact, the essential progress claimed by this paper. If the manifold is known (the "Cousin data"), we can apply Cousin's method[16,17] to write down a function $C(\sigma,\tau)$, holomorphic in $\mathscr{B}$ and vanishing precisely on $\mathscr{Z}_f$. We do this and then construct $f(\sigma,\tau)$ explicitly in Sec. III. In Sec. IV, we consider amplitudes with Mandelstam analyticity and introduce and discuss some assumptions needed to imitate the treatment of Secs. II and III. We then construct, from the "zero trajectories" known in the physical region, the Cousin data in the Mandelstam domain. In Sec. V, we show that the construction of the Cousin function of Ref. 16 can be generalized to the Mandelstam domain and thus obtain $A(s,t,u)$ explicitly. Conclusions are presented in Sec. VI, together with a review of the assumptions delimiting the class of amplitudes used in the paper.

II. THE DIRECT PRODUCT OF TWO UNIT DISKS. CONSTRUCTION OF THE COUSIN DATA

We consider a function $f(\sigma,\tau)$, holomorphic in σ and τ in $\mathscr{B}_\epsilon\equiv\{|\sigma|<1+\epsilon\otimes|\tau|<1+\epsilon\}$, for some $\epsilon>0$ and real analytic:

$$f^*(\sigma^*,\tau^*)=f(\sigma,\tau). \tag{2.1}$$

We assume its modulus squared, $M(\sigma,\tau)=|f(\sigma,\tau)|^2$, is known in the set $\mathscr{P}\equiv\mathscr{P}_\sigma\cup\mathscr{P}_\tau\equiv\{|\sigma|=1;$ $\operatorname{Im}\tau=0,|\tau|<1\}\cup\{\operatorname{Im}\sigma=0,|\sigma|<1;|\tau|=1\}$. The values of $M(\sigma,\tau)$ for $|\tau|=1$, σ real, can be extended analytically to the interior of the disk $|\sigma|<1+\epsilon$, by means of

$$M(\sigma,\tau)=f(\sigma,\tau)f^*(\sigma^*,\tau)\equiv M_\tau(\sigma). \tag{2.2}$$

One obtains similarly $M_\sigma(\tau)$ by extending the values of $M(\sigma,\tau)$ for $|\sigma|=1$, τ real. [Clearly, if we perform the mapping $\sigma\to s$, $\tau\to t$, $s=4\sigma/(1+\sigma)^2$, and the same for t, $f(\sigma(s),\tau(t))$ is holomorphic in a Mandelstam domain with two cuts and admits of extensions to the second sheet through the cut; the images of $\mathscr{P}_\sigma$ and $\mathscr{P}_\tau$ contain the physical regions of the s and t channels; the modulus of f is known on them, if it is known in the physical region, by analytic continuation; cf. Introduction.]

For each σ on $|\sigma|=1$, we can solve the equation

$$f(\sigma,\tau)=0 \tag{2.3}$$

and obtain a certain number of roots $\tau_i(\sigma)$, $|\tau_i|<1+\epsilon$. Assuming that, for no σ, $|\sigma|=1$, $f(\sigma,\tau)\equiv 0$, there can be only a finite number of such roots in $|\tau|\leqslant 1$, at each σ, $|\sigma|=1$. The functions $\tau_i(\sigma)$ are holomorphic in the neighborhood of every point σ, $|\sigma| = 1$, where all roots of $f(\sigma,\tau) = 0$ are distinct. As σ moves along $|\sigma| = 1$, they can acquire at most branch points of finite order (see Appendix A), if they stay inside $|\tau| < 1 + \epsilon$.

At a branch point σ_0, $|\sigma_0| = 1$, several functions $\tau_{i1}(\sigma)$, $\tau_{i2}(\sigma)$, ... ,$\tau_{in}(\sigma)$ acquire the same value. We define the *continuation* of one function $\tau_{i1}(\sigma)$ past σ_0 by analytic continuation along a small semicircle $|\sigma - \sigma_0| = \epsilon$, contained in $|\sigma| < 1$. We call "zero trajectory" $\tau_i(\sigma)$ the whole function obtained along $|\sigma| = 1$ by analytic continuation from a function element $\tau_i^0(\sigma)$ with the convention above. It can obviously be many-valued on $|\sigma| = 1$.

The trajectory function $\tau_i(\sigma)$ gives rise to a curve γ_i: $\tau = \tau_i(\sigma)$ in the plane, as σ moves along $|\sigma| = 1$ and τ stays inside $|\tau| < 1$. Such a curve is closed if, by analytic continuation along $|\sigma| = 1$ (with the convention above at branch points), we return to the same function element with which we started (after a number of turns). The curve γ_i should not be confused with the image set of $\tau_i(\sigma)$ in $|\tau| < 1$; it is possible that γ_i consists of the same set covered several times as is the case, e.g., with the zeros of $f(\sigma,\tau) = \tau - \frac{1}{2}\sigma^4$. The curve γ_i is specified only if its parametrization $\tau_i(\sigma)$, $|\sigma| = 1$, is given. We assign an orientation to the γ_i's by letting σ move counterclockwise along $|\sigma| = 1$.

It is of interest to consider also the set of points in $|\tau| < 1$ which are the images of $|\sigma| = 1$ under $\tau = \tau_i(\sigma)$, without reference to the trajectory function. It is possible that several trajectory functions have the same image set [e.g., the trajectory functions of $f(\sigma,\tau) = (2\sigma\tau - 1)(\sigma - 2\tau)$ have the same image set: the circle $|\tau| = \frac{1}{2}$]. It is thus reasonable to denote the distinct image sets by $\bar{\gamma}_l$, $l = 1,2, \ldots$, and, if necessary, the curves parametrized by $\tau_i(\sigma)$ and giving rise to $\bar{\gamma}_l$, by $\gamma_{l1}, \gamma_{l2}, \ldots \gamma_{ln}$. We say then that the set $\bar{\gamma}_l$ has multiplicity n, and to allow for the possible different orientations of the γ_{li}, we distinguish between two "oriented" multiplicities, $m_1, m_2, m_1 + m_2 = n$.

In Appendix B we justify in detail the following intuitively obvious picture of the curves γ_i and of the image sets $\bar{\gamma}_l$: (a) every γ_i has a continuous tangent, except possibly for a finite number of points where it might have cusps; (b) the γ_i's have finite length; (c) each γ_i has either a "beginning" at some τ_a and an end at τ_e, $|\tau_a| = 1$, $|\tau_e| = 1$, or it is closed (and lies entirely in $|\tau| \leqslant 1$); (d) there are a finite number of γ_i's; (e) two γ_i's can either intersect at a finite number of points or coincide; (f) let $\mathscr{L}_\tau = \cup \bar{\gamma}_i$; the complement of $\mathscr{L}_\tau$ in $|\tau| < 1$ consists of a finite number of domains $\mathscr{D}_i$.

We may state that "two points of a curve γ_{i1} are distinct" if they correspond to different values of the parameter σ, $|\sigma| = 1$. We can thus define along the image set $\bar{\gamma}_i$ containing γ_{i1} an "inverse" function $\sigma_{i1}(\tau)$ as being equal to the value of the parameter σ at the point τ; even if $\bar{\gamma}_i$ contains just one curve γ_{i1}, $\sigma_{i1}(\tau)$ may be many-valued [as in the example of $f(\sigma,\tau) = \tau - \frac{1}{2}\sigma^4$]. However, expressions like $\int_{\gamma_{i1}} \sigma_{i1}(\tau)\, d\tau$ have a well-defined meaning: they are $\int_{|\sigma| = 1} \sigma (d\tau_{i1}/d\sigma)\, d\sigma$ and the integration may cover $|\sigma| = 1$ several times.

Clearly, if $\bar{\gamma}_i$ is multiple, several (possibly multivalued) inverse functions $\sigma_{ik}(\tau)$ may be defined. The functions $\sigma_{ik}(\tau)$ are locally holomorphic on $\bar{\gamma}_i$, except for isolated branch points, where they are Hölder continuous.

Starting from one function element $\sigma_{i1}^0(\tau)$, defined in the neighborhood of τ_0 on $\bar{\gamma}_i$, we can construct by analytic continuation along γ_{i1} the whole $\sigma_{i1}(\tau)$. The only difficulty is at branch points of $\sigma_{i1}(\tau)$. We understand then that the continuation is performed along the image c' through the function $\tau_{i1}(\sigma)$, defining γ_{i1}, of a small semicircle $|\sigma - \sigma_0| = \epsilon$, $|\sigma| < 1$, $|\sigma_0| = 1$ in the sense induced by the clockwise sense in the σ plane.

The phrase "analytic continuation" of $\tau_i^0(\sigma)$ along $|\sigma| = 1$ or of $\sigma_i^0(\tau)$ along γ_i will be used in the text with the convention above at branch points. For clarity, two examples are now presented and sketched in Fig. 1. Consider $f(\sigma,\tau) = \tau^2 - (\sigma - 1)$ first. For σ near 1 on $|\sigma| = 1$, there are two zeros in $|\tau| < 1$: $\tau_{1,2} = \pm(\sigma - 1)^{1/2}$, which approach each other as $\sigma \to 1$. The image of the small semicircle c (Fig. 1) consists of two arcs c' and c'' in the τ plane. The curves γ_i, $i = 1,2$, are uniquely defined. The inverse functions $\sigma_{1,2}(\tau) = \tau^2 + 1$ are regular at $\tau = 0$. Consider next the zeros of $f(\sigma,\tau) = \tau - (\sigma - 1)^3$. We obtain for σ near 1 a smooth curve $\tau_1(\sigma)$ as image of $|\sigma| = 1$ in the τ plane. However, there is a branch point of the inverse function at $\tau = 0$. The image of the semicircle c consists of the (slightly deformed) contour c' in the τ plane, and we understand that $\sigma_1(\tau)$ "beyond" $\tau = 0$ is the continuation of $\sigma_1(\tau)$ "before" $\tau = 0$ along the path c' (Fig. 1).

As τ moves around $|\tau| = 1$, we denote also by $\sigma_i(\tau)$ the various roots of $f(\sigma,\tau) = 0$ contained in $|\sigma| \leqslant 1$. We can divide the unit circle $|\tau| = 1$ into a finite number of intervals (τ_j, τ_{j+1}), $j = 1,2, \ldots ,N$, $\tau_N = \tau_1$, so that the number n_j of roots of $f(\sigma,\tau) = 0$ in $|\sigma| \leqslant 1$ is constant for fixed τ in (τ_j, τ_{j+1}). We disregard zeros of $f(\sigma,\tau)$ on $|\sigma| = 1 \otimes |\tau| = 1$ which are isolated in $\partial\mathscr{B}$ (see Appendix C).

We can now part the various functions $\sigma_i(\tau)$ associated with γ_i and with intervals (τ_i, τ_{i+1}) into classes. Two functions $\sigma_k(\tau)$, $\sigma_l(\tau)$ belong to the same class if they can be obtained from each other by analytic continuation along the curves $\{\gamma_j\}$ and the circle $|\tau| = 1$. This partition clearly in-

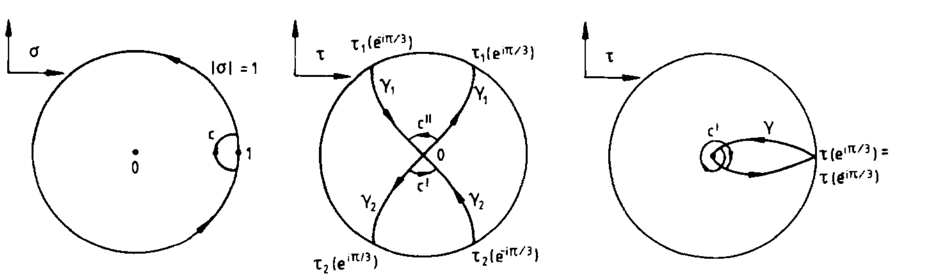


FIG. 1. Paths of analytic continuation of $\tau_i(\sigma)$ and $\sigma_i(\tau)$ at irregular points of γ_i (Sec. II).

duces also a separation of the curves γ_i into classes.

Explicitly, assume we start with an element $\sigma_1^0(\tau)$ of the function $\sigma_1(\tau)$ at τ_1, $|\tau_1| = 1$, and continue it counterclockwise along $|\tau| = 1$. We might reach a point τ_0, $\sigma_1(\tau_0) = \sigma_0$, so that $|\sigma_1(\tau_0)| = 1$. If τ_0 is a regular point of $\sigma_1(\tau)$, $d\sigma_1/d\tau \neq 0$ and finite, then $\sigma_1(\tau)$ realizes a conformal mapping of a sufficiently small convex neighborhood V_0 of τ_0 onto a neighborhood U_0 of σ_0. The image of the piece of circle: $\mathscr{C} = V_0 \cap (|\tau| = 1)$ under $\sigma_1(\tau)$ is a curve $\mathscr{C}'$ in the σ plane. Along $\mathscr{C}'$, $|\sigma_1(\tau)| < 1$ for τ before τ_0 and either $|\sigma_1(\tau)| < 1$ or $|\sigma_1(\tau)| > 1$ for τ past τ_0, and sufficiently close to it. In the former case, we continue further $\sigma_1(\tau)$ on $|\tau| = 1$ [as in the case of $f(\sigma,\tau) = (\sigma - 1/2) - \tau/2$]. In the second case, consider the inverse mapping $\tau = \tau_1(\sigma)$ and the image under τ_1 of a piece of circle $U_0 \cap (|\sigma| = 1)$. Part of this image must coincide with the image set of a curve γ_2: $\tau = \tau_2(\sigma)$, ending at τ_0 and having $\sigma_1(\tau)$ as inverse function, for τ near τ_0. Thus, $\sigma_1(\tau)$ is identical with $\sigma_2(\tau)$, the inverse function associated with γ_2. We continue now $\sigma_1(\tau)$ along γ_2 and can verify that, as we move away from $|\tau| = 1$, the value $\sigma_1(\tau)$ moves counterclockwise on $|\sigma| = 1$ [see Fig. 2(a) for the example $f(\sigma,\tau) = \sigma\tau + \sigma + \tau - 1$].

In the general case (without restrictions on $d\sigma_1/d\tau$), we continue $\sigma_1(\tau)$ analytically clockwise on a small circle around τ_0, starting at τ "before" τ_0, where $|\sigma_1(\tau)| < 1$. Either we reach in this process a first curve γ_2 on which $|\sigma_1(\tau)| \equiv 1$ — then $\sigma_1(\tau) \equiv \sigma_2(\tau)$, the inverse function associated to γ_2 [see Fig. 2(b) for the example $f(\sigma,\tau) = \tau - 1 - (\sigma - 1)^2$] or we do not meet any such curve [as for $f = \tau - 1 + (\sigma - 1)^2$]. In the former case, we continue $\sigma_1(\tau)$ along γ_2, in the latter along $|\tau| = 1$. If we continue along γ_2, we must reach again $|\tau| = 1$, since γ_2 has an "end" (see Appendix B). We continue again counterclockwise on $|\tau| = 1$ and repeat the procedure until we recover the function we started with. This must be possible, since there are only a finite number of roots of $f(\sigma,\tau) = 0$ at fixed $\tau, |\tau| = 1$.

The closed path of several curves γ_i and pieces of the circle $|\tau| = 1$, needed to return $\sigma_1(\tau)$ to its original value is called Γ_1. With this procedure, each curve γ_i falls into one class Γ_k and only in one. A closed γ_i makes up a class by itself.

We make next a simplifying assumption, which we discuss and remove in Appendix C:

(A) There exist only a finite number of points τ, $|\tau| = 1$, for which one of the roots $\sigma(\tau)$ of $f(\sigma,\tau) = 0$ has unit modulus $(|\sigma(\tau)| = 1)$.

Clearly, the "end points" of γ_i are among these points. In fact, the other ones are isolated in $\partial\mathscr{B}$, i.e., they are not limit points of other zeros of $f(\sigma,\tau)$ on $\partial\mathscr{B}$. As indicated in Appendix C, we can ignore them in the following (since they are isolated in $\mathscr{B}$).

We are now in a position to prove the following.

Theorem 2.1. Define

$$\mathscr{N}(\tau) = \sum_j \frac{1}{2\pi i}\int_{\gamma_j} \frac{d\tau'}{\tau' - \tau} + \sum_j \frac{n_j}{2\pi i}\int_{\substack{\tau_j \\ |\tau'|=1}}^{\tau_{j+1}} \frac{d\tau'}{\tau' - \tau} \tag{2.4}$$

with the sense of integration counterclockwise on $|\tau'| = 1$ and the sense of γ_j induced by the counterclockwise motion on $|\sigma| = 1$; n_j is the number of roots of $f(\sigma,\tau) = 0$ in $|\sigma| \leqslant 1$ for $|\tau| = 1$, $\tau \in (\tau_j, \tau_{j+1})$. Let further $n(\tau)$ be the number of roots of $f(\sigma,\tau) = 0$ in $|\sigma| < 1$ for fixed $\tau \in (\mathrm{C}\mathscr{Z}_\tau) \cap (|\tau| < 1)$ $(\equiv \cup_i \mathscr{D}_i)$. Then:

(a) $\mathscr{N}(\tau) \equiv 0$, for $|\tau| > 1$;
(b) $\mathscr{N}(\tau) = N_i$, a constant integer, for $\tau \in \mathscr{D}_i$,
(c) $n(\tau) = \overline{N}_i$, a constant integer, for $\tau \in \mathscr{D}_i$;
(d) $N_i = \overline{N}_i$.

Before proceeding to the proof, we point out that, according to this theorem, we find unambiguously the number of roots of $f(\sigma,\tau) = 0$ in $|\sigma| < 1$ for τ fixed, $\tau \in \mathscr{D}_i$, only from information concerning the zero trajectories, available on the boundary $\partial\mathscr{B}$: $\{|\sigma| \leqslant 1 \otimes |\tau| = 1\} \cup \{|\sigma| = 1 \otimes |\tau| \leqslant 1\}$. This is the first step in the construction of the Cousin data.

Proof: We can rewrite (2.4) as

$$\mathscr{N}(\tau) = \sum_k \frac{1}{2\pi i}\int_{\Gamma_k} \frac{d\tau'}{\tau' - \tau} \tag{2.5}$$

with Γ_k the closed loops introduced before, needed to return $\sigma_i(\tau)$ to its original value. Since all Γ_k are contained in $|\tau| < 1$, (a) and (b) follow by application of the residue theorem. (Notice that closed loops contained in $|\tau| < 1$ may contribute negative values.) The fact that the γ_i have finite length gives meaning to (2.4). To prove (c), we write

$$n(\tau) = \frac{1}{2\pi i}\oint_{|\sigma|=1} \frac{\partial f(\sigma,\tau)/\partial\sigma}{f(\sigma,\tau)}\, d\sigma. \tag{2.6}$$

For τ in $\mathscr{D}_i$, $f(\sigma,\tau) \neq 0$, if $|\sigma| = 1$, since $\mathscr{D}_i \subset \mathrm{C}\mathscr{Z}_\tau$. Since the right-hand side of (2.6) is continuous in τ and $n(\tau)$ is an integer, it must be a constant, as long as τ does not reach a curve γ_i (or $|\tau| = 1$). Thus $n(\tau) = \overline{N}_i$ if $\tau \in \mathscr{D}_i$.

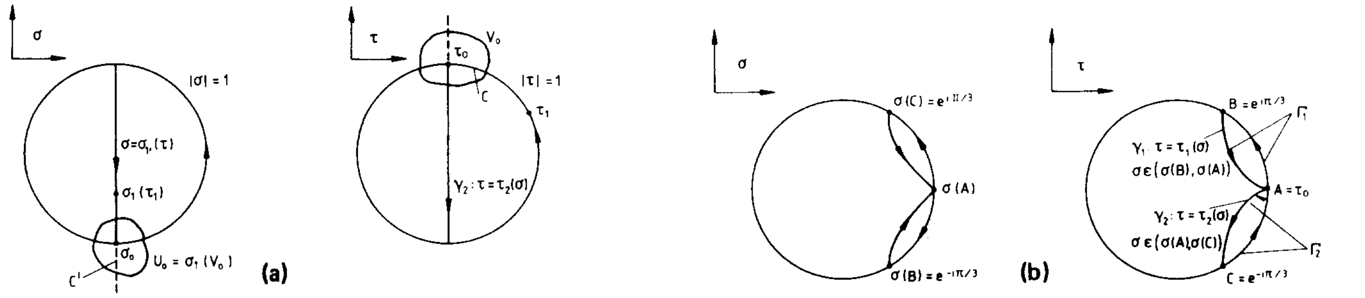


FIG. 2. The continuation of $\sigma_i(\tau)$ at points, where $|\sigma_i(\tau)| = 1$ and the path described by $\sigma_i(\tau)$ in the σ plane (Sec. II); (a) $f(\sigma,\tau) = \sigma\tau + \sigma + \tau - 1$; (b) $f(\sigma,\tau) = \tau - 1 - (\sigma - 1)^2$.

To prove (d), we first consider a domain $\mathscr{D}_1$ adjacent to the boundary $|\tau| = 1$ along (τ_1, τ_2) and show that (d) is true for this domain. Indeed, since $\mathscr{N}(\tau)\equiv 0$ for $|\tau| > 1$, it follows from the Plemelj discontinuity formulas (Ref. 18, p. 43) that, for τ in the domain $\mathscr{D}_1$, $\mathscr{N}(\tau) = n_1$. By assumption (A), we can choose τ_0, $\tau_0 \in (\tau_1, \tau_2)$, so that all $n(\tau_0) = n_1$ roots of $f(\sigma,\tau_0) = 0$ lie in the interior of $|\sigma| < 1$. But, then, $n(\tau)$ defined by (2.6) for τ in a neighborhood U_{τ_0} of τ_0 stays constant in U_{τ_0}, as shown above (if U_{τ_0} is sufficiently small). Then, however, $n(\tau)$ is constant in all of $\mathscr{D}_1$, and $n(\tau) = n_1$. Therefore, (d) is true for τ in $\mathscr{D}_1$.

To prove (d) for a domain $\mathscr{D}_2$ adjacent to $\mathscr{D}_1$, we compute first $\mathscr{N}(\tau)$ for τ in $\mathscr{D}_2$ by means of the Plemelj formulas from (2.4) and then show that, by crossing the boundary between $\mathscr{D}_1$ and $\mathscr{D}_2$, $n(\tau)$ varies by the same amount at $\mathscr{N}(\tau)$. Assume that the boundary $\bar{\gamma}_1$ between $\mathscr{D}_1$ and $\mathscr{D}_2$ is simple, given by a curve γ_1 with an inverse function $\sigma_1(\tau)$ single-valued on $\bar{\gamma}_1$[19]; assume also that γ_1 is such that $\mathscr{D}_2$ lies on its left. Then the discontinuity of $\mathscr{N}(\tau)$ is $N_2 - N_1 = 1$. If $\sigma(\tau)$ is k-valued rather than one-valued, then $N_2 - N_1 = k$ and if $\bar{\gamma}_1$ is multiple and has multiplicities $m_1, \bar{m}_1$(m_1 when it is oriented so that $\mathscr{D}_2$ lies on its left), $N_2 - N_1 = k_1 + k_2 + \cdots k_{m_1} - k'_1 - k'_2 - \cdots - k'_{\bar{m}1}$, where k_i, k'_j indicate the many valuedness of the inverse functions $\sigma_{ik}(\tau)$.

For the computation of $n(\tau)$ in $\mathscr{D}_2$, assume again first that $\bar{\gamma}_1$ is simple and $\sigma_1(\tau)$ is single-valued. Consider then a point $\tau_0 \in \bar{\gamma}_1$ and a neighborhood U_{τ_0} of it, where the inverse function $\sigma_1(\tau)$ is holomorphic and $d\sigma_1/d\tau(\tau_0)\neq 0$. As τ approaches $\bar{\gamma}_1$, $|\sigma_1(\tau)| \to 1$ and, if $\mathscr{D}_2$ lies on the left of γ_1, $|\sigma_1(\tau)| < 1$ in $\mathscr{D}_2 \cap U_{\tau_0}$, $|\sigma_1(\tau)| > 1$ in $\mathscr{D}_1 \cap U_{\tau_0}$ (conformal mappings preserve orientations of angles). Thus, as we cross $\bar{\gamma}_1$ from $\mathscr{D}_1$, a new zero $\sigma_1(\tau)$ "enters" the unit disk $|\sigma| < 1$. On the other hand, consider the other $\tilde{n}(\tau_0)$ zeros of $f(\sigma,\tau_0)$ lying in $|\sigma| < 1$. By continuity, they stay inside $|\sigma| < 1$ for τ in a sufficiently small neighborhood of τ_0. Alternatively,[20] one can verify that

$$\tilde{n}(\tau) = \frac{1}{2\pi i}\oint_{\mathscr{C}} \frac{\partial f(\sigma,\tau)/\partial\sigma}{f(\sigma,\tau)}\, d\sigma \tag{2.7}$$

is continuous across $\bar{\gamma}_1$ in a neighborhood (small enough) of τ_0; in (2.7) the contour $\mathscr{C}$ encloses the $\tilde{n}(\tau_0)$ roots of $f(\sigma,\tau_0) = 0$ in $|\sigma| < 1$, is contained in $|\sigma| < 1$, and avoids a neighborhood of the point $\sigma_1(\tau_0)$. Thus, in $U_{\tau_0} \cap \mathscr{D}_1$, $\tilde{n}(\tau) = \tilde{n}(\tau_0) = n_1$, whereas in $U_{\tau_0} \cap \mathscr{D}_2$, $n(\tau) = \tilde{n}(\tau) + 1 = n_1 + 1$. These equalities hold then in all of $\mathscr{D}_1$ and $\mathscr{D}_2$, and this finishes the proof if $\bar{\gamma}_1$ is simple. If $\sigma_1(\tau)$ is not one-valued or $\bar{\gamma}_1$ is multiple, we have to allow for $k_1 + k_2 + \cdots k_{m_1}$ zeros to approach $|\sigma| = 1$ as τ approaches γ_1 from $\mathscr{D}_2$ and for $k'_1 + k'_2 \cdots k'_{\bar{m}1}$ zeros if γ_1 is approached from $\mathscr{D}_1$ but the reasoning is strictly unchanged. This ends the proof of Theorem 2.1.

Let $\sigma_l^{(i)}(\tau)$, $l = 1,2,\cdots$, be the N_i roots of $f(\sigma,\tau)$ for τ inside $\mathscr{D}_i$. Since N_i is constant in $\mathscr{D}_i$, it follows that the functions

$$S_{i,k}(\tau) \equiv \sum_{l=1}^{N_i} (\sigma_l^{(i)}(\tau))^k \tag{2.8}$$

are holomorphic for τ in $\mathscr{D}_i$. This follows from the symmetry of these sums under permutation of the $\sigma_l^{(i)}(\tau)$ and can also be checked from the representation:

$$S_{i,k}(\tau) = \frac{1}{2\pi i}\oint_{|\sigma|=1} \sigma^k \frac{\partial f(\sigma,\tau)/\partial\sigma}{f(\sigma,\tau)}\, d\sigma \tag{2.9}$$

[cf. Eq. (A7) of Appendix A].

We can now prove:

Theorem 2.2: Define, with the same sense of integration as in (2.4),

$$\Omega^{(k)}(\tau) = \sum_j \frac{1}{2\pi i}\int_{\gamma_j} \frac{(\sigma_j(\tau'))^k}{\tau' - \tau}\, d\tau' + \sum_j \frac{1}{2\pi i}\int_{\tau_j, |\tau'|=1}^{\tau_{j+1}} \frac{\Sigma_{l=1}^{n_j} (\sigma_{pl}(\tau'))^k}{\tau' - \tau}\, d\tau'. \tag{2.10}$$

The summation in the second integral is performed over the n_j roots $\sigma_{pl}(\tau')$ of $f(\sigma,\tau) = 0$, $\tau' \in (\tau_j, \tau_{j+1})$. Then:

$$\Omega^{(k)}(\tau) = S_{i,k}(\tau) \tag{2.11}$$

for τ in $\mathscr{D}_i$ and

$$\Omega^{(k)}(\tau) = 0 \tag{2.12}$$

for $|\tau| > 1$.

Theorem 2.2 achieves the construction of the Cousin data: from information available on the boundary of the analyticity domain, we construct the symmetric combinations of the roots $\sigma_k^{(i)}(\tau)$ of $f(\sigma,\tau) = 0$ for τ in each $\mathscr{D}_i$. If one knows these symmetric combinations and the numbers N_i of the roots in $\mathscr{D}_i$ by (2.4), one can construct the roots explicitly (see below) at each τ, $|\tau| < 1$. Equation (2.12) can be seen as a condition for the "correct" resolution of the discrete ambiguity (see below).

Proof: It is enough to consider $S_{j,1}(\tau)$. We prove first that, if $\bar{\gamma}_j$ is the boundary between $\mathscr{D}_j$ and $\mathscr{D}_{j+1}$, and has multiplicity one and a single-valued inverse function $\sigma_j(\tau)$ ($\mathscr{D}_{j+1}$ lies to the left of γ_j)

$$S_{j+1,1}(\tau) - S_{j,1}(\tau) = \sigma_j(\tau). \tag{2.13}$$

For this, we simply repeat the argument at the end of Theorem 2.1, i.e., show the continuity of

$$\tilde{S}_j(\tau) = \frac{1}{2\pi i}\oint_{\mathscr{C}} \sigma \frac{\partial f/\partial\sigma(\sigma,\tau)}{f(\sigma,\tau)}\, d\sigma \tag{2.14}$$

across γ_j, for $\mathscr{C}$ chosen as shown there, i.e., a contour contained in $|\sigma| < 1$ enclosing all the roots in $|\sigma| < 1$ of $f(\sigma,\tau) = 0$, for $\tau_0 \in \bar{\gamma}_j$, and avoiding a neighborhood of the point $\sigma_j(\tau_0)$. [$\tilde{S}_j$ is the sum over N_j roots of $f(\sigma,\tau) = 0$.] If the curve γ_j is multiple or has a many-valued inverse function $\sigma_{jk}(\tau)$, Eq. (2.13) generalizes in a straightforward manner to

$$S_{j+1,1}(\tau) - S_{j,1}(\tau) = \sum_{l=1}^{m_{jl}} \sigma_{jl}(\tau) - \sum_{l=m_{jl}+1}^{m_j} \sigma_{jl}(\tau), \tag{2.13'}$$

where the various inverse functions pertaining to $\bar{\gamma}_j$ have been ordered so that the first m_{jl} correspond to the sense on γ_j leaving $\mathscr{D}_{j+1}$ to the left.

We write now Cauchy's theorem for each $S_{i,1}(\tau)$,

$i = 1,2, \cdots,$

$$\frac{1}{2\pi i}\oint_{\partial \mathscr{D}_i} \frac{S_{i,1}(\tau')}{\tau' - \tau} d\tau' = \begin{cases} S_{i,1}(\tau), & \tau \in \mathscr{D}_i \\ 0, & \tau \in \mathrm{C}\mathscr{D}_i \end{cases}, \tag{2.15}$$

where $\partial \mathscr{D}_i$ is the boundary of $\mathscr{D}_i$ and integration is performed in a counterclockwise sense. We now claim that, adding Eq. (2.15) for all i, for a fixed $\tau \in \mathscr{D}_k$, we obtain a value, $S_1(\tau)$, which is precisely the value assumed by $\Omega^{(1)}(\tau)$, Eq. (2.10) at that $\tau \in \mathscr{D}_k$. Indeed, we recognize the contributions of the circle $|\tau| = 1$ are the same in both $S_1(\tau)$ and $\Omega^{(1)}(\tau)$. Consider now the inner boundary $\bar{\gamma}_j$ (simple and with single-valued inverse) separating $\mathscr{D}_j$ and $\mathscr{D}_{j+1}$. Both $S_1(\tau)$ and $\Omega^{(1)}$ contain a term corresponding to the integral over this boundary of $\sigma_j(\tau')/(\tau' - \tau)$; we only have to verify that the sense is the same. Assume that, with respect to the sense in Eq. (2.10), $\mathscr{D}_{j+1}$ lies to the left of γ_j. Then, Eq. (2.13) holds. It follows that, in $S_1(\tau)$, the integral is performed in the sense induced by the counterclockwise sense in $\mathscr{D}_{j+1}$, that is, so that $\mathscr{D}_{j+1}$ lies to the left of the integration path in (2.15). Thus the senses in $S_1(\tau)$ and $\Omega^{(1)}(\tau)$ are the same. There is no difficulty in generalizing this reasoning to arbitrary $\bar{\gamma}_j$. We conclude $S_1(\tau) \equiv \Omega^{(1)}(\tau)$ and the properties of $\Omega^{(1)}(\tau)$ in the theorem follow from Eq. (2.15). This ends the proof.

Given the sums $S_{i,k}(\tau)$ and the numbers N_i, we can construct in each $\mathscr{D}_i$ the symmetric combinations:

$$\begin{aligned} \alpha_{1,i}(\tau) &= -\sum_{k=1}^{N} \sigma_k^{(i)}(\tau) \quad (\equiv S_{i,1}(\tau)), \\ \alpha_{2,i}(\tau) &= \sum_{k<l} \sigma_k^{(i)}(\tau)\sigma_l^{(i)}(\tau), \\ &\;\;\vdots \\ \alpha_{N_i,i}(\tau) &= \prod_k \sigma_k^{(i)}(\tau) * (-1)^{N_i}. \end{aligned} \tag{2.16}$$

These are the coefficients of a polynomial $P_i(\sigma,\tau)$, of degree N_i in σ. this polynomial vanishes in $\mathscr{D}_i$ precisely at those points where $f(\sigma, \tau)$ vanishes and thus gives a complete description of the Cousin data. Explicit formulae are known (Newton's formulae, Ref. 21, p. 74) to compute the $\alpha_{k,i}(\tau)$ from the $S_{k,i}(\tau)$. It is amusing that, by using them and the notations of Eqs. (2.4) and (2.10), it is possible to write an expression for $P_i(\sigma,\tau)$ valid for all $|\tau| < 1$. Let $N = \max_{|\tau|<1} \mathscr{N}(\tau)$. Then we construct

$$\mathscr{F}(\sigma, \tau) = \frac{1}{\sigma^{N-\mathscr{N}(\tau)}} \sum_{p=0}^{N} \alpha_p(\tau)\sigma^{N-p} \tag{2.17}$$

with

$$\alpha_0(\tau) = 1, \alpha_1(\tau) = -\Omega^{(1)}(\tau) \tag{2.18}$$

$$\alpha_p(\tau) = \frac{(-1)^p}{p!} \begin{vmatrix} \Omega^{(1)}(\tau) & 1 & 0 \cdots 0 \\ \Omega^{(2)}(\tau) & \Omega^{(1)}(\tau) & 2 \cdots 0 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ & & p-1 \\ \Omega^{(p)}(\tau) & \Omega^{(p-1)}(\tau) & \cdots \Omega^{(1)}(\tau) \end{vmatrix}.$$

Resorting to Newton's formulas, one can check that, indeed, $\mathscr{F}(\sigma,\tau) = P_i(\sigma,\tau)$, $\tau \in \mathscr{D}_i$. Formula (2.18) ensures that $\alpha_p(\tau)$ vanishes for $p > \mathscr{N}(\tau)$ (see Ref. 21, p. 74) and the factor $\sigma^{N-\mathscr{N}(\tau)}$ cancels the $n - \mathscr{N}(\tau)$ unwanted zeros of the numerator at $\sigma = 0$. The function $\mathscr{F}(\sigma,\tau)$ achieves, in a compact notation, the construction of the Cousin data, from information available on the boundary of $|\sigma| \leqslant 1 \otimes |\tau| \leqslant 1$.

Before closing this section, we comment on the significance of Eq. (2.12). If we start the construction of the amplitude from the modulus of $f(\sigma, \tau)$, available on the set $\mathscr{P}$, we know the location of γ_i and the values of $\sigma_i(\tau)$ on $|\tau| = 1$ only up to a twofold ambiguity. For each solution of this ambiguity, we can construct the functions $\Omega^{(k)}(\tau)$ [Eq. (2.10)]. However only few solutions will obey (2.12). Only if (2.12) is fulfilled, are we sure that the value of $\Omega^{(k)}(\tau)$ in a domain $\mathscr{D}_1$, adjacent to $|\tau| = 1$, approaches indeed, as $|\tau| \to 1$, the value $S_{1,k}(\tau)$, obtained from the known zero trajectories. In other words, for a correct solution of the discrete ambiguity, (a) $S_{1,k}(\tau)$, known on an interval of $|\tau| = 1$ should admit of an analytic continuation to all of $\mathscr{D}_1$ and (b) this continuation should be identical with the value assumed in $\mathscr{D}_1$ by $\Omega^{(k)}(\tau)$. The fact that requirement (a) is not satisfied for an arbitrary resolution of the discrete ambiguity has been pointed out (essentially) in Refs. 14 and 15. In order to choose a correct solution of the discrete ambiguity, one required that, in simple cases, the sum and product of two zero trajectories be analytic in a domain as large as possible around the physical region ($|\tau| = 1$). Clearly, Eq. (2.12) gives the correct mathematical expression for this method. Notice that its formulation does not require any continuation of $f(\sigma, \tau)$ to the second sheet. Also, it is not guaranteed that only one resolution of the discrete ambiguity satisfies (2.12). There may be several correct solutions (see Ref. 12).

III. THE DIRECT PRODUCT OF TWO UNIT DISKS. CONSTRUCTION OF THE AMPLITUDE

The first part of this section is concerned with the construction of a function, $C(\sigma,\tau)$, the "Cousin function," holomorphic in both σ and τ in $\mathscr{B} = \{|\sigma| < 1 \otimes |\tau| < 1\}$ and such that it vanishes in $\mathscr{B}$ precisely at those points where $f(\sigma,\tau)$ vanishes. The second part deals with the construction of the amplitude itself.

The construction of $C(\sigma,\tau)$ follows to some extent Ref. 16, p. 259 ff. The proof of its validity is different and adapted to the needs of Sec. V.

We define first

$$\mathscr{K}_1(\sigma,\tau) \equiv \exp\left[-\sum_i \frac{1}{2\pi i}\int_{\gamma_i} \frac{\ln(\sigma - \sigma_i(\tau'))}{\tau' - \tau} d\tau'\right]. \tag{3.1}$$

The integration is performed in the sense assigned to γ_i in Sec. II. The determinations of the logarithm are to be chosen as indicated in Appendix D. The purpose is to make $\mathscr{K}_1(\sigma,\tau)$ one-valued and real analytic. We show in Appendix D that this is possible, if the amplitude $f(\sigma,\tau)$ from which the zero trajectories are derived is real analytic. However, the choice of branches will generate discontinuities of $2k_j\pi i$, k_j an integer, in the numerators of the integrands of Eq. (3.1) at certain points τ_{0j}, on the curves γ_i; the positions of τ_{0j} and the magnitude of the discontinuities are uniquely determined by the prescriptions of Appendix D, given the curves $\gamma_j : \tau = \tau_j(\sigma)$. As a consequence, the function $\mathscr{K}_1(\sigma,\tau)$ acquires poles at fixed values of $\tau = \tau_{0j}$ in $|\tau| < 1$.

To make this point clear, we give a (simple) example and refer for the general discussion to Appendix D. Consider $f(\sigma,\tau) = \sigma/2 - \tau$; the disk $|\tau| < 1$ contains just one curve γ_1: the circle $|\tau| = \frac{1}{2}$; the inverse function is $\sigma_1(\tau') = 2\tau'$, $|\tau'| = \frac{1}{2}$. For σ real in Eq. (3.1), $-1 < \sigma < 1$, it is natural to choose that determination of the logarithm for which $\operatorname{Im}\ln(\sigma - \sigma_1(-\frac{1}{2})) = \operatorname{Im}\ln(\sigma + 1) = 0$. This implies, by continuity along $|\tau'| = \frac{1}{2}$, that $\operatorname{Im}\ln(\sigma - \sigma_1(\frac{1}{2} + i\epsilon)) = -\pi$, $\operatorname{Im}\ln(\sigma - \sigma_1(\frac{1}{2} - i\epsilon)) = \pi$ $(\epsilon > 0)$. One can verify (cf. Appendix E, Example 1) that $\mathscr{K}_1(\sigma,\tau)$ is indeed real analytic with this choice. The integral in (3.1) along γ_1 has a logarithmic singularity at $\tau = \frac{1}{2}$; for $(\tau - \frac{1}{2})$ small, it behaves like $\ln(\frac{1}{2} - \tau)$. This leads to a pole of $\mathscr{K}_1(\sigma,\tau)$ at $\tau = \frac{1}{2}$, for all σ.

It is thus natural to define

$$\mathscr{K}(\sigma,\tau) = \prod_j (\tau - \tau_{0j})^{k_j} K_1(\sigma,\tau), \tag{3.2}$$

where the product runs over the possible τ_{0j}, where discontinuities of the numerator of the integrand occur and k_j are the integers (positive) giving the magnitude of the discontinuity in the sense of integration. Clearly, the factors $(\tau - \tau_{0j})^{k_j}$ are supposed to remove the poles of $\mathscr{K}_1(\sigma,\tau)$ at τ_{0j} (see Appendix D). In the example above, $k_1 = +1$, $\tau_{01} = \frac{1}{2}$.

We define further, with $\mathscr{F}(\sigma,\tau)$ of Eq. (2.17)

$$C(\sigma,\tau) = \mathscr{F}(\sigma,\tau)\mathscr{K}(\sigma,\tau) \tag{3.3}$$

and shall prove the following.

Theorem 3.1: $C(\sigma,\tau)$ is (i) holomorphic in $\mathscr{B}$; (ii) holomorphic in τ, $|\tau| < 1$, for σ on $|\sigma| = 1$; (iii) continuous in $(|\sigma| \leqslant 1 \otimes |\tau| < 1)$; (iv) is real analytic; (v) is such that $f(\sigma,\tau)/C(\sigma,\tau)$ has no zeros in $\mathscr{B}$; if $|\sigma| = 1$, this ratio has no zeros in $|\tau| < 1$.

Proof: (i) To prove holomorphy of $C(\sigma,\tau)$ in $\mathscr{B}$, we study a single factor $k_1(\sigma,\tau)$ of $\mathscr{K}(\sigma,\tau)$, corresponding to a curve γ_1 with a single valued inverse $\sigma_1(\tau)$:

$$k_1(\sigma,\tau) = \exp\left[-\frac{1}{2\pi i}\int_{\gamma_1} \frac{\ln(\sigma - \sigma_1(\tau'))}{\tau' - \tau}\, d\tau'\right]. \tag{3.4}$$

Clearly, $k_1(\sigma,\tau)$ is holomorphic in the neighborhood of every point (σ,τ) in $\{|\sigma| < 1\} \otimes \{\tau \text{ plane}\setminus\gamma_1\}$. Because $\{|\sigma| < 1\}$ is simply connected, it follows that $k_1(\sigma,\tau)$ is one-valued and holomorphic in this domain. Consider that part of γ_1 which is the boundary between two domains $\mathscr{D}_1$ and $\mathscr{D}_2$, $\mathscr{D}_2$ lying on the left of γ_1. Let $F_2(\sigma,\tau) \equiv \mathscr{F}(\sigma,\tau)$, $\tau \in \mathscr{D}_2$, $F_1(\sigma,\tau) \equiv \mathscr{F}(\sigma,\tau)$, $\tau \in \mathscr{D}_1$. Equations (2.18) imply that the coefficient of the leading power in σ in all the determinations of $\mathscr{F}(\sigma,\tau)$ is unity. For τ on γ_1:

$$F_2(\sigma,\tau) = (\sigma - \sigma_1(\tau))F_1(\sigma,\tau). \tag{3.5}$$

Applying Plemelj's formulas to $k_1(\sigma,\tau)$, for $\tau \to \gamma_1$, one gets (if $\tau \in \gamma_1$ is not a cusp of γ_1):

$$\lim_{\substack{\tau\to\gamma_1\\ \tau\in\mathscr{D}_2}} k_1(\sigma,\tau) = (\sigma - \sigma_1(\tau))^{-1/2} \times \exp\left[-\frac{1}{2\pi i} P\int_{\gamma_1} \frac{\ln(\sigma - \sigma_1(\tau'))}{\tau' - \tau}\, d\tau'\right], \tag{3.6}$$

$$\lim_{\substack{\tau\to\gamma_1\\ \tau\in\mathscr{D}_1}} k_1(\sigma,\tau) = (\sigma - \sigma_1(\tau))^{1/2} \times \exp\left[-\frac{1}{2\pi i} P\int_{\gamma_1} \frac{\ln(\sigma - \sigma_1(\tau'))}{\tau' - \tau}\, d\tau'\right].$$

From (3.5) and (3.6), we see that $\lim_{\tau\to\gamma_1, \tau\in\mathscr{D}_2} C(\sigma,\tau) = \lim_{\tau\to\gamma_1, \tau\in\mathscr{D}_1} C(\sigma,\tau) = c(\sigma,\tau)$, $c(\sigma,\tau)$ being defined only for τ on γ_1. Thus, defining $C(\sigma,\tau) = c(\sigma,\tau)$ on γ_1, $C(\sigma,\tau)$ is continuous as a function of two real variables at each point of γ_1 (see Ref. 18, p. 38), as we approach it from $\mathscr{D}_2$ or from $\mathscr{D}_1$. We still have to verify that the function so defined is actually holomorphic in $|\sigma| < 1 \otimes (\mathscr{D}_2 \cup \mathscr{D}_1)$. To this end (see, e.g., Ref. 22, p. 309), let $\mathscr{C}$ be a closed curve in $\mathscr{D}_2 \cup \mathscr{D}_1$, intersecting γ_1 in two points P_1 and P_2, and define the function:

$$\overline{C}(\sigma,\tau) = \frac{1}{2\pi i}\oint_{\mathscr{C}} \frac{C(\sigma,\tau')}{\tau' - \tau}\, d\tau'. \tag{3.7}$$

$\overline{C}(\sigma,\tau)$ is holomorphic and thus continuous in τ in the domain D enclosed by $\mathscr{C}$. Adding and subtracting in (3.7) the integrals along the segment P_1P_2 of γ_1 of $c(\sigma,\tau')/(\tau' - \tau)$, we conclude, by Cauchy's theorem, that $\overline{C}(\sigma,\tau)$ is equal to $C(\sigma,\tau)$ in both parts of D, on the right and the left of γ_1. It follows that $\lim_{\tau\to\gamma_1}\overline{C}(\sigma,\tau) = \lim_{\tau\to\gamma_1} C(\sigma,\tau) = c(\sigma,\tau)$ and thus $\overline{C}(\sigma,\tau)$ coincides everywhere in D with our definition of $C(\sigma,\tau)$. Therefore, $C(\sigma,\tau)$ is holomorphic in τ at all points of γ_1 on the boundary between $\mathscr{D}_1$ and $\mathscr{D}_2$ at each fixed σ, $|\sigma| < 1$.

If $\bar{\gamma}_1$ does not have a single-valued inverse or the logarithm has a discontinuity, there is no change of the argument at points of $\bar{\gamma}_1$ where there is no cusp, except that we must modify $k_1(\sigma,\tau)$, Eq. (3.4) by inclusion of factors $(\tau - \tau_{0i})^{k_i}$, to remove unwanted poles [Eq. (3.2)].

At an intersection τ_0 of two γ_i's, $i = 1,2$, so that $d\sigma_i(\tau)/d\tau \neq 0$ and finite, $i = 1,2$, we consider two factors $k_i(\sigma,\tau)$, $i = 1,2$ [Eq. (3.4)]. If the four adjacent domains are $\mathscr{D}_l$, $l = 1, \ldots, 4$, one checks as before that the four functions $F_l(\sigma,\tau)k_1(\sigma,\tau)k_2(\sigma,\tau)$, $l = 1, \ldots, 4$ are the analytic continuations of each other through γ_1, γ_2 and that they are bounded at τ_0. They admit thus of a unique analytic extension at $\tau = \tau_0$, obtained by taking the limit $\tau \to \tau_0$ from any of the $\mathscr{D}_l$'s.

The discussion of Appendix E shows that $C(\sigma,\tau)$ is holomorphic in τ even at cusps of γ_i, at each fixed σ, $|\sigma| < 1$.

Further, for fixed τ, the expressions (3.6) and (E1) are holomorphic functions of σ, $|\sigma| < 1$. Using Hartogs' theorem (see Ref. 16, p. 227 and Ref. 23), it follows that $C(\sigma,\tau)$ is indeed a holomorphic function of both σ and τ, in $|\sigma| < 1 \otimes |\tau| < 1$.

(ii) If $|\sigma| = 1$, $C(\sigma,\tau)$ is not well defined, since the argument of the logarithm in (3.4) may vanish. We *define* $C(\sigma_0, \tau)$ for $|\sigma_0| = 1$, $\tau_1(\sigma_0) = \tau_0$, $\tau \notin \gamma_1$, $\tau_0 \in \gamma_1$ by taking the limit $\sigma \to \sigma_0$, $|\sigma| < 1$, along any direction nontangent to $|\sigma| = 1$, at fixed τ. Since the integration along γ_1 is performed so that

$\sigma_1(\tau')$ moves counterclockwise on $|\sigma| = 1$, it is natural to define

$$\arg(\sigma_0 - \sigma_1(\tau_0 + 0)) - \arg(\sigma_0 - \sigma_1(\tau_0 - 0)) = \pi. \quad (3.8)$$

This convention defines $\mathscr{k}_1(\sigma,\tau)$ properly for σ on $|\sigma| = 1$ and τ not on γ_1 (the symbols ± 0 refer to the sense of integration on γ_1). The function thus defined is holomorphic in τ, τ not on γ_1. For τ on γ_1, we can use Eq. (3.6) for $|\sigma| < 1$ and let $\sigma \to \sigma_0$, $|\sigma_0| = 1$, using Eq. (3.8). There is no difficulty to conclude that $C(\sigma_0, \tau)$ is holomorphic in τ, $\tau \in \gamma_1$, except if τ is such that $\tau_1(\sigma_0) = \tau$. Then the principal part integral becomes meaningless and a separate study is needed. We perform it in Appendix E and conclude that $C(\sigma_0, \tau)$ vanishes in fact at such points and is holomorphic there.

(iii) The continuity of $C(\sigma,\tau)$ at the boundary points can be read off the definition (3.1)–(3.3) of $C(\sigma,\tau)$ and its extension to $|\sigma| = 1$, except for those points with $|\sigma| = 1$, $\tau = \tau_i(\sigma) \in \gamma_i$, $|\tau| < 1$. The continuity at such points is shown in Appendix E.

(iv) One must check that, for σ,τ real, $C(\sigma,\tau)$ is real. The reality of $\mathscr{K}(\sigma,\tau)$ is proven in Appendix D (for the choice of branches of the logarithm indicated there). Further, since the roots of $f(\sigma,\tau) = 0$ at real τ are either real or fall in complex conjugate pairs, the sums $S_{i,k}(\tau)$ [Eq. (2.8)], are real, and this is also true of the coefficients $\alpha_{i,k}(\tau)$ [Eq. (2.16)]. Thus $C(\sigma,\tau)$ is indeed real for σ,τ real.

(v) If $|\sigma_0| = 1$, we must show that $C(\sigma_0, \tau)$ vanishes at those points $\tau \in \gamma_i$ so that $\sigma_i(\tau) = \sigma_0$. As stated above, this is not transparent from Eq. (3.6), but we show in Appendix E that this is nevertheless, the case.

In $\mathscr{B}$, $\mathscr{K}(\sigma,\tau)$ does not vanish, so that $C(\sigma,\tau)$ vanishes where $\mathscr{F}(\sigma,\tau) = 0$. For each $\tau \in \mathscr{D}_i$, $\tau \notin \gamma_i$, $\mathscr{F}(\sigma,\tau)$ vanishes by construction in $|\sigma| < 1$ wherever $f(\sigma,\tau)$ vanishes, with the same order and nowhere else. Then, it follows that $f(\sigma,\tau)/C(\sigma,\tau)$ is free of zeros in $|\sigma| < 1 \otimes \mathscr{D}_i$. For $\tau \in \gamma_i$, the zeros of $f(\sigma,\tau)$ in $|\sigma| < 1$ are given by the determination of $\mathscr{F}(\sigma,\tau)$ in the domain $\mathscr{D}_i$ lying to the right of γ_i. However, as seen in the proof of Theorem 2.2 [cf. eq. (3.5)], these zeros are the same as those obtained from the values of $\mathscr{F}(\sigma,\tau)$ in $\mathscr{D}_{i+1}$ as $\tau \to \gamma_i$. Thus, the zeros of $C(\sigma,\tau)$ in $|\sigma| < 1$ are well defined for τ on γ_i and $f(\sigma,\tau)/C(\sigma,\tau)$ is nonvanishing in $|\sigma| < 1$ even for τ on γ_i.

This concludes the proof of Theorem 3.1.

We can now proceed to the construction of $f(\sigma,\tau)$. We have seen that the function

$$E(\sigma,\tau) = f(\sigma,\tau)/C(\sigma,\tau) \quad (3.9)$$

is holomorphic in $\mathscr{B}$ and nonvanishing in $\{|\sigma| \leqslant 1 \otimes |\tau| < 1\}$. For $|\sigma| = 1$, $|\tau| < 1$, we know from analytic continuation of the modulus of $f(\sigma,\tau)$ away from $\mathscr{P}_\sigma$ [cf. Eq. (2.2)] the values of the combination [$\sigma = \exp(i\alpha)$]:

$$E(e^{i\alpha},\tau)E^*(e^{i\alpha}, \tau^*) = M_{\sigma = e^{i\alpha}}(\tau)/(C(e^{i\alpha},\tau)C^*(e^{i\alpha},\tau^*))$$
$$= g_1(e^{i\alpha},\tau). \quad (3.10)$$

Since, by Theorem 3.1, $C(e^{i\alpha},\tau)$ is a holomorphic function of τ, it follows that $g_1(e^{i\alpha},\tau)$ is a known, real holomorphic function of τ in $|\tau| < 1$.

We now define

$$L(\sigma,\tau) = \ln(\pm E(\sigma,\tau)), \quad (3.11)$$

where the sign is chosen to make $L(\sigma,\tau)$ real holomorphic in $\mathscr{B}$. We rewrite then (3.10) as

$$L(e^{i\alpha},\tau) + L(e^{-i\alpha},\tau) = \ln g_1(e^{i\alpha},\tau), \quad (3.12)$$

where we have used $L^*(\sigma,\tau^*) = L(\sigma^*,\tau)$. At each fixed τ, it is apparent that $L(e^{i\alpha},\tau)$ can be extended analytically to $|\sigma| < 1$ and $L(e^{-i\alpha},\tau)$ to $|\sigma| > 1$. We expect in fact we can express $L(\sigma,\tau)$ by means of a Cauchy integral over its boundary values, $L(e^{i\alpha},\tau)$. Assume for the moment we can do this. It is then also true that

$$\oint_{|\sigma'|=1} \frac{L(1/\sigma',\tau) - L(0,\tau)}{\sigma' - \sigma} d\sigma' = 0. \quad (3.13)$$

We obtain then from Eq. (3.12) that

$$L(\sigma,\tau) + L(0,\tau) = \frac{1}{2\pi} \oint \frac{\ln g_1(e^{i\alpha},\tau)}{1 - \sigma e^{-i\alpha}} d\alpha. \quad (3.14)$$

From (3.14) we obtain $L(\sigma,\tau)$ in all of $\mathscr{B}$:

$$L(\sigma,\tau) = \frac{1}{2\pi} \oint \frac{e^{i\alpha} + \sigma}{e^{i\alpha} - \sigma} \ln g_1(e^{i\alpha},\tau)\, d\alpha. \quad (3.15)$$

From Eqs. (3.15) and (3.11) we can calculate $E(\sigma,\tau)$ in all of $\mathscr{B}$ up to a sign ambiguity. This achieves the desired construction of $f(\sigma,\tau)$:

$$f(\sigma,\tau) = \pm C(\sigma,\tau)E(\sigma,\tau). \quad (3.16)$$

The sequence of equations used implicitly in (3.16) is (3.15), (3.10), (3.11), (3.9), (3.3), (3.2), (3.1), (2.18), (2.17), (2.16), and (2.5).

To justify the use of Cauchy's formula for $L(\sigma,\tau)$ in Eqs. (3.13) and (3.14), one needs to verify that $\oint |L(re^{i\alpha},\tau)|\, d\alpha$ stays bounded as $r \to 1$ (see Ref. 24, p. 41). This follows in turn if $\oint |\ln C(re^{i\alpha},\tau)| d\alpha$ is bounded, as $r \to 1$, and we only have to verify this for the exponents of each factor $\mathscr{k}_i(\sigma,\tau)$, corresponding to the various curves γ_i in Eq. (3.1). This is done in Appendix E, Subsection (d).

Now, by symmetry, we could have written an equation similar to (3.10) for $|\tau = \exp(i\beta)| = 1$, $|\sigma| < 1$:

$$E(\sigma,e^{i\beta})E^*(\sigma^*,e^{i\beta}) = M_{\tau = e^{i\beta}}(\sigma)/(C(\sigma,e^{i\beta})C^*(\sigma^*,e^{i\beta}))$$
$$\equiv g_2(\sigma,e^{i\beta}), \quad (3.17)$$

where $g_2(\sigma,e^{i\beta})$ is holomorphic in $|\sigma| < 1$. Indeed, if $|\tau| = 1$, but not an end point of a γ_i, $C(\sigma,\tau)$ vanishes in $|\sigma| < 1$ only at those points where $\mathscr{F}(\sigma,\tau)$, Eq. (2.17) vanishes, that is, by construction, wherever $f(\sigma,\tau)$ vanishes. As a consequence, $g_2(\sigma,e^{i\beta})$ is also free of zeros in $|\sigma| < 1$ for almost all τ on $|\tau| = 1$. The problem is that the expression $C(\sigma,e^{i\beta})$ is not well defined for all β. We show namely in Appendix E that, at the end points τ_{ei} of the curve γ_i, $C(\sigma,\tau)$ is singular, in a manner that is dependent on σ. However, for each fixed σ, $\oint |\ln C(\sigma, re^{i\beta})|\, d\beta$ is bounded ar $r \to 1$, as one verifies from Eq. (E17) and one can consequently still write a dispersion relation for $L(\sigma,\tau)$ at fixed σ. We obtain then

$$L(\sigma,\tau) = \frac{1}{2\pi} \oint \frac{e^{i\beta} + \tau}{e^{i\beta} - \tau} \ln g_2(\sigma,e^{i\beta})\, d\beta. \quad (3.18)$$

The fact that the result $L(\sigma,\tau)$ must be the same as in Eq. (3.15) shows that the functions M_σ and M_τ used in (3.10) and (3.17) cannot be prescribed independently. The equality of (3.15) and (3.18) gives a condition that has to be fulfilled by the modulus functions on $\mathscr{P}_\sigma$ and $\mathscr{P}_\tau$ in order that the problem admits of a solution.

IV. THE MANDELSTAM DOMAIN. CONSTRUCTION OF THE COUSIN DATA

In order to adapt the procedure of Sec. II and III to the construction of amplitudes $A(s,t,u)$, holomorphic in the Mandelstam domain $\mathscr{M}$, one has to overcome the difficulties that $\mathscr{M}$ has not the simple topological structure of the direct product of two domains in the s and t planes (which might appear as a major obstacle) and, more formally, get rid of the assumption that $A(s,t,u)$ can be continued some finite distance on the second sheet.

In this section, we deal with the latter point and introduce a number of assumptions on $A(s,t,u)$, sufficient to ensure the validity of the operations of Secs. II and III, without invoking second sheet continuations. It turns out that the construction of the Cousin data can then proceed, essentially in the same way as in Sec. II. The topological difference between $\mathscr{M}$ and the domain $\mathscr{B}$ of Secs. II and III is apparent in the construction of the Cousin function $C(s,t,u)$; the way in which the author believes this difficulty may be circumvented is described in Sec. V.

The Mandelstam domain $\mathscr{M}$ consists of the set of points $(s,t) \in \mathbb{C}^2$, lying in the complement of $C_s^0 \cup C_t^0 \cup C_u^0$, with $C_s^0 = \{(s,t)|s \in C_s\}, C_s = \{s|\text{Im}\, s = 0, s \geqslant 4m^2\}$ and similarly for C_t^0, C_u^0, $s+t+u=4m^2$, m = mass of the pion. We denote by $D_s = \{s|s \in \complement C_s\}$ and, with similar meaning, D_t, D_u. It is convenient to regard $\mathscr{M}$ as $D_s \otimes D_t \cap D_u \otimes D_t$.

An amplitude $A(s,t,u)$ is said to have Mandelstam analyticity if (a) it is holomorphic in the Mandelstam domain, (b) as s approaches $s_0 \in C_s$, $\text{Im}\, s>0$, the limiting values $A(s_0,t)$ exist and are a holomorphic function of t in a plane cut for $t \geqslant 4m^2$, $t \leqslant -s_0$; (c) property (b) holds also if C_s is replaced by C_t, C_u and $A(s_0,t)$ by the corresponding limiting functions. Further, $A(s,t,u)$ is real analytic: $A(s,t,u) = A^*(s^*,t^*,u^*)$; since in general $A(s_0+i\epsilon,t) \neq A(s_0-i\epsilon,t)$, $s_0 \in C_s$, we denote by C_s^+ and C_s^- the upper and lower "lips" of the "cut" C_s. For each s in $D_s \cup C_s^+ \cup C_s^-$, we denote by Q_s the interior of the corresponding cut t plane, i.e., the interior of the set of values of t, so that $(s,t) \in \overline{\mathscr{M}}$. Similar meaning is attached to Q_t and Q_u.

The modulus of $A(s,t,u)$ is measured in the physical region of the three channels, assumed to extend down to s,t or $u=4m^2$ (normal thresholds). By analytic continuation in angle at fixed energy in each channel (see Introduction) we are free to suppose that $|A(s,t,u)|$ is known in an "extended physical region" $\mathscr{P}_s \cup \mathscr{P}_t \cup \mathscr{P}_u$, where $\mathscr{P}_s = \{(s,t) \in \mathbb{R}|s \geqslant 4m^2, u \leqslant 4m^2, t \leqslant 4m^2\}$, etc.

The hypothesis that $f(\sigma,\tau)$ can be extended to the second sheet was used in Sec. II mainly to justify the picture of the curves γ_i: $\tau = \tau_i(\sigma)$ in $|\tau|<1$. The restrictions we are introducing next concerning $A(s,t,u)$ essentially state directly the validity of the properties of γ_i, as derived in Appendix B.

(A1) The amplitude $A(s,t)$ is continuous as a function of four real variables at all points on the (finite) boundary of $\mathscr{M}$ [i.e., all finite (s,t,u) with at least one of them on the corresponding cut]. The derivatives $\partial A/\partial s$ (fixed t), $\partial A/\partial t$ (fixed s), and $\partial/\partial s(\partial A/\partial t)$ are continuous (as above) at the points on the boundary of $\mathscr{M}$ with $s \in C_s$, except possibly for a finite number of points $s_{t,i}$, where $\partial A/\partial s$, $\partial/\partial s(\partial A/\partial t)$ might be unbounded. A similar statement is true for that part of the boundary of $\mathscr{M}$ with $t \in C_t$ or $u \in C_u$.

Let us notice that the statement about the continuity of the derivatives at the boundary means, e.g., that at fixed $t \in \overline{D}_t$, the holomorphic function of $s, s \in Q_t$, $\partial A/\partial s(s,t)$ has a well-defined limit at points of the cut C_s and implies (see, e.g., Ref. 24, p. 42, Theorem 3.11) that the value of this limit can be computed from the variation of the complex function $A(s,t)$ along the cut. In particular, the derivative with respect to s in the direction normal to the cut exists and is equal to the derivative along the cut (the Cauchy–Riemann equations are valid for $s \in C_s$).

As a consequence of (A1), we can make some statements about the behavior with respect to s of the isolated solutions of the equation $A(s,t)=0$, lying in the interior Q_s of the cut complex t plane, as s moves along C_s or in its neighborhood. A solution $t(s)$ is isolated for s in some interval (or domain) I if, for each $s \in I$, there exists a disk d_r: $|t-t(s)|<r(s)$, $d_r \subset Q_s$, so that $A(s,t)$ vanishes in d_r only at $t=t(s)$.

Lemma 1: If $A(s_0,t)$, $s_0 \in C_s^+$, $s_0 \neq s_{t,i}$, has a simple root at t_0, $t_0 \in Q_{s_0}$, there exists a neighborhood U_r of s_0: $U_r = \{s||s-s_0|<r, \text{ Im}\, s>0\}$ and a unique solution $t(s)$ of the equation $A(s,t)=0$ for $s \in \overline{U}_r$ satisfying $t_0 = t(s_0)$, $A(s,t(s)) \equiv 0$, $s \in \overline{U}_r$. The solution is isolated for $s \in \overline{U}_r$. The function $t(s)$ is holomorphic for $s \in U_r$ and with a continuous derivative for $s \in \overline{U}_r$.

This can be proved from the representation:

$$t(s) = \frac{1}{2\pi i}\oint_{|t-t_0|=r_1} t\,\frac{\partial A/\partial t}{A}(s,t)\,dt \tag{4.1}$$

obviously valid for $s=s_0$ and r_1 chosen so that we enclose just the zero at t_0 and $A(s_0,t) \neq 0$ for $|t-t_0|=r_1$. Since $A(s,t)$ is continuous on the closed set $(|s-s_0| \leqslant a$, $\text{Im}\, s \geqslant 0) \otimes (|t-t_0| \leqslant r_2)$ $(r_2>r_1)$ (for some $a>0$), it is uniformly continuous there and thus we can find r, valid for all t on $|t-t_0|=r_1$ so that $A(s,t) \neq 0$ for $|t-t_0|=r_1$ and $s \in \overline{U}_r$: $\{s||s-s_0| \leqslant r, \text{ Im}\, s \geqslant 0\}$. Therefore, Eq. (4.1) defines for $s \in \overline{U}_r$ a function $t(s)$, which is holomorphic for $s \in U_r$ and with continuous derivative for $s \in \overline{U}_r$ [using assumption (A1)].

To show that $t(s)$ defined by (4.1) satisfies $A(s,t(s)) \equiv 0$ for $s \in \overline{U}_r$, we notice first that, for $s \in \overline{U}_r$, the holomorphic function of t, $A(s,t)$ has only one zero inside $|t-t_0|=r_1$. This follows from the fact that for $s \in \overline{U}_r$, by the same argument as above

$$\frac{1}{2\pi i}\oint_{|t-t_0|=r_1} \frac{\partial A/\partial t}{A}(s,t)\,dt = 1. \tag{4.2}$$

Evaluating (4.1) by the residue theorem, we verify that, indeed, $A(s,t(s)) \equiv 0$ for $s \in \overline{U}_r$. Equation (4.2) shows also that $t(s)$ is an isolated solution of $A(s,t)=0$, $s \in \overline{U}_r$.

If s_0 is a "threshold" $s_{t,i}$, the reasoning of Lemma 1 can still be done, but we only conclude that $t(s)$ is continuous at $s=s_0$.

Thus through any simple zero of $A(s_0,t)$ there passes a differentiable (if $s_0 \neq s_{t,i}$) "zero trajectory" function $t=t_i(s)$, $s \in C_s^+$. The curves Γ_i: $t=t_i(s)$, $s \in C_s^+$, have even a continuous tangent at all points where $dt_i(s)/ds \neq 0$. These curves can be assigned an orientation, by letting s move to increasing values if $s \in C_s^+$ or decrease if $s \in C_s^-$.[25]

We now define the inverse functions $s_i(t)$ for isolated zero trajectories:

Lemma 2: Let $t = t_1(s)$ be an isolated zero trajectory, defined for $|s - s_0| < r$, $s \in C_s^+$, $s_0 \neq \{s_{t,i}\}$, and $dt_1(s)/ds(s = s_0) \neq 0$. Then: (i) for any sufficiently small R, the curve $\Gamma_1: t = t_1(s)$ parts the disk $V_R: |t - t_1(s_0)| < R$ into two disjoint simply connected domains; (ii) let V_R^+ be that part of V_R lying to the left of Γ_1; if R is sufficiently small, there exists a unique function $s_1(t)$, holomorphic in V_R^+ and with continuous derivative in $\overline{V}_R^+$ so that, for s on C_s^+, $t_1(s) \in \Gamma_1 \cap \overline{V}_R^+$, $s_1(t_1(s)) \equiv s$.

For part (i), we use the fact that dt_1/ds is continuous at s_0, and, therefore, $dt_1/ds \neq 0$ for $|s - s_0|$ sufficiently small, $s \in C_s^+$. Then Γ_1 is near s_0 a smooth arc in the sense of Ref. 18. Property (i) follows from statement 1°, p. 9 of Ref. 18.

To prove (ii), we show that the function $t_1(s)$ defined on $\overline{U}_r$ of Lemma 1 realizes a conformal mapping of U_r (for r small enough) onto a domain d_r lying to the left of Γ_1. We shall then see that $t_1(s)$ maps actually $\overline{U}_r$ one-to-one onto $\overline{d}_r$ and $s_1(t)$ is simply the inverse function of this mapping. Finally, we show that $d_r \supset V_R^+$ for R small enough.

From Lemma 1, one deduces that $dt_1(s)/ds$ is a continuous function of $s, s \in \overline{U}_r$. Since $dt_1/ds(s = s_0) \neq 0$, we can choose r so that $dt_1/ds \neq 0$ for $s \in \overline{U}_r$. In particular we can suppose that, e.g., $\mathrm{Re}\, dt_1/ds > 0$, for $s \in \overline{U}_r$. It follows that, if $s', s'' \in \overline{U}_r$, $s' \neq s''$, then $t_1(s') \neq t_1(s'')$. Indeed:

$$\begin{aligned}&\mathrm{Re}((t_1(s'') - t_1(s'))/(s'' - s'))\\ &\quad = \int_0^1 \mathrm{Re}\, \frac{dt_1}{ds}(s' + \lambda(s'' - s'))\, d\lambda > 0. \qquad (4.3)\end{aligned}$$

Thus, $t_1(s)$ applies conformally U_r onto its image domain d_r and applies $\overline{U}_r$ one-to-one onto its image. Since $t_1(s)$ is continuous in $\overline{U}_r$, the image of $\overline{U}_r$ is $\overline{d}_r$. Therefore, $\overline{d}_r$ contains a piece of Γ_1. Further, using the continuity of dt_1/ds in $\overline{U}_r$, the Cauchy–Riemann equations and a mean value theorem, one verifies that if Δs is real and sufficiently small, $t(s_0 + i\Delta s) - t(s_0)$ lies to the left of $t(s_0 + \Delta s) - t(s_0)$, i.e., d_r lies to the left of Γ_1.

Now, the inverse function of a one-to-one continuous mapping of a compact set is continuous itself (see Ref. 26, p. 310) and thus there exists a continuous function $s_1(t)$, defined on $\overline{d}_r$, holomorphic in d_r and such that on Γ_1, $s_1(t_1(s)) \equiv s$. The function $s_1(t)$ is differentiable where $dt_1/ds \neq 0$, i.e., in $\overline{d}_r$. Choose $R_0 < \min|t_1(s) - t_1(s_0)|$, for s on the semicircle $|s - s_0| = r$. Then $V_R^+ \subset d_r$ if $R < R_0$, and this finishes the proof.

We next wish to follow trajectory functions, not necessarily isolated, on larger intervals of C_s^+; we must allow for their intersections or for the possibility of multiple zeros. The following Lemma can be proved exactly as Lemma 1.

Lemma 3: If $A(s_0,t)$ has N zeros inside the bounded domain $D \subset Q_{s_0}(s_0 \neq s_{t,i})$, there exists a neighborhood U_r: $|s - s_0| < r$, $\mathrm{Im}\, s > 0$, of s_0, so that $A(s_0,t)$ has N zeros in D for all s in $\overline{U}_r$; these zeros are the roots of an equation

$$P(s,t) = t^n + a_1(s)t^{n-1} + \cdots + a_n(s) = 0, \qquad (4.4)$$

where the functions $a_1(s), \ldots, a_n(s)$ are holomorphic in U_r, with a continuous derivative in $\overline{U}_r$.

To obtain the statement about the coefficients $a_i(s)$, we use the representation

$$\sum_{i=1}^{N} (t_i(s))^k = \frac{1}{2\pi i} \oint_{\partial D} t^k \frac{\partial A/\partial t}{A}(s,t)\, dt \qquad (4.5)$$

and express the $a_i(s)$ in terms of the sums in (4.5) by means of Newton's formulas, Eq. (2.18).

There are a number of consequences that can be drawn from this lemma.

(i) Consider first an "intersection" of several trajectories: $A(s_0,t)$ has a zero of the nth order at t_0, but for s near s_0 it might have only isolated roots $t_i(s)$. Then the polynomial (4.4) can be written relatively to a domain D: $|t - t_0| < r$, enclosing no other zeros of $A(s_0,t)$, as

$$P(s,t) = (t - t_0)^n + a_1'(s)(t - t_0)^{n-1} + \cdots + a_n'(s) \qquad (4.6)$$

with the coefficients a_k' vanishing at s_0 [cf. Eq. (A1) of Appendix A]. Since the a_k' have a continuous derivative, $P(s,t)$ cannot vanish outside a circle $|t - t_0| = \epsilon$, if we choose $|s - s_0|$ so small that $|a_k'(s)| < \epsilon^k/(2n)$. Thus, the individual isolated trajectories must satisfy $\lim_{s \to s_0} t_i(s) = t_i(s_0) = t_0$, i.e., are continuous at s_0. The problem is whether we can follow them meaningfully on both sides of s_0.

(ii) If a zero $t = t(s)$ of $A(s,t)$ has higher multiplicity over intervals of s, then we verify just as in Lemma 1, that the function $t = t(s)$ is holomorphic in U_r, with continuous derivative in $\overline{U}_r$ (if $\overline{U}_r$ does not contain "thresholds"). The function $t(s)$ has an inverse $s(t)$, constructed as in Lemma 2, if $dt/ds \neq 0$.

(iii) In Appendix A, we show that, given a point (s_0,t_0) in $\mathscr{M}$ with $A(s_0,t_0) = 0$, we can part in a certain neighborhood $U_0 \otimes V_0$ of (s_0,t_0) the set of zeros of $A(s,t)$ passing through (s_0,t_0) in several components; each component is described in $U_0 \otimes V_0$ by an irreducible pseudopolynomial and is uniquely determined by just one root $t = t(s)$ defined on $U_1 \subset U_0$, $s_0 \notin U_1$, through analytic continuation in U_0. It may be that the point s_0 is not a branch point of the function $t(s)$. Given a domain $U_r \otimes V_t \subset \mathscr{M}$, there will exist in general several points $s_{o,k} \in U_r$, which are branch points of functions $t(s)$, $t(s) \in V_t$, describing locally the zeros of $A(s,t)$ in $\mathscr{M}$. We shall postulate by (A2) that such points do not accumulate to C_s, for any choice of V_t.

(iv) If this is granted, we can define unambiguously the continuation past s_0 of a zero trajectory $t = t_1(s)$, defined for $s \in C_s^+$, $s < s_0$, by performing an analytic continuation along a semicircle $|s - s_0| = \epsilon$, $\mathrm{Im}\, s > 0$, until we reach points on C_s^+, with $s > s_0$. Since $A(s,t_1(s)) \equiv 0$ on this path, we can "tie" together zero trajectories $t_i(s)$ on larger intervals of energy, and talk about one "big" trajectory function; we denote the latter also by $t_i(s)$. The phrase "analytic continuation of $t_i(s)$ along C_s" will be used having in mind this procedure.

(v) All the statements above are true if $s \in C_s^+$ is a "threshold" $s_{t,k}$ provided the words "continuous derivative in $\overline{U}_r$" are replaced by "continuous in $\overline{U}_r$."

(A2) The set of points $(\bar{s},\bar{t}) \in \mathscr{M}$ with $A(\bar{s},\bar{t}) = 0$, where $\bar{s}$ is a branch point of a root $t = t(s)$ of $A(s,t)$ does not accumulate at $C_s^0(C_s^0 = \{(s,t)|s \in C_s\})$. Further, there are only a finite number of points $s_k, s_k \in C_s$, which do not have the property that $\lim_{s\to s_k} dt_i/ds$ exists and is finite and nonzero, for any trajectory $t_i(s)$ lying in Q_s. At such points, a change of variable can be done, $\zeta_{i,k} = (s - s_k)^{\nu_{i,k}}$ so that $\lim_{\zeta\to 0} dt_i/d\zeta_{i,k}$ is finite and nonzero. The same is true for $t \in C_t$, $u \in C_u$ and the corresponding trajectory functions.

Several comments concerning (A2) follow.

(i) The functions $t_i(s)$ which can now be defined even at points where they do not correspond to isolated trajectories are, by (A2), Hölder-continuous functions of $s, s \in C_s$. For sufficiently small r, they are holomorphic in $U_r(s_0)$ and continuous in $\overline{U}_r(s_0)$, $s_0 \in C_s$ [by comment (i) following Lemmas 3]; $U_r(s_0) = \{s||s - s_0| < r, \operatorname{Im} s > 0\}$.

(ii) We need to assume that[27] $\lim_{\zeta\to 0} dt/d\zeta$ is nonzero only from one side of s_1, since we can show that it must be the same from the other side and that, in fact, $dt/d\zeta$ is continuous as a function of ζ in the image $\overline{U}_\zeta$ through $\zeta = (s - s_1)^\nu$ of $\overline{U}_r(s_1)$. To prove this, we have to show that $dt/d\zeta$ is bounded in $\overline{U}_\zeta$ by const/ζ^p, for some $p > 0$, since the statement follows then by an application of the Phragmen–Lindelöf theorem to $dt/d\zeta$ at $\zeta = 0$. To see this, we notice that: (a) the derivative of $(s - s_1)^k t(s)$ is continuous on $[s_1 - r, s_1 + r]$, for k sufficiently large; (b) the function $(s - s_1)^k t(s)$ is continuous in $\overline{U}_r$; (c) we can represent $(s - s_1)^k t(s)$ by means of a Poisson formula in terms of its boundary values in U_r, and we can obtain its derivative by differentiating the formula and then integrating by parts (see Ref. 24, p. 43). The result is that $d/ds((s - s_1)^k t(s))$ can be written, up to factors as a Poisson integral, over its continuous boundary values, and therefore is itself continuous, and thus bounded in $\overline{U}_r$. This gives the desired bound on dt/ds.

(iii) It could be that trajectories intersect at some point (s_0,t_0), $s_0 \in C_s$, but, nevertheless, $\lim_{s\to s_0} dt_i/ds$ exists and is finite along one (or several) of them. The same argument as above shows that dt_i/ds is then a continuous function in $\overline{U}_r(s_0)$ (defined as above). The existence of the inverse function associated to each trajectory is then proved exactly as in Lemma 2.

(iv) The argument of comment (ii) above is valid also at thresholds $s_{t,i}$.

(v) Assumption (A2) allows us to define the inverse function $s_1(t)$ on the curve $\Gamma_1 : t = t_1(s)$ even in the neighborhood of a point s_0 where $dt_1/ds = 0$ or is not finite. Indeed, we can show that $t_1(\zeta)$ makes a one-to-one conformal mapping of $\overline{U}_\zeta$ onto its image. If $\overline{U}_\zeta$ is convex, the proof of Lemma 2 carries through unchanged (the curve Γ_1 is "piecewise smooth," according to Ref. 18). If it is not convex, it is easy to verify that, if $\lim_{\zeta\to 0} dt_1/d\zeta$ is finite, then the change of variables $\theta = (t - t_1(s_0))^{1/\nu}$ makes $\lim_{s\to s_0} d\theta/ds$ finite, and we can repeat the argument of the lemma for the inverse function $s_1(\theta)$. We can then go over to $\zeta(t)$.

(vi) The curves $\Gamma_i : t = t_i(s)$ have at most a finite number of cusps; the angle α of the cusp is related to the power $\nu_{i,k}$ of $(s - s_k)^{\nu_{i,k}}$ needed to render $\lim_{\zeta_{i,k}\to 0} dt_i/d\zeta_{i,k}$ finite by $\alpha = \nu_{i,k}\pi - 2n\pi$, with n the smallest number for which $0 < \alpha < 2\pi$.

(vii) The inverse functions $s_i(t)$ are Hölder-continuous functions along Γ_i; they are defined in a neighborhood of Γ_i on their left and can be obtained there starting from an element $s_i^0(t)$ by "analytic continuation along Γ_i." As in Sec. II, this means that, at a point t_0 where $ds_i/dt = 0$, we must continue $s_i(t)$ along a circular path starting clockwise on the left of Γ_i and proceed until the complex number $s_i(t) - s_i(t_0)$ rotates by an angle π. When this has happened, the argument t has reached the branch of Γ_i lying beyond t_0, and the function so obtained coincides with the inverse function already defined on that piece of Γ_i.

(A3) As s moves along C_s^+, there are only a finite number of zero trajectories $t = t_i(s)$ in the complex t plane. The same is true for $t \in C_t^+$, $u \in C_u^+$.

This assumption means (i) for each $s \in C_s^+$, there exists only a finite number $N_z(s)$ of zeros $t_i(s)$ of $A(s,t)$ in the complex t plane Q_s, including the cuts; (ii) the number $N_z(s)$ stays bounded as $s \to \infty$; (iii) among the intervals $I_n(s)$ of C_s^+ on which the zero trajectories $t_n(s)$ are defined ("exist on the physical sheet of the t plane"), there does not exist a sequence $I_{n_k}(s) = (s_k, s_{k'})$ of intervals with $s_k \to \infty$; in fact, there are only a finite number of such intervals $I_n(s)$, of finite or infinite length, (iv) (A3) is true for $s \in C_s^-$ by real analyticity.

We extend (A3) by an assumption analogous to (A), Sec. II: (A3') The set of points $(s,t) \in C_s \otimes C_t$, where $A(s,t) = 0$ consists of isolated points. (The same is true for $C_u \otimes C_t$, $C_s \otimes C_u$.)

We study now the behavior of the trajectory functions $t_i(s)$, $s \in C_s^+$, as their values approach C_t ($t \geqslant 4m^2$, $\operatorname{Im} t = 0$). From (A1) alone, we can prove:

Lemma 4: If for some sequence $s_n \to s_0$, $s_n < s_0$, $s_n \in C_s^+$, $\lim_{n\to\infty} t(s_n) = t_0$, $t_0 \in C_t$, then $\lim_{n\to\infty} t(s'_n) = t_0$, for any other sequence $s'_n \to s_0$, $s'_n < s_0$.

The (simple) proof of this lemma is given in Appendix G.

As a consequence of this lemma, one can verify that the curve $\Gamma_1 : t = t_1(s)$ has even a tangent at the end point (s_0,t_0), if $\partial A/\partial t\,(s_0,t_0) \neq 0$, $\partial A/\partial s(s_0,t_0) \neq 0$. Indeed, in this case, along the trajectory

$$\frac{dt}{ds} = \frac{-\partial A/\partial s}{\partial A/\partial t\,(s,t)} \tag{4.7}$$

and we use (A1) and Lemma 4 to show that the limit of the right-hand side exists, as $s \to s_0$.

The next assumption concerns the behavior of the trajectories at end points:

(A4) If (s_0,t_0) is an end point of a trajectory $t = t(s)$, then $\lim_{s\to s_0}(s - s_0)^\alpha\, dt/ds$ exists and is finite and nonzero for some α with $\operatorname{Re} \alpha < 1$. Further, we assume that the set of points with $dt/ds = -1$ does not accumulate at $(s_0,t(s_0))$.

As a first consequence of (A4), a trajectory $t = t(s)$ with an end point at (s_0,t_0) is Hölder continuous at s_0 and so is its inverse function $s = s(t)$ at t_0.

We now wish to investigate the continuation of the zero trajectory $t(s)$ clockwise along a small semicircle around s_0. We prove, namely:

Lemma 5: Let $t_1(s)$ be the continuation of the trajectory function with the same name, defined for $s < s_0$, $s \in C_s^+$, $\operatorname{Im} t_1(s) > 0$, inside U_r: $|s - s_0| < r$, $\operatorname{Im} s > 0$ $[t_1(s_0) = t_0 \in C_t]$. Then, for any r, sufficiently small, either $t_1(s)$ can be continued throughout U_r [and then (s_0,t_0) is a "false" end point of the trajectory] or the continuation stops along the image $\overline{\Gamma}_2$:$s = s_2(t)$ of a trajectory defined in the t channel for $t > t_0$ and obeying $\lim_{t \to t_0} s_2(t) = s_0$. In this case, $\lim_{s \to \overline{\Gamma}_2} t_1(s)$ exists and coincides there with the inverse function $t_2(s)$ associated to $\overline{\Gamma}_2$.

Proof: We assume $\operatorname{Im} t_1(s) > 0$, so that $t_0 \in C_t^{\,+}$. We first subject the choice of r, defining U_r, to a number of conditions: (i) we can find R, in fact as small as we like, so that $A(s_0,t)$ does not vanish for $|t - t_0| = R$, $\operatorname{Im} t \geqslant 0$. This follows from the analyticity in t of $A(s_0,t)$, its continuity on $C_s \otimes C_t$ and (A3′). Further, by (A1) we can find $r = r(R)$ so small that, for all $s \in U_r$, $A(s,t) \neq 0$ for $|t - t_0| = R$. As a consequence of this choice, all possible continuations of the trajectory functions $t = t(s)$, $s \in U_r$, stay either contained in $V_R: |t - t_0| < R$, $\operatorname{Im} t > 0$ or reach the boundary of V_R on $\operatorname{Im} t = 0$, $t \in [t_0 - R, t_0 + R]$. (ii) Using (A2), dt/ds is finite and $\neq 0$ for $s \in U_r$, $t \in V_R$, if r and R are sufficiently small; otherwise, there would exist a set of points (s',t'), where $dt/ds = 0$ or not finite accumulating to C_s^0 or C_t^0. (iii) We can choose R so small that the image of no trajectories $s = s_i(t)$, defined for $t \in C_t^{\,+}$, $t \in [t_0 - R, t_0 + R]$, crosses U_r, unless these trajectories satisfy $\lim_{t \to t_0} s_i(t) = s_0$. Further, R and r can be chosen so that none of these trajectories have two end points in $[s_0 - r, s_0 + r] \otimes [t_0 - R, t_0 + R]$. As a consequence, the images of the various trajectories $\overline{\Gamma}_i$:$s = s_i(t)$ that satisfy $\lim_{t \to t_0} s_i(t) = s_0$ are open curves when restricted to U_r. (iv) The functions $s = s_i(t)$ have a continuous and nonvanishing derivative for $t \in [t_0 - R, t_0)$ or $(t_0, t_0 + R]$. As a consequence, the curves Γ_i cross any semicircle $|s - s_0| = r_1 < r$ in one point, for one value $t \in [t_0 - R, t_0)$ or $(t_0, t_0 + R]$ and in particular have no multiple points in U_r. (v) $A(s,t) \neq 0$ if $(s,t) \in [s_0 - r, s_0 + r] \otimes [t_0 - R, t_0 + R]$, $(s,t) \neq (s_0,t_0)$. Let r_0 be a number satisfying these conditions and U_{r_0} the corresponding neighborhood.

Consider now the continuation of $t(s)$ along a path $\mathscr{P}_r$: $|s - s_0| = r$, $\operatorname{Im} s > 0$, $r < r_0$, starting on $[s_0 - r_0, s_0)$. Continuation along $\mathscr{P}_r$ stops only if, for some sequence $s_n \to \bar{s}$, $t(s_n) \to \bar{t}$, $\operatorname{Im} \bar{t} = 0$. If we can find no such sequence for any semicircle $\mathscr{P}_r$, then $t(s)$ can be continued throughout U_{r_0} and we are in the first case of the lemma. If such a sequence exists, then, with the same reasoning as in Lemma 4, we can verify that, for any other sequence $s_n' \to \bar{s}$, $\lim_{n \to \infty} t(s_n') = \bar{t}$; by (v) above [possible by (A3′)] $\operatorname{Im} \bar{s} > 0$ and [by (A1)] there exists one (or several) t channel trajectories $s = s_i(t)$ passing through $(\bar{s},\bar{t})$. They all satisfy $\lim_{t \to t_0} s_i(t) = s_0$.

Assume for simplicity there exists just one such trajectory: $s = s_2(t)$, $\bar{s} = s_2(\bar{t})$. Now, let $t(\mathscr{P})$ be the image of $\mathscr{P}_r$ under $t(s)$ in V_R; we can define along $t(\mathscr{P})$ the inverse function $s = s_1(t)$ [by (ii) above]. We now show that $s_1(t)$ coincides with the analytic continuation to complex t of the trajectory function $s_2(t)$. This will establish that $t_1(s)$ coincides with the inverse function of that trajectory, as asserted by the lemma.

To prove this, we make a reasoning analogous to Lemma 1: we surround $\bar{s}$ by a circle c_1 of radius r_1, $c_1 \subset \overline{U}_{r_0}$, so that $A(s,\bar{t}) \neq 0$, for s on c_1. There exists then a semidisk V_1: $|t - \bar{t}| < r_1'$, $\operatorname{Im} t > 0$, $V_1 \subset V_R$, so that $A(s,t) \neq 0$ for $s \in c_1$, $t \in \overline{V}_1$ and for each $t \in \overline{V}_1$ there exists just one zero in d_1: $\{s \mid |s - s_0| < r_1\}$, given by the analytic continuation of the trajectory $s_2(t)$:

$$s_2(t) = \frac{1}{2\pi i} \oint_{|s - s_0| = r_1} s \, \frac{\partial A/\partial s}{A}(s,t)\, ds. \tag{4.8}$$

The set of points $(s_2(t),t)$, $t \in V_1$, given by (4.8) exhausts the zeros of $A(s,t)$ in $d_1 \otimes V_1$. But $A(s,t)$ vanishes on the continuum $(s_1(t),t)$, with $t \in t(\mathscr{P})$ and this intersects $d_1 \otimes V_1$. This means that $s_1(t) \equiv s_2(t)$ (see Fig. 3). If there are several trajectories passing through $(\bar{s},\bar{t})$ we can only conclude that $s_1(t)$ coincides with the continuation of one of them, but this is sufficient.

The function $s_2(t)$ maps V_1 onto a domain lying to the left of $\overline{\Gamma}_2$:$s = s_2(t)$. The semicircle $\mathscr{P}_r$ lies to the left of $\overline{\Gamma}_2$ if the latter is such that $s_2(t)$ moves away from s_0, as t increases. Thus $s = s_2(t)$ is defined on $[t_0, t_0 + a]$, $a > 0$ (i.e., $t \in (t_0, t_0 + a)$ and $\lim_{\substack{t \to t_0 \\ t > t_0}} s_2(t) = s_0$.

Consider now the domain U delimited by $\overline{\Gamma}_2$: $s = s_2(t)$, the semicircle $|s - s_0| = r$, $\operatorname{Im} s \geqslant 0$, and the line $\operatorname{Im} s = 0$, $s < s_0$. Since $\overline{\Gamma}_2$ has no multiple points, U is simply connected. If we show that $t_1(s)$ is holomorphic at any point of U, then its continuation along any path in U leads to the same value and the proof of the lemma is finished. To this end, define $t_1(s)$ at points of U by analytic continuation starting from $s < s_0$, $s \in C_s^{\,+}$, along semicircles $|s - s_0| = \epsilon < r$, for decreasing ϵ. Assume that, in this process, we find $\bar{s}_1 \in U$, so that $\lim_{s \to \bar{s}_1} \operatorname{Im} t_1(s) = 0$. As before, we conclude there must exist a trajectory $\overline{\Gamma}_3$: $s = s_3(t)$, passing through $\bar{s}_1, t > t_0$, so that, for $s \in \overline{\Gamma}_3$, $t_1(s) = t_3(s)$, the inverse function of $\overline{\Gamma}_3$. By assumption, $\overline{\Gamma}_3$ intersects $|s - s_0| = r$ in one point situated at the right of $\overline{\Gamma}_2$; consequently, $\overline{\Gamma}_3$ intersects $\overline{\Gamma}_2$ at a point P with coordinates $(\bar{r},\theta)$, $\bar{r} < r$. In the neighborhood of P, the function $t_1(s)$ must coincide with both $t_2(s)$ and $t_3(s)$, the inverse functions of $\overline{\Gamma}_2$ and $\overline{\Gamma}_3$. Then $\overline{\Gamma}_2 = \overline{\Gamma}_3$ and $t_1(s)$ can be defined in all of U. This ends the proof of Lemma 5.

We now turn to the behavior of the trajectories $t_i(s)$ as s or t or both tend to infinity. We assume:

(A5) (i) If, as $|s| \to \infty$ along some direction $e^{i\theta_0}$, $0 \leqslant \theta_0 < \pi$, $t_i(s) \to t_0$, $|t_0|$ finite, then, for some α with $\operatorname{Re} \alpha > 0$, the function $C(s) = s^\alpha (t_i(s) - t_0)$ tends to $C_0 \neq 0$ in all directions $0 \leqslant \theta < \pi$. The function $t_i(s)$ admits of an analytic continuation counterclockwise along any circle of sufficiently large radius in the s plane until $u_i(s) = 4m^2 - s - t_i(s)$ reaches $C_u^{\,-}$. (ii) The function $C(s)$ admits of an asymptotic expansion uniformly valid in $0 \leqslant \theta < \pi$: $C(s) = C_0 + C_1/s^\beta + O(1/s^{\beta + \epsilon})$, $\operatorname{Re} \beta$, $\epsilon > 0$, $C_1 \neq 0$; the same is true for the derivative of $C(s)$, up to $O(1/s^{\beta + \epsilon + 1})$; (iii) if, as $|s| \to \infty$ along $e^{i\theta}$, both $|t_i(s)|, |u_i(s)| \to \infty$, there exists an angular interval $\phi_1 < \theta < \phi_2$, $\theta = \arg s$, where $t_i(s)$ can be analytically continued, on sufficiently large circles, until $s, t_i(s)$ or $u_i(s)$ reach

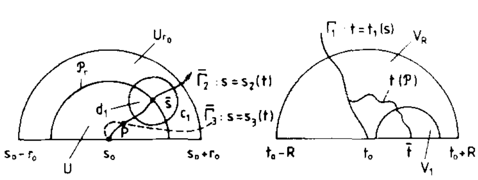


FIG. 3. To the proof of Lemma 5.

$C_s, C_t,$ or C_u. The asymptotic expansion $t_i(s) = s^\delta(C_0 + C_1/s^\delta + O(1/s^{\delta+\epsilon}))$ holds uniformly in this angular interval, Re $\delta, \epsilon > 0$, $C_1 \neq 0$ and the same is true for the derivative dt_i/ds; (iv) properties (i)–(iii) are valid if s,t,u are interchanged in any order.

One may verify that (A5) (i) is true if the amplitude and its derivative $\partial A/\partial t$ obey for large $\nu = s - u$ a Regge formula, uniformly in all directions of the complex ν plane, in a neighborhood of t_0:

$$A(\nu, t) = \gamma(t)\nu^{\alpha(t)} + \delta(t)\nu^{\beta(t)} + O(\nu^{\beta(t)-k}) \tag{4.9}$$

with Re $\alpha(t) > 0$, Re $\beta(t) < 0$, $\alpha(t)$, $\beta(t)$, $\gamma(t)$, $\delta(t)$ holomorphic in $|t - t_0| < r$, Re $(\beta(t) - k) <$ Re $\beta(t_0)$ for $|t - t_0| < r, k > 0$. A zero $t(\nu)$ of $A(\nu,t)$ for large ν, $t(\nu) \to t_0$ as $|\nu| \to \infty$ can occur only if $\gamma(t_0) = 0$. The validity of (A5) (i) follows then by evaluating

$$t(\nu) = \frac{1}{2\pi i}\oint_{|t-t_0| = r_1 < r} t\,\frac{\partial A/\partial t}{A}(\nu,t)\,dt. \tag{4.10}$$

We next make an assumption similar to (4.9) about the amplitude itself.

(A6) For each fixed $s, s \in D_s \cup C_s^+ \cup C_s^-$,

(i) $|A(s,t)| < C|t|^{N_1(s)}$, with $N_1(s)$ a function that is finite for all finite s; (ii) for t on C_t or C_u, the phase $\phi_s(t)$ of $A(s,t)$ can be defined by continuity along C_t or C_u, except for a finite number of points. The discontinuities at these points can be obtained by continuation of $A(s,t)$ along any sufficiently small semicircles around these points. The phase so defined obeys $|\phi_s(t)| < C_1(s)$, $C_1(s)$ finite for finite s. An analogous statement is true for each t and u in the corresponding sets.

The functions $N_1(s)$, $C_1(s)$ are allowed to increase indefinitely, as s increases.

As in Sec. II, we consider now more closely the pattern of curves Γ_i: $t = t_i(s)$, $s \in C_s$, or Γ'_j: $t = t_j(u)$, $u \in C_u$. Without loss of generality, we assume as in Sec. II that trajectories $t =$ const, $u =$ const or $s =$ const, do not occur. We denote by $\overline{\Gamma}_i$, $\overline{\Gamma}'_j$ the sets of image points in the domain D_t (t plane$\setminus C_t$) of the functions $t_i(s)$, $s \in C_s$, and $t_j(u)$, $u \in C_u$. Along $\overline{\Gamma}_i$, $\overline{\Gamma}'_j$ one can define the inverse functions $s_i(t)$, $u_j(t)$ as we have just seen. They may be many-valued. As in Sec. II, the sets $\overline{\Gamma}_i$, $\overline{\Gamma}'_j$ may have multiplicity higher than unity [i.e., $t = t_i(s)$ may be a multiple zero of $A(s,t) = 0$ or two curves Γ_i: $t = t_i(s)$, Γ_j: $t = t_j(s)$ may have the same image set, although different domains of definition and possibly different orientations]. As we have seen in Sec. II, multiplicities higher than unity cause strictly inessential complications, and we assume from now on that the $\overline{\Gamma}_i$'s or the $\overline{\Gamma}'_j$'s used in our argument are simple curves, with single-valued inverse, so that we shall not need to distinguish between Γ_i and the image set $\overline{\Gamma}_i$.

Assume now that a trajectory $t = t_1(s)$ "enters" the physical sheet of the t plane at $s_0 \in C_s^+$. As s increases along C_s^+, the value of the function $t_1(s)$ may either stay inside Q_s, or return to the cut C_t or reach a point $\bar{u}_0 \geqslant 4m^2$ on the u cut C_u, for some $\bar{s}_0 > s_0$. In this last case, the image $t = t_1(s)$ in D_t approaches a point $\bar{t}_0 = 4m^2 - \bar{s}_0 - \bar{u}_0 \leqslant -4m^2$ on the real axis of the t plane. We cannot follow the trajectory $t = t_1(s)$ for $s > \bar{s}_0$ in general and thus the curve Γ_1: $t = t_1(s)$ appears to have in D_t an "open end" at t_0. Loosely speaking, we would like it to "return" somehow to C_t, so that we can part D_t in subdomains $\mathscr{D}_i$ as we did with the disk $|\tau| < 1$ in Sec. II and thus repeat the arguments we used there.

We can "move away" from $\bar{t}_0$ by applying Lemma 5 to the function[28] $u_1(s) = 4m^2 - s - t_1(s)$ and conclude that it can be continued around $\bar{s}_0$ in the s plane until one reaches either Im $s = 0$, $s > \bar{s}_0$, or a curve $\hat{\Gamma}'_2$: $s = s_2(u)$, $u \in C_u$, $\lim_{u \to \bar{u}_0} s_2(u) = \bar{s}_0$, where it will coincide with the inverse function $u = \bar{u}_2(s)$. In the former case, the curve Γ_1 in the t plane is continued by $t = t_1(s) = 4m^2 - s - u_1(s)$ for $s > \bar{s}_0$. In the second case, define the function $t_2(u) = 4m^2 - u - s_2(u)$ for $u \in C_u$, for which $\lim_{u \to \bar{u}_0} t_2(u) = \bar{t}_0$. Then, the curve Γ'_2: $t = t_2(u)$ has in the t plane an end point precisely, where Γ_1: $t = t_1(s)$ has its end. Moreover, its orientation is (Lemma 5) such that $s_2(u)$ moves away from $\bar{s}_0$ if u moves in its natural sense (increasing on C_u^+, decreasing on C_u^-). Thus t moves away from $\bar{t}_0$ on Γ'_2 and Γ'_2 can be regarded as a "continuation" of Γ_1. This can be made more precise.

Lemma 6: In case the analytic continuation of $u_1(s)$ stops at $\hat{\Gamma}'_2$: $s = s_2(u)$ the inverse function $s_1(t)$ defined on Γ_1 can be analytically continued around $\bar{t}_0$ to the left of Γ_1, for $|t - \bar{t}_0|$ sufficiently small, until one reaches Γ'_2; there $s_1(t)$ coincides with the function $\bar{s}_2(t) = 4m^2 - t - u_2(t)$, with $u_2(t)$ the inverse function of $t = t_2(u)$.

To prove this statement, consider the equation

$$t = 4m^2 - s - \bar{u}_2(s), \tag{4.11}$$

where $\bar{u}_2(s)$ is the inverse function associated to $\hat{\Gamma}'_2$: $s = s_2(u)$. By Lemma 5, the right-hand side of (4.11) is holomorphic in the domain $\mathscr{U}$ bounded by Im $s = 0$, $s < \bar{s}_0$, the curve $\hat{\Gamma}'_2$: $s = s_2(u)$, $u \in C_u$, and an arc $|s - \bar{s}_0| = r$, Im $s > 0$. Near each s in $\mathscr{U}$, if $\tilde{t}$ is the value of the right-hand side, Eq. (4.11) has a unique solution $s = s(t)$, holomorphic in a neighborhood of $\tilde{t}$, since, by (A4) $d\bar{u}_2/ds \neq -1$, if r is small enough. For s on Im $s = 0$, $s < \bar{s}_0$, $\bar{u}_2(s) = u_1(s)$, and (4.11) reduces to $t = t_1(s)$, which defines the inverse function $s = s_1(t)$. Thus $s(t)$ is the analytic continuation of $s_1(t)$. Let us find the solution of (4.11) for s on $\hat{\Gamma}'_2$. The value of the right-hand side is $t = 4m^2 - s_2(u) - u = t_2(u)$, $u \in C_u$, i.e., t lies on Γ'_2. The equation $t = 4m^2 - s_2(u) - u$ defines the inverse function $u_2(t)$, $t \in \Gamma'_2$: $t \equiv 4m^2 - s_2(u_2(t)) - u_2(t)$. We claim that $\bar{s}_2(t) \equiv s_2(u_2(t))$ verifies (4.11) identically. Indeed, $4m^2 - s_2(u_2(t)) - \bar{u}_2(s_2(u_2(t))) = 4m^2 - s_2(u_2(t)) - u_2(t) \equiv t$. Thus, on $\hat{\Gamma}'_2$, $s(t) \equiv \bar{s}_2(t)$. Now, as s moves clockwise around s_0 in Im $s > 0$, starting at $s < s_0$, the right-hand side of (4.11) moves to the left of Γ_1. If $\mathscr{P}$ is the path described by it until s reaches Γ'_2, the continuation of $s_1(t)$ along $\mathscr{P}$ yields $\bar{s}_2(t)$.[29]

In Fig. 4, we show the pattern of the curves Γ_i in the t plane for the amplitude $A(s,t,u) = 2(4m^2 - s)^{1/2} + (4m^2 - t)^{1/2} + (4m^2 - u)^{1/2} + 2$, which we discuss in Appendix J.

There is another way in which a curve Γ_1: $t = t_1(s)$ might have an "open end" in the t plane: as $s \to \infty, s \in C_s^+$, $t_1(s)$ might tend to a finite (complex) limit t_0. Then, by (A5) (i), there exists a curve Γ'_2: $t = t_2(u)$, $u \in C_u^-$, with the property that $\lim_{u \to \infty} t_2(u) = t_0$. The angle between the tangents to the curves Γ_1, Γ'_2 at their common point is, in general,

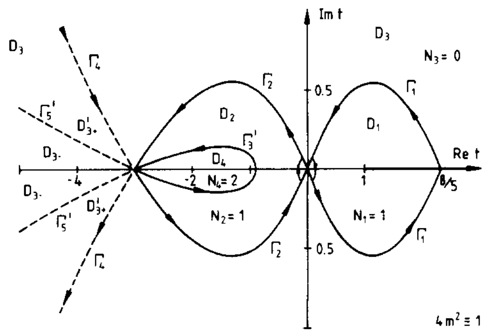


FIG. 4. The pattern of trajectories in the t plane for $A(s,t,u) = 2(4m^2 - s)^{1/2} + (4m^2 - t)^{1/2} + (4m^2 - u)^{1/2}$.

different from π. Indeed, by (A5) (ii), $\lim_{s\to\infty} s^\alpha(t_i(s) - t_0)$ is independent of the direction of s and thus, for t near t_0, $(t - t_0)|_{t\in\Gamma_2'} = \exp(-i\pi\alpha)(t - t_0)|_{t\in\Gamma_1}$.

The inverse function $s_1(t)$ is the solution of the equation

$$s^\alpha(t - t_0) = C(s) \tag{4.12}$$

for t on Γ_1. By (A5) (ii) the derivative dt/ds does not vanish if $|s|$ is large enough, and, consequently, Eq. (4.12) can be solved for $|t - t_0|$ sufficiently small. Equation (4.12) can be solved also for t on Γ_2' and for t on neither Γ_1 or Γ_2': this defines the analytic continuation of $s_1(t)$ around t_0, until we reach Γ_2'. We obtain there $\bar{s}_2(t) = 4m^2 - t - u_2(t)$, with $u_2(t)$ the inverse function of Γ_2'.

We are now finally in a position to imitate the treatment of Sec. II: we join several curves Γ_i, Γ_i' with each other and with pieces of the sets C_t^+, C_t^- by considering the analytic continuation (in the sense described in this section) along C_t, Γ_i, Γ_i' of a zero trajectory function $s(t)$ defined in the t channel. We have seen that if, at some t_0, $s(t)$ or $u(t) = 4m^2 - t - s(t)$ becomes real and larger than $4m^2$, either we can continue beyond t_0 or there exists a curve Γ_1: $t_1 = t_1(s)$ [or Γ_1': $t = t_1(u)$], so that the inverse function $s_1(t)$ defined on it is the analytic continuation of $s(t)$. Assume we continue now $s_1(t)$ along Γ_1: the continuation can be done until Γ_1 reaches C_t or $u_1(t)$ reaches C_u, or $|s_1(t)| \to \infty$ as $t \to t_0$, or $|s_1(t)| \to s_0$ (maybe infinite), as $t \to \infty$ on Γ_1. In this last case, by (A5) (i), (iii), and (iv) we can continue $s_1(t)$ along an arc of a circle of large enough radius in the t plane, until t reaches C_t or $u_1(t) = 4m^2 - t - s_1(t)$ reaches C_u, i.e., t reaches a curve Γ_i': $t = t_i(u)$. There $s_1(t)$ coincides with $4m^2 - t - u_i(t)$, with $u_i(t)$ the inverse function of Γ_i'. In all the former cases we can also continue $s_1(t)$ further. By (A3), we must return to the original function, after a finite number of steps. We call $\widetilde{\Gamma}_1$ the closed path made up of the set of curves Γ_i, Γ_i' and intervals (t_i, t_{i+1}) of C_t^+, C_t^- that are needed to return $s(t)$ to its original value. There can exist closed paths entirely contained in D_t (as in Fig. 4). As in Sec. II, a sense is attached to a path $\widetilde{\Gamma}_k$: that of increasing (decreasing) t for $t \in C_t^+$ (C_t^-) and that defined on each component Γ_i,Γ_i'. Every curve Γ_i, Γ_i' falls uniquely into one path $\widetilde{\Gamma}_k$. We call $\mathscr{L}_t$ the union of all image points of Γ_i, Γ_j'. We make now a last simplifying assumption.

(A7) A curve Γ_i does not intersect itself or another curve $\Gamma_j(\Gamma_j')$ an infinite number of times.

As a consequence of (A7), the complement of $\mathscr{L}_t$ consists of a finite number of domains $\mathscr{D}_i$ (some of them unbounded). All roots $s(t)$ of the equation $A(s,t) = 0$ are bounded if t is the interior of some domain $\mathscr{D}_k$. Indeed, if $|s(t)| \to \infty$ as $t \to t_0$, then we have seen that (A5) (i) allows us to find the inverse function $t(s)$, satisfying Eq. (4.11) as $|s| \to \infty$ for all directions $0 \leqslant \theta < \pi$ (or $\pi < \theta \leqslant 2\pi$ by real analyticity). We let then $s \to \infty$ along C_s^+; this generates Γ_i: $t = t_i(s)$ with $\lim_{s\to\infty} t_i(s) = t_0$, so that $t_0 \notin \operatorname{int} \mathscr{D}_k$.

Before stating the analog of the theorems of Sec. II, we introduce two notations.

(i) We consider the function

$$W(t) = \exp(-\theta(t)) = \exp(-(4m^2 - t)^{1/4}), \tag{4.13}$$

where the root is chosen with a cut along $t > 4m^2$ and positive real for $t < 4m^2$. The modulus of $W(t)$ decreases more quickly than any power in any direction of the complex t plane, including the cuts. Also, the integral

$$I(\Gamma_i) = \int_{\Gamma_i} |\exp(-\theta(t))||dt| \tag{4.14}$$

converges for every Γ_i. This is a consequence of (A1), (A2), (A4), (A5): $I(\Gamma_i)$ can be transformed into an integral along C_s or C_u; the Jacobian $|dt/ds|$ can diverge at most like $1/(s - s_0)^n$ $(n > 0)$ at finite s_0 or like s^k, $k > 0$ as $s \to \infty$. In both cases, (4.14) converges.

(ii) We divide C_t^+, C_t^- into intervals (t_j, t_{j+1}), $-\mathrm{N} \leqslant \mathrm{j} \leqslant \mathrm{N}$, $t_0 = 4m^2$, so that the number n_j of roots $s_l(t)$, $l = 1, \ldots, n_j$ of $A(s,t) = 0$, $s_l(t) \in Q_t$, $t \in C_t$, is constant for $t \in (t_j, t_{j+1})$. Negative (positive) indices of t_j refer to C_t^- (C_t^+).

We can now state:

Theorem 4.1: Define:

$$\mathscr{N}(t) = \frac{1}{2\pi i W(t)} \left(\sum_j \int_{\Gamma_j(\Gamma_j')} \frac{W(t')}{t' - t} dt' + \sum_j n_j \int_{t_j}^{t_{j+1}} \frac{W(t')}{t' - t} dt' \right) \tag{4.15}$$

with the sense of $\Gamma_j(\Gamma_j')$ used above and $W(t)$ of Eq. (4.13). Then:

(i) for $t \in \mathscr{D}_k$, $\mathscr{N}(t) = N_k$, an integer;

(ii) if the boundary of $\mathscr{D}_k$ contains $(t_j, t_{j+1}) \in C_t$, $N_k = n_j$;

(iii) let $N(t)$ be the number of roots of $A(s,t)$ in $\mathscr{M}$, at a given $t \in \mathscr{D}_k$; $N(t)$ is independent of t, for $t \in D_k$: $N(t) = \overline{N}_k$;

(iv) $N_k = \overline{N}_k$.

Proof: From the remark following Eq. (4.14), $\mathscr{N}(t)$ is well defined. If we group together all Γ_i, Γ_i' and intervals (t_j, t_{j+1}) that make up a closed path $\widetilde{\Gamma}_l$, we can write

$$\mathscr{N}(t) W(t) = \sum_l \frac{1}{2\pi i} \int_{\widetilde{\Gamma}_l} \frac{W(t')}{t' - t} dt'. \tag{4.16}$$

The integral over each $\widetilde{\Gamma}_l$ gives rise in every $\mathscr{D}_k$, by the residue theorem, to an integer times $W(t)$. Then (i) follows by addition over the $\widetilde{\Gamma}_l$.

It follows from the real analyticity of $A(s,t)$ that to each interval (t_j, t_{j+1}) on C_t^+, there corresponds an interval (t_{-j-1}, t_{-j}) on C_t^-, with $t_{-j-1} = t_{j+1}$, $t_{-j} = t_j$, and

$n_{-j} = n_j$. Applying the Plemelj discontinuity formulas across (t_j, t_{j+1}) to $\mathcal{N}(t)W(t)$, we obtain:

$$\mathcal{N}(t_+)\exp(-\theta(t_+)) - \mathcal{N}(t_-)\exp(-\theta(t_-)) = n_j \cdot 2i\,\mathrm{Im}\,(\exp(-\theta(t_+))), \tag{4.17}$$

where $\theta(t_+)$ is the value of $\theta(t)$ on C_t^+. Equation (4.17) is an identity for $t \in (t_j, t_{j+1})$ and can be satisfied for integer $\mathcal{N}(t_+)$, $\mathcal{N}(t_-)$ only if $\mathcal{N}(t_+) = \mathcal{N}(t_-) = n_j$. This proves (ii).

To prove (iii) and (iv), we consider, at fixed $\bar{t} \in C_t^+$, $\bar{t} \in (t_1,t_2)$ a contour $\mathscr{C}_L$ running along and a small distance away from the s and u cuts and closing by large pieces of circle of radius L in the complex s plane; L is chosen so that $\mathscr{C}_L$ encloses all n_1 zeros of $A(s,\bar{t})$. Now,

$$N(t) = \frac{1}{2\pi i}\oint_{\mathscr{C}_L} \frac{\partial A/\partial s}{A}(s,t)\,ds \tag{4.18}$$

is, for $t = \bar{t}$, independent of L (for L sufficiently large). Let $\mathscr{D}_1$ be the domain adjacent to (t_1,t_2). Since $A(s,\bar{t}) \neq 0$ for s on $\mathscr{C}_L$, we can find a small neighborhood $\mathscr{U}$ of $\bar{t}$, so that $A(s,t) \neq 0$ for $t \in \mathscr{U} \cap \bar{\mathscr{D}}_1$, $s \in \mathscr{C}_L$. Then, by (A1), we conclude that $N(t)$ is independent of t, for $t \in \mathscr{U} \cap \mathscr{D}_1$, and $N(t) = N(\bar{t}) = n_1$. Altering slightly the contour $\mathscr{C}_L$, we can verify with the same argument that $N(t)$ is actually constant in the neighborhood of any point of $\bar{\mathscr{U}} \cap \bar{\mathscr{D}}_1$ and, in fact, in a neighborhood (contained in $\mathscr{D}_1$) of any point of $\mathscr{D}_1 \cup (t_1,t_2)$. Thus, $N(t) = N_1 = n_1$ for $t \in \mathscr{D}_1$. Since $\mathcal{N}(t) = n_1$ for $t \in \mathscr{D}_1$, (iii) and (iv) are proved for domains adjacent to C_t (and not containing points t with $\mathrm{Im}\,t = 0$, $t \leqslant -4m^2$—see below).

Consider now a boundary Γ_1 between $\mathscr{D}_1$ and $\mathscr{D}_2$ and assume $\mathscr{D}_2$ lies to the left of Γ_1. Then, as t approaches Γ_1 from $\mathscr{D}_2$, one of the roots $s(t)$ of $A(s,t) = 0$ approaches $s_0 \in C_s$ [or $u_0 \in C_u$ if the boundary is $\Gamma_1' : t = t_1(u)$, $u \in C_u$]. By surrounding the remaining $N_1(\bar{t}_1)$ zeros, $\bar{t}_1 \in \Gamma_1$, by a contour running like $\mathscr{C}_L$ and avoiding a neighborhood of s_0, one concludes, as in Sec. II, that, if Γ_1 has unit multiplicity

$$\lim_{\substack{t \to \bar{t}_1 \\ t \in \mathscr{D}_2}} N(t) - \lim_{\substack{t \to \bar{t}_1 \\ t \in \mathscr{D}_1}} N(t) = 1. \tag{4.19}$$

This is precisely the discontinuity obtained from (4.15) by the Plemelj formulas. The fact that $N(t) = N(\bar{t}_1) + 1 = N_2 =$ constant for $t \in \mathscr{D}_2$ can be shown as for $t \in \mathscr{D}_1$, provided $\mathscr{D}_2$ does not contain a piece of the real axis $\mathrm{Im}\,t = 0$, $t < -4m^2$.

We are careful about this last point, since for t real, $t < -4m^2$, the contour $\mathscr{C}_L$ can no longer be a simple closed curve: the s and u cuts overlap. However, $\mathcal{N}(t)$ has no discontinuity across $\mathrm{Im}\,t = 0$, $t < -4m^2$, unless a curve Γ_i runs along it. Thus, we only have to convince ourselves that, for t_0 on $\mathrm{Im}\,t = 0$, $t_0 < -4m^2$, $t_0 \in \mathscr{D}_2$, $A(s,t)$ has indeed the same number of zeros as for $\mathrm{Im}\,t \neq 0$. To verify this (rather obvious) statement, it is enough to observe that, e.g., as $t \to t_0$, $\mathrm{Im}\,t \to 0$, $\mathrm{Im}\,t > 0$, no root $s(t)$ of $A(s,t) = 0$ can stay enclosed in the flat rectangle $(4m^2, -t_0, -t_0 - i\,\mathrm{Im}\,t, 4m^2 - i\,\mathrm{Im}\,t)$ contained between the overlapping cuts. Indeed, if this were the case, t_0 would belong to a Γ_i. Thus, for $\mathrm{Im}\,t$ small we can modify the contour $\mathscr{C}_L$ to two disjoint semicircles and their diameters below and above the cuts (see Fig. 5). For $N(t)$ defined via (4.18) through an integral over

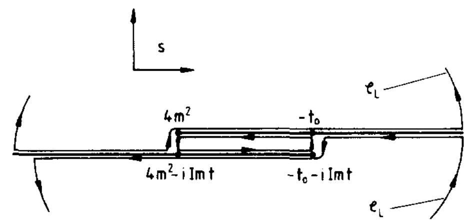


FIG. 5. The contour $\mathscr{C}_L$ for overlapping cuts.

these two contours, the continuity as $\mathrm{Im}\,t \to 0$ is obvious. Since the procedure can now be extended from $\mathscr{D}_2$ to any finite number of domains $\mathscr{D}_i$, the proof of Theorem 4.1 is finished.

Following Sec. II, we now consider the symmetric combinations of the roots $s_j^{(i)}(t)$, $u_j^{(i)}(t)$ of $A(s,t) = 0$, for t inside a domain $\mathscr{D}_i$, $j = 1,2,\ldots,N_i$. The sums

$$S_{i,k}(t) = \sum_{j=1}^{N_i} (s_j^{(i)}(t))^k, \quad U_{i,k}(t) = \sum_{j=1}^{N_i} (u_j^{(i)}(t))^k \tag{4.20}$$

are holomorphic and one-valued functions of t, for $t \in \mathscr{D}_i$. As t approaches t_0 on a boundary of Γ_i, one of the roots $s_j^{(i)}(t)$ may diverge, however, not worse than const$/(t - t_0)^p$, $p > 0$, according to (A5) (i). To each $\bar{t}_{0,j} \in D_t$, where one of the roots $s_l^{(n)}(t)$ diverges, we associate a factor $B(t,\bar{t}_{0,j})^{p_j}$, defined by

$$B(t,\bar{t}_{0,j}) = \frac{z - z_{0,j}}{1 - zz_{0,j}^*}, \quad z = \frac{1 - (1 - t/4m^2)^{1/2}}{1 + (1 - t/4m^2)^{1/2}}. \tag{4.21}$$

If $\bar{t}_{0,j} \in C_t$, we associate to it the same $B(t,t_{0,j})$, however, with t in the definition of z replaced by $t_1 = (1 - t/4m^2)^{1/2}$. In (4.21), $z_{0,j}$ are the images through $z(t)$ [or $z(t_1(t))$] of $\bar{t}_{0,j}$. According to (A3), there are only a finite number of points $t_{0,j}$. Further, if $|t| \to \infty$ in an unbounded domain $\mathscr{D}_i$, then by (A5) (iii) the functions $S_{i,k}(t)$, $U_{i,k}(t)$ are polynomially bounded. We conclude that the functions

$$\begin{aligned}\tilde{S}_{i,k}(t) &= \prod_j (B(t,t_{0,j})^{p_j})^k \exp(-\theta(t)) S_{i,k}(t) \\ &\equiv G_k(t) S_{i,k}(t)\end{aligned} \tag{4.22}$$

are one valued and holomorphic in $\mathscr{D}_i$, continuous in $\bar{\mathscr{D}}_i$ and vanish as $|t| \to \infty$ in an unbounded $\mathscr{D}_i$ [similar definition for $\tilde{U}_{i,k}(t)$]. We can now state

Theorem 4.2: Define

$$\Omega_s^{(k)}(t) = \frac{1}{2\pi i G_k(t)}\left(\sum_l \int_{\Gamma_l,\Gamma_l'} \frac{(s_l(t'))^k G_k(t')}{t' - t}\,dt + \sum_j \int_{t_j}^{t_{j+1}} \frac{\sum_{l=1}^{n_j} (s_{p_l}(t'))^k G_k(t')}{t' - t}\,dt'\right), \tag{4.23}$$

$$\Omega_u^{(k)}(t) = \frac{1}{2\pi i G_k(t)}\left(\sum_l \int_{\Gamma_l,\Gamma_l'} \frac{(u_l(t'))^k G_k(t')}{t' - t}\,dt' + \sum_j \int_{t_j}^{t_{j+1}} \frac{\sum_{l=1}^{n_j} (u_{p_l}(t'))^k G_k(t')}{t' - t}\,dt'\right). \tag{4.24}$$

Then, if $t \in \mathscr{D}_i$

$$\Omega_s^{(k)}(t) = S_{i,k}(t), \ \Omega_u^{(k)}(t) = U_{i,k}(t). \tag{4.25}$$

The integrations in (4.23) and (4.24) are performed in the same way as in (4.15); $s_{p_1}(t'), \ldots, s_{p_{n_j}}(t')$ are the n_j roots of $A(s,t) = 0$ in Q_t for $t \in (t_j, t_{j+1})$.

Proof: In each $\mathscr{D}_i$ one can obtain $S_{i,k}(t)$ by means of a Cauchy integral along the boundary $\partial\mathscr{D}_i$

$$\frac{1}{2\pi i G_k(t)} \oint_{\partial\mathscr{D}_i} \frac{S_{i,k}(t')G_k(t')}{t'-t}\,dt' = \begin{cases} S_{i,k}(t), & t \in \mathscr{D}_i \\ 0, & t \in \complement\, \mathscr{D}_i \end{cases} \tag{4.26}$$

since $\widetilde{S}_{i,k}(t)$ is continuous in $\overline{\mathscr{D}}_i$. As in Sec. II, one shows that, if $\mathscr{D}_i$ lies to the left of Γ_i (assumed simple)

$$\widetilde{S}_{i,k}(t) - \widetilde{S}_{i-1,k}(t) = (s_i(t))^k G_k(t). \tag{4.27}$$

Equation (4.25) follows then by addition of Eqs. (4.26) using the constraints (4.27) and verifying that the sense of the integrals is correct. The same can be done for $U_{i,k}(t)$. This ends the proof.

The analog of Eq. (2.12) is

$$\lim_{t \to t_0 \in C_t^{+(-)}} \Omega_s^{(k)}(t) = S_{i,k}(t_0 \pm i\epsilon),$$
$$\tag{4.28}$$
$$\lim_{t \to t_0 \in C_t^{+(-)}} \Omega_s^{(k)}(t) = U_{i,k}(t_0 \pm i\epsilon).$$

The values on the left-hand sides of Eq. (4.28) can be obtained directly from experiment up to a discrete ambiguity. In general, an arbitrary resolution of the discrete ambiguity in the s,t,u channels will lead to a violation of Eqs. (4.28). These latter represent conditions to be satisfied by the zero trajectories in order that the construction of amplitudes with Mandelstam analyticity be possible.

Theorem 4.2 achieves in fact the construction of the Cousin data in the Mandelstam domain: from knowledge of the zero trajectories on the boundary of $\mathscr{M}$, we have obtained the values of the symmetric sums (4.20) in each domain $\mathscr{D}_i$. We can now construct two functions $\mathscr{F}_s(s,t)$, $\mathscr{F}_u(u,t)$ by means of Eqs. (2.17) and (2.18) of Sec. II with the appropriate replacements and $N = \max \mathscr{N}(t)$. In each domain $\mathscr{D}_i$, these functions equal a polynomial in s or u of degree N_i, with coefficients that are holomorphic in t. These polynomials (one of them suffices) describe completely the zeros of $A(s,t)$ in the cut complex s plane Q_t at each t in $\mathscr{D}_i$. For reference:

$$\mathscr{F}_s(s,t) = \frac{\sum_{p=0}^{N} \alpha_p(t) s^{N-p}}{s^{N-\mathscr{N}(t)}} \tag{4.29}$$

with $\alpha_p(t)$ defined as a function of $\Omega_s^{(k)}(t)$ by (2.18).

V. THE MANDELSTAM DOMAIN. CONSTRUCTION OF THE AMPLITUDE

In the first part of this section, we construct the Cousin function $C(s,t)$ which vanishes in $\mathscr{M}$ at those points where the amplitude vanishes and nowhere else and, in contrast to $\mathscr{F}_s(s,t)$, $\mathscr{F}_u(u,t)$ of the previous section, is holomorphic in all of $\mathscr{M}$. In the second part, we construct the amplitude $A(s,t)$, similarly to Sec. III.

We define first:

$$K_0(s,t) = \exp\left(- \sum_i \frac{1}{2\pi i w(t)} \int_{\Gamma_i} \frac{\ln(s - s_i(t'))}{t'-t} w(t')dt' \right)$$
$$\times \exp\left(- \sum_j \frac{1}{2\pi i w(t)} \int_{\Gamma'_j} \frac{\ln(u_j(t') - u)}{t'-t} w(t')dt' \right)$$
$$\equiv K_{0s}(s,t)K_{0u}(s,t) \tag{5.1}$$

with (a possible choice)

$$w(t) = W(t) = \exp(-(4m^2 - t)^{1/4}). \tag{5.2}$$

In (5.1), the curves Γ_i are obtained from the s channel zero trajectories $\Gamma_i: t = t_i(s)$, $s \in C_s^+$, $s \in C_s^-$, whereas $\Gamma'_j: t = t_j(u)$, $u \in C_u^+$ are obtained from the u channel. The sense of integration is the one of the previous section (i.e., s increasing, $s \in C_s^+$; decreasing, $s \in C_s^-$); the weight $w(t)$ guarantees convergence of the integrals.

The determination of the logarithm is chosen as follows: (i) for a curve $\Gamma_i: t = t_i(s')$, $s' \in C_s^+$, we take $\ln(s - s_i(t'))$ positive imaginary for $s < 4m^2$; we can continue in s throughout D_s, so that $\ln(s - s_i(t'))$ has a right-hand cut for $s > s_i(t')$. As a consequence, the integral

$$I_i(s,t) = \frac{1}{2\pi i} \int_{\Gamma_i} \frac{\ln(s - s_i(t'))}{t'-t} \frac{w(t')}{w(t)}\,dt' \tag{5.3}$$

has, at fixed t, a cut for $s > \min s_i(t'), t' \in \Gamma_i$; (ii) by the real analyticity of $A(s,t)$, to each $\Gamma_i: t = t_i(s')$, $s' \in C_s^+$, there corresponds a curve $\Gamma_j \equiv \Gamma_i^*: t = t_j(s')$, $s' \in C_s^-$, $t_j(s') = t_i^*(s'^*)$, $s' \in C_s^-$; for $t' \in \Gamma_i^*$, we choose $\ln(s - s_j(t'))$ to be negative imaginary for $s < 4m^2$. We continue this choice in D_s so that $\ln(s - s_j(t'))$ has a right-hand cut in s, for $s > s_j(t')$. As a consequence, the same expression (5.3) taken along Γ_i^*, has, at fixed t, a cut for $s > \min s_j(t')$, $t' \in \Gamma_i^*$ and is equal at real s, t to the complex conjugate of (5.3). A minus sign comes from the opposite sense of integration on Γ_i^*. (iii) For a curve $\Gamma'_j: t = t_j(u')$, $u' \in C_u^+$, we take $\ln(u_j(t') - u)$ to be real for $u < 4m^2$. The continuation in D_u is made so that $\ln(u_j(t') - u)$ has a right-hand cut in u, for $u > u_j(t')$. (iv) For $t' \in \Gamma_j'^*$, defined as above in the s channel, $\Gamma_j'^* \equiv \Gamma'_k: t = t_k(u)$, $u \in C_u^-$, $t_k(u) = t_j^*(u^*)$, we choose $\ln(u_k(t') - u)$ again as real for $u < 4m^2$, and continue it with a right-hand cut in D_u. As a consequence, the integrals

$$I'_i(s,t) = \frac{1}{2\pi i} \int_{\Gamma'_i} \frac{\ln(u_i(t') - u)}{t'-t} \frac{w(t')}{w(t)}\,dt' \tag{5.4}$$

have a right-hand cut for $u > \min u_i(t')$, at fixed t, and the same holds for the integral over $\Gamma_i'^*$. As for the curves Γ_i, the integral over Γ'_i is for real t,u minus the complex conjugate of the integral over $\Gamma_i'^*$.

By this definition of the logarithm, we ensured that $K_0(s,t)$ is real for s,t,u real. However, there may appear end point singularities of $I_i(s,t)$, $I'_i(s,t)$, Eqs. (5.3) and (5.4) leading to zeros or poles of $K_0(s,t)$ as a consequence of these definitions: (i) If the continuation of a curve Γ_i is the curve Γ_i^* [as is possible if, e.g., a zero $t = t(s)$ of $A(s,t)$ becomes real as $s \to 4m^2$, $\mathrm{Im}\, s = -\epsilon < 0$ and is then continued analytically to $\operatorname{Im} s = \epsilon > 0$], then there is a discontinuity in the argument of $\ln(s - s_i(t'))$ in (5.3) as $t' \to t_0$, t_0 real, $s_i(t_0) = 4m^2$: for $s < 4m^2$, real, $\ln(s - 4m^2) = i\pi$, if we take the limit from Γ_i and $\ln(s - 4m^2) = -i\pi$, if we take the limit from Γ_i^*. One can verify this leads to a zero of $K_0(s,t)$ at $t = t_0$ for all s. No such discontinuity occurs when Γ_i' is continued by $\Gamma_i'^*$. (ii) Assume now a curve Γ_i terminates at a point t_0 on $\operatorname{Im} t = 0$, $t < -4m^2$. It is then "continued", in the sense of Lemma 6 by a curve $\Gamma_j'(\Gamma_j'^*)$ and, by real analyticity there must also exist a curve Γ_i^* "beginning" at t_0. This situation leads to a zero of $K_{0s}(s,t)$ at t_0 for all s; however, the sense of integration cannot be *a priori* known so that, in general, a pole might also appear at t_0. The curves Γ_j' with an end at t_0 lead in a similar way to a zero or pole of $K_{0u}(u,t)$ for all u. In $K_0(s,t)$ these singularities may cancel, so that we can only state that, in general, $K_0(s,t) \sim (t - t_{0i})^{2k_i}, k_i = -1,0,1$ at these points. (iii) Let $\bar t_{0j}$ be those end points of curves $\Gamma_i: t = t_i(s)$, $t_i(s) \to \bar t_{0j}$ as $s \to \infty$, $s \in C_s^+$ or $s \in C_s^-$. By (A5) (i) (and real analyticity) they are also end points of a curve Γ_j', $u \in C_u^-$, or $u \in C_u^+$. We show in Appendix H [part (ii)] that $K_0(s,t) \sim (t - t_{0j})^{-2}$ at these points.

We define thus:

$$K(s,t) = \prod_i (t - t_{0,i})^{p_i} \prod_j (t - t_{0j})^2 K_0(s,t). \tag{5.5}$$

In (5.5), $p_i = -1$ for points of the category (i), $p_i = -2k_i$ for points of category (ii).

Theorem 5.1: Define

$$C(s,t) = \mathscr{F}_s(s,t) K(s,t). \tag{5.6}$$

Then: (i) $C(s,t)$ is holomorphic in $\mathscr{M}$; (ii) is real analytic; (iii) for $s \in C_s$, it is holomorphic as a function of t in the corresponding cut t plane; (iv) property (iii) holds for $u \in C_u$; (v) it is continuous at those $(s,t,u) \in \bar{\mathscr{M}}$ with $s \in C_s$ or $u \in C_u$, and the other coordinate in Q_s or Q_u; (vi) $A(s,t)/C(s,t)$ is nonvanishing in $\mathscr{M}$; for $s \in C_s$ or $u \in C_u$, it is nonvanishing in the corresponding cosine plane Q_s or Q_u.

Proof: At each fixed $t \notin \Gamma_i$, $K_{0s}(s,t)$ is holomorphic in the neighborhood of any point $s \in D_s$. Since D_s is simply connected, $K_{0s}(s,t)$ is holomorphic and one valued there. Similarly, $K_{0u}(u,t)$ is holomorphic and one valued in D_u. We conclude that $K_0(s,t)$ is holomorphic and one valued in $D_s \cap D_u$ at each $t \notin \Gamma_i$.

We can differentiate (5.1) with respect to t at each $t \notin \Gamma_i$ and thus $K_0(s,t)$ is in fact holomorphic in $\mathscr{D}_i \otimes D_s \cap \mathscr{D}_i \otimes D_u$, with respect to both s and t. (See Ref. 23.) The function $\mathscr{F}_s(s,t)$ is manifestly holomorphic in s and t in the domain above and thus, so is $C(s,t)$, Eq. (5.6).

At fixed s, $K_0(s,t)$ has discontinuities along Γ_i and Γ_i'. The poles and zeros in t have been removed by the factors in (5.5). There exists apart from this a cut for $t = -s - x$, $\operatorname{Im} x = 0$, $x > 0$, coming from the logarithm in the exponent of $K_{0u}(u,t)$. The function $\mathscr{F}_s(s,t)$ has discontinuities along Γ_i, Γ_i', and a cut for $t \geqslant 4m^2$, present in $\Omega_s^{(k)}(s,t)$, Eq. (4.15). To establish holomorphy of $C(s,t)$ in $\mathscr{M}$, we show now that the discontinuities across Γ_i of $\mathscr{F}_s(s,t)$ and $K(s,t)$ remove each other in their product.

At points interior to D_s, there is no difficulty in applying Plemelj's formulas to $C(s,t)$ since the integrands are Hölder continuous. If $\mathscr{D}_{i+1}$ lies to the left and $\mathscr{D}_i$ to the right of Γ_i, which is assumed simple, and if t is not a cusp, we repeat the argument of Sec. III to obtain

$$\begin{aligned}
\lim_{\substack{t \to \Gamma_i \\ t \in \mathscr{D}_{i+1}}} \mathscr{F}_s(s,t) K(s,t) &= F_{i+1}(s,t) \bar K(s,t) (s - s_i(t))^{-1/2} \\
&\quad \times \exp\left(-P \frac{1}{2\pi i} \int_{\Gamma_i} \frac{\ln(s - s_i(t'))}{t' - t} \frac{w(t')}{w(t)} dt' \right) \\
&= F_i(s,t) \bar K(s,t) (s - s_i(t))^{1/2} \\
&\quad \times \exp\left(-P \frac{1}{2\pi i} \int_{\Gamma_i} \frac{\ln(s - s_i(t'))}{t' - t} \frac{w(t')}{w(t)} dt' \right) \\
&= \lim_{\substack{t \to \Gamma_i \\ t \in \mathscr{D}_i}} \mathscr{F}_s(s,t) K(s,t),
\end{aligned} \tag{5.7}$$

where we have used the fact that $F_{i+1}(s,t) = (s - s_i(t)) \times F_i(s,t)$, $t \in \Gamma_i$, and $F_k(s,t)$ is the determination of $\mathscr{F}_s(s,t)$ in $\mathscr{D}_k$; we have written in (5.7) $K(s,t) = \bar K(s,t) k_i(s,t)$ with $k_i(s,t)$ the factor of K_{0s} corresponding to Γ_i. If t is a cusp of Γ_i, application of (E1) shows that $C(s,t)$ is holomorphic there. Thus, $C(s,t)$ is holomorphic in t at all points of Γ_i, except possibly the end points. Since at fixed t on Γ_i, Eq. (5.7) is holomorphic and one valued for s in D_s, we have succeeded in enlarging the analyticity domain of $C(s,t)$ to $(\mathscr{D}_i \cup \mathscr{D}_{i+1}) \otimes D_s \cap (\mathscr{D}_i \cup \mathscr{D}_{i+1}) \otimes D_u$.

Consider now a curve Γ_j': only $K_{0u}(u,t)$ has a discontinuity there; however, we have to take care that $\mathscr{F}_s(s,t)$ is a polynomial in s rather than u. We use then the fact that, for t on Γ_j', $u_j(t) - u = s - \bar s_j(t)$ [$\bar s_j(t)$ is defined in Lemma 6] and obtain

$$\begin{aligned}
\lim_{\substack{t \to \Gamma_j' \\ t \in \mathscr{D}_{j+1}}} \mathscr{F}_s(s,t) K(s,t)
&= F_{j+1}(s,t) \bar K(s,t) \psi(s,t) / (u_j(t) - u)^{1/2} \\
&= F_{j+1}(s,t) \bar K(s,t) \psi(s,t) / (s - \bar s_j(t))^{1/2} \\
&= F_j(s,t) (u_j(t) - u)^{1/2} \bar K(s,t) \psi(s,t) \\
&= \lim_{\substack{t \to \Gamma_j' \\ t \in \mathscr{D}_j}} \mathscr{F}_s(s,t) K(s,t)
\end{aligned} \tag{5.8}$$

with

$$\psi(s,t) = \exp\left(-P \frac{1}{2\pi i} \int_{\Gamma_j'} \frac{\ln(s + t - \bar s_j(t') - t')}{t' - t} dt' \right). \tag{5.9}$$

The reason why we chose opposite signs for s and u in the logarithms appearing in K_{0s} and K_{0u} is now apparent: the coefficient of the highest power of s in each of the determinations of $\mathscr{F}_s(s,t)$ is unity and we would have obtained a minus sign on the right-hand side of (5.8) had we not made this choice. We have achieved thus analyticity of $C(s,t)$ in $(\mathscr{D}_j \cup \mathscr{D}_{j+1}) \otimes D_s \cap (\mathscr{D}_j \cup \mathscr{D}_{j+1}) \otimes D_u$, and we can clearly obtain by this method analyticity in $D_t \otimes D_s \cap D_t \otimes D_u$, except for isolated planes $t = \text{const}$. These correspond to intersections of several Γ_i's or end points of Γ_i, Γ_i' (with the definitions of the previous section, such end points can lie in D_t).

The case of the intersection of several trajectories is treated precisely as in Sec. II (possibly with the care indicated in Appendix E for cusps). We have still to discuss (i) those points $\bar{t}$ on Im $t=0$, $t<-4m^2$ at which a curve $\Gamma_i: t=t_i(s)$ is "continued" (in the sense of Lemma 6) by $\Gamma'_j: t=t_j(u)$; (ii) those points $\bar{t}_{0,j}\in D_t$, with the property that $\lim_{s\to\infty} t_i(s)=\bar{t}_{0,j}$.

We show in Appendix H that $C(s,t)$ is holomorphic in s and t even in the neighborhood of such points, and thus $C(s,t)$ is holomorphic in all of $\mathscr{M}$.

Statement (ii) of the theorem is obvious, since the branches of the logarithm in $K_0(s,t)$ have been chosen so as to make it real for s,t real and $\Omega_s^{(k)}(t)$ is real analytic by its property (4.25).

Take now $s=s_0$, $s_0\in C_s$. The argument of the logarithm in $K_{0s}(s,t)$ in (5.1) has to be defined with care, since we distinguish[30] between curves $\Gamma_i: t=t_i(s')$, $s'\in C_s^+$, $\Gamma_j^{(*)}: t=t_j(s')$, $s'\in C_s^-$: If $s_0\in C_s^+$, $t'\in\Gamma_i$: $\arg(s_0-s_i(t'))=0$ if $s_0>s_i(t')$, $\arg(s_0-s_i(t'))=\pi$ if $s_0<s_i(t')$; if $t'\in\Gamma_j^{(*)}$, then $\arg(s_0-s_j(t'))=-2\pi$, if $s_0>s_j(t')$, but $=-\pi$ if $s_0<s_j(t')$. If $s_0\in C_s^-$, $t'\in\Gamma_i$: $\arg(s_0-s_i(t'))=2\pi$ if $s_0>s_i(t')$, $\arg(s_0-s_i(t'))=\pi$ if $s_0<s_i(t')$; if $t'\in\Gamma_j^{(*)}$, $\arg(s_0-s_j(t'))=0$, if $s_0>s_j(t')$, and $\arg(s_0-s_j(t'))'=-\pi$ if $s_0<s_j(t')$. With this, the functions $C(s_0,t)$ are well defined and even analytic at all points t, with $t\neq t_i(s_0)$ for some i. The behavior at $t_0=t_i(s_0)$ is explored as in Appendix E: if $t_i(s_0)$ is a regular point of Γ_i, we can repeat the proof without change. If $t_0=t_i(s_0)$ is a cusp, we can still do the same, provided we show that $\lim_{t\to t_0}(si_{in}(t)-s_0)/(t-t_0)_n^p, p>0$ is the same from all directions as $t\to t_0$. This follows from (A2) and the continuity of $d\xi/dt$ at the points of Γ_i, with the notation of (E9).

As for (iv), we define first the logarithm for u in C_u (we do not distinguish between $\Gamma'_i, \Gamma_i'^{(*)}$): if $u\in C_u^+$, $\arg(u_i-u)=0$ if $u<u_i(t')$ and $\arg(u_i-u)=-\pi$ if $u>u_i(t')$; if $u\in C_u^-$, $u_i(t')\in\Gamma_i$, $\arg(u_i-u)=0$ if $u<u_i(t')$ and $\arg(u_i-u)=\pi$ if $u>u_i(t')$. The property in (iv) is then proved precisely as for $s\in C_s$, noting that $s_i(t)-s=u-u_i(t)$ for t on Γ_i.

The proof of (v) follows Appendix E in detail and the proof of (vi) is the same as that of the analogous statement in Theorem 3.1: $K(s,t)$ is constructed so as to be nonvanishing in $\mathscr{M}$. This concludes the proof of Theorem 5.1.

We can now construct

$$E(s,t)=A(s,t)/C(s,t), \tag{5.10}$$

which is holomorphic and free of zeros in $\mathscr{M}$ and also in the cosine planes corresponding to s or u on the cuts. Moreover, its modulus is known in the "extended physical region" $\mathscr{P}_s\cup\mathscr{P}_t\cup\mathscr{P}_u$. At each fixed s,t or u on the cut, its extension to the corresponding cosine plane is also assumed to be known. We can compute

$$L(s,t)=\ln(\pm E(s,t)). \tag{5.11}$$

For a function $f(s,t)$ continuous and nonvanishing in $\mathscr{M}$, we define $\ln f(s,t)$ along each ray $s=4m^2/3+\lambda(s_1-4m^2/3)$, $t=4m^2/3+\lambda(t_1-4m^2/3)$, $0<\lambda<\infty$, passing through $s=t=u=4m^2/3$ and (s_1,t_1,u_1) by continuity from the value $\ln f(4m^2/3, 4m^2/3)$; one can verify that $\mathscr{M}$ is such that, if $(s_1,t_1,u_1)\in\mathscr{M}$, then the ray passing through (s_1,t_1,u_1) is entirely contained in $\mathscr{M}$. With this definition $\ln f(s,t)$ is continuous in $\mathscr{M}$ and, if $f(s,t)$ is holomorphic in $\mathscr{M}$, so is $\ln f(s,t)$ [which can be written as $\int_{\text{ray}}^{(s_1,t_1)} f'(s(\lambda),t(\lambda))/f(s(\lambda),t(\lambda))d\lambda$.] We can choose the sign of $E(s,t)$ in (5.11) so that $L(s,t)$ is real for s,t,u real.

Then, at fixed $s\in C_s^+$,

$$L(s,t)+L^*(s,t^*)=\ln(E(s,t)E^*(s,t^*))+2k\pi i=h_1(s,t). \tag{5.12}$$

With the above definition of the logarithm, $\ln(E(s,t)E^*(s,t^*))$ can be chosen real for t real, $s\in C_s^+$ and thus $k=0$. Clearly, at fixed $u\in C_u^+$, it is true that

$$L(u,t)+L^*(u,t^*)=\ln(E(u,t)E^*(u,t^*))=h_2(u,t). \tag{5.13}$$

By real analyticity, $L^*(s,t^*)=L(s^*,t)$, etc. We define

$$l(s,t)=\frac{L(s,t)}{(4m^2-s)^{1/2}(4m^2-u)^{1/2}} \tag{5.14}$$

with the root negative imaginary above the cut $s\geqslant 4m^2$ ($u\geqslant 4m^2$). We can then write (5.12) and (5.13) as

$$l(s,t)-l(s^*,t)=\frac{ih_1(s,t)}{(s-4m^2)^{1/2}(4m^2-u)^{1/2}},\quad s\in C_s^+, \tag{5.15}$$

$$l(u,t)-l(u^*,t)=\frac{ih_2(u,t)}{(4m^2-s)^{1/2}(u-4m^2)^{1/2}},\quad u\in C_u^+. \tag{5.16}$$

We show below that for $t\in\mathscr{D}_i$, $l(s,t)$ satisfies an unsubtracted dispersion relation as a consequence of assumption (A6). If this is granted, we can write, using (5.15) and (5.16)

$$l(s,t)=\frac{1}{2\pi}\int_{4m^2}^{\infty}\frac{h_1(s',t)ds'}{(s'-4m^2)^{1/2}(4m^2-u')^{1/2}(s'-s)}+\frac{1}{2\pi}\int_{4m^2}^{\infty}\frac{h_2(u',t)du'}{(u'-4m^2)^{1/2}(4m^2-s')^{1/2}(u'-u)}. \tag{5.17}$$

This defines $E(s,t)$ completely, and the construction of $A(s,t)$ is finished:

$$A(s,t)=\pm C(s,t)E(s,t). \tag{5.18}$$

We must still show that $l(s,t)$ satisfies the dispersion relation (5.17). To this end, we notice first that, at each fixed t, as a consequence of (A6), the amplitude $A(s,t)$ can be written as

$$A(s,t)=\Omega_t(s)\Omega_t(u)P(s,t) \tag{5.19}$$

with $P(s,t)$ a polynomial in s, and [$\phi_t(s)$ is the phase of $A(s,t)$, $s\in C_s$]

$$\Omega_t(s)=\exp\left[\frac{s}{\pi}\int_{4m^2}^{\infty}\frac{\phi_t(s')}{s'(s'-s)}ds'\right]. \tag{5.20}$$

The argument that the representation (5.19) is valid under assumption (A6) is standard; with more general assumptions than here it can be found in, e.g., Ref. 31. Further, $C(s,t)$ can be written as $\mathscr{F}_s(s,t)K(s,t)$ and $\mathscr{F}_s(s,t)$ reduces for $t\in\mathscr{D}_i$ to a polynomial in s, $F_i(s,t)$. By construction, $F_i(s,t)$ and $P(s,t)$

have the same roots at fixed t, so that we can write

$$E(s,t) = c_0(t)\Omega_t(s)\Omega_t(u)/K(s,t) \tag{5.21}$$

with $c_0(t)$ a constant at each fixed t. None of the factors on the right-hand side vanishes for $s \in Q_t$ and our purposes will be attained if we show that, for $f(s,t)$ equal in turn to $\Omega_t(s)$, $\Omega_t(u)$, $K(s,t)$,

$$l_f(s,t) = \frac{\ln f(s,t)}{(4m^2 - s)^{1/2}(4m^2 - u)^{1/2}} \tag{5.22}$$

satisfies a dispersion relation. To this end, it is enough to show that: (i) $\operatorname{Im} l_f(s,t) < K/|s|^\alpha$, for some $\alpha > 0$ and that, (ii), as $|s| \to \infty$ in all directions of the complex plane

$$|l_f(s,t)| < \text{const}\, \frac{|\ln|\sin\theta||}{|s|^\alpha} = B(|s|,\theta). \tag{5.23}$$

If this is so, then, by (i) the right-hand side of Eq. (5.17) can be written for each $\operatorname{Im} l_f$ and defines a function $\bar{l}_f(s,t)$ which obeys the bound (5.23) as $|s| \to \infty$, in all complex directions. The difference $\Delta_f = l_f(s,t) - \bar{l}_f(s,t)$ is thus a function holomorphic for all finite s and satisfying the bound (5.23). Writing then that $|\Delta_f(s,t)|$ is less than the modulus of a function having no zeros inside the circle of radius $k\,|s|$, $k > 1$ and having a modulus on that circle equal to $B(k\,|s|,\theta)$, Eq. (5.23), we can verify that $|\Delta_f(s,t)|$ can in fact be bounded by arbitrarily small numbers, provided k is chosen sufficiently large. Thus, $\Delta_f(s) \equiv 0$.

We must therefore verify (i) and (ii) above. From (5.20) and (A6), for s sufficiently large, $\operatorname{Re}\ln\Omega_t(s) < \text{const} \times \ln s$, for $s \in C_s$ and the same is true for $\Omega_t(u)$. Also, one can verify that, for s complex, if (A6) (ii) is true

$$|\ln\Omega_t(s)| \leqslant \text{const}\, \ln(|s|/|\sin\theta|), \tag{5.24}$$

which implies (5.23) for $f = \Omega_t(s)$, $\Omega_t(u)$. There might still exist isolated points where $\operatorname{Re}(\ln\Omega_t(s))$ increases logarithmically [the isolated zeros of $A(s,t)$]. Then, the difference Δ_Ω between l_Ω and $\bar{l}_\Omega$ in the argument above might have isolated singularities at these points; however, these are such that $\lim_{s\to s_0}(s - s_0)\, \Delta_\Omega(s) = 0$($s_0$ is such a point); this means $\Delta_\Omega(s)$ is holomorphic at s_0 and there is no change in the proof. Thus, if $f = \Omega_t(s)$, $\Omega_t(u)$, $l_f(s,t)$ does satisfy a dispersion relation. To show that l_K satisfies it, we shall establish for the exponent $I(s,t)$ of $K_{0s}(s,t)$ the bound

$$|I(s,t)| < \hat{K}(r,t)|\ln|s||\tag{5.25}$$

with r the distance from t to the boundary $\partial\mathscr{D}_i$ of $\mathscr{D}_i$. This bound is valid for all s in $D_s \cup C_s^+ \cup C_s^-$ and its verification finishes the proof if $t \in \mathscr{D}_i$. It is enough to consider the contribution $I_i(s,t)$ of a single curve Γ_i; we write

$$|I_i(s,t)| \leqslant \frac{1}{2\pi r|w(t)|}\int_{a_i}^{b_i} \left|\ln(s - s')||w(t_i(s'))||\frac{dt_i}{ds'}\right| ds', \tag{5.26}$$

where b_i can be finite or infinite. If b_i is finite, then (5.25) is straightforward. If b_i is not finite, we write

$$\begin{aligned}|I_i(s,t)| &\leqslant \frac{1}{2\pi r}\left[\int_{a_i}^{|s|-1} + \int_{|s|-1}^{|s|+1} + \int_{|s|+1}^{\infty}\right] \\ &\quad\times\left(\frac{|\ln(s - s')|}{|w(t)|}\,|w(t_i(s'))||\frac{dt_i}{ds'}\,|ds'\right) \\ &= I_{i1}(s,t) + I_{i2}(s,t) + I_{i3}(s,t).\end{aligned} \tag{5.27}$$

Clearly,

$$\begin{aligned}|I_{i1}(s,t)| &\leqslant \frac{1}{2\pi r}\max_{a<s'<|s|-1}\left|\ln(s - s')|\,\frac{1}{|w(t)|}\right. \\ &\quad\left.\times\int_a^\infty |w(s')||\frac{dt_i}{ds'}||ds'| \leqslant \frac{M_i(t)}{2\pi r}|\ln|s|\right|.\end{aligned} \tag{5.28}$$

For $I_{i2}(s,t)$, we use (for $s \neq s'$)

$$\begin{aligned}|\ln(s - s')| &\leqslant |\ln|s - s'|| + \text{const};\ |\ln|s - s'|| \\ &\leqslant \ln\max(1/|s' - |s||, |s' + |s||) \\ &\leqslant -\ln|s' - |s|| + \ln|s' + |s||,\end{aligned}$$

$$|I_{i2}(s,t)| \leqslant \frac{M_i'(t)}{2\pi r} + \frac{\ln|s|}{2\pi r}M_i''(t). \tag{5.29}$$

For the third integral, we write

$$\begin{aligned}|\ln(s - s')| &\leqslant \ln|s - s'| + \text{const} \leqslant \ln(|s| + s') + \text{const} \\ &\leqslant \ln s' + \text{const},\end{aligned} \tag{5.30}$$

where we have used that $|s| < s'$, $|s - s'| > 1$. The resulting integral converges and tends to zero as $|s| \to \infty$. This proves (5.25) for $t \in \mathscr{D}_i$. With more careful estimates, we show in Appendix I that the factor $\hat{K}(r,t)$ in (5.25) does not get unbounded as t approaches Γ_i, but rather stays finite if t is not such that $|s_i(t')| \to \infty$ as $t' \to t$. In the latter case, we can only show that $|I(s,t)| < \hat{K}(t)\ln^2|s|$, which is still enough to justify (5.24). This is also shown in Appendix I.

This ends the construction of $A(s,t)$.

Clearly, we could have obtained $A(s,t)$ starting from the modulus in the s and t channels instead of s and u. As in Sec. III, there is the difficulty that $C(s,t)$ is singular at the end points of the curves Γ_i but, at each fixed s, $|\ln C(s,t)|$ is integrable, so that the construction can still be done. The condition that the amplitude obtained by the two routes is the same gives a continuous set of sum rules, to be fulfilled by the modulus in the t channel, if the problem is to have a solution.

An example of the construction of $A(s,t)$ in $\mathscr{M}$ is given in Appendix J.

VI. CONCLUSIONS

Equations (5.18), (5.17), (5.14), (5.11), (5.6), (5.5), (5.1), (4.29), (4.23), (2.18), and (4.15) allow an explicit construction of the amplitude $A(s,t,u)$, holomorphic in the Mandelstam domain, starting from its modulus in the three channels and a certain resolution of the discrete ambiguity in all channels. This achieves the aim stated in the Introduction.

The sequence of equations above has been derived starting from an amplitude $A(s,t,u)$ obeying analyticity in two variables in the Mandelstam domain and certain restrictions (A1)–(A7), Sec. IV. In the course of the paper, it turned out that the data of the problem—the modulus and the position of the zero trajectories—must be suitably restricted in order to be compatible with the analyticity properties of $A(s,t,u)$. These restrictions are given by conditions (4.28) for the position of the zero trajectories and by the equality of the results of the two possible constructions (5.18), starting from the modulus distribution in the s and u channels and s and t channels in turn.

In fact, we can state that equations (4.28) are sufficient conditions for picking out a resolution of the discrete ambiguity that is consistent with two-variable analyticity. Indeed, given oriented curves Γ_i, Γ'_j [i.e., functions $t_i(s)$, $t_j(u)$, $s \in C_s, u \in C_u$] and sets of n_j functions $s_{k_l}(t)$, $l = 1, \ldots, n_j$ defined on intervals (t_j, t_{j+1}) of C_t, we can construct a function $C(s,t)$, holomorphic in two variables in the Mandelstam domain and having the prescribed zero trajectories, as s,t or u take on physical values. The function $C(s,t)$ is given by Eq. (5.6).

In general, Eqs. (4.28) are not sufficient to eliminate all but one solution of the discrete ambiguity. It is easy to find examples where coherent reflections are possible, so that the identities (4.28) are left unimpaired. It is interesting to notice that, because of the reflection symmetry of the patterns of zeros in D_t, another solution of the discrete ambiguity leaves the pattern completely unchanged, except for the reversal of the sense of some curves $\Gamma_i(\Gamma'_i)$ and the replacement of the inverse functions $\sigma_i(\tau)$ of these curves by their complex conjugate; one might also need the replacement of some functions $\sigma_i(\tau)$ on C_t by their conjugate.

It is presumably true that if several solutions to (4.28) are available, and the given modulus of $A(s,t)$ satisfies the consistency conditions for one of the solutions, then it satisfies them for all the others. However, a proof of this statement is lacking at present.

The construction of amplitudes holomorphic in the Mandelstam domain required the use of a number of assumptions. We introduced them proceeding by analogy with the construction of Sec. II and III of amplitudes $f(\sigma,\tau)$, holomorphic in a neighborhood of $\overline{\mathscr{B}} = \{|\sigma| \leqslant 1 \otimes |\tau| \leqslant 1\}$. Let us recall these assumptions and their role shortly.

Assumption (A1) concerns the continuity of the amplitude at boundary points; it ensures that we can talk locally about zero trajectories and that their images in the t plane have a continuous tangent vector, except for some special points. Assumption (A2) allows one to "continue" zero trajectories on larger intervals of energy and specifies the nature of the cusps. Assumption (A3) is the most drastic one: there are a finite number of zero trajectories. This is in particular violated if the amplitude saturates the Froissart bound[32,33] or obeys some form of geometrical scaling.[34] If (A3) is dropped, one has to account for the convergence of the various series occurring in (4.15) and (4.23). Assumption (A3′) states, e.g., that there are no zero trajectories running along the cuts of the t plane as s moves on C_s. By a treatment similar to Appendix C, (A3′) may be dropped. Assumption (A4) restricts the behavior of the trajectories at end points and ensures that continuations of the inverse functions around these points can be performed (Lemma 6). Assumption (A5) restricts the behavior of zeros at infinity and (A6) restricts the behavior of both the modulus of the amplitude at large s,t,u and of its phase along the cuts. Finally (A7) is rather artificial and ensures that only a finite number of domains occur in the complement of the images of the curves Γ_i in D_t. Presumably this assumption is unnecessary.

The problem of constructing amplitudes starting from experimental information, in such a way that they are compatible with analyticity in two variables, has been considered in the past in other works from a somewhat different point of view. In Ref. 35 the authors set up a family of functions, holomorphic in $\mathscr{M}$, so that any amplitude satisfying an unsubtracted Mandelstam representation may be expended in terms of them. Applications to $\pi\pi$ scattering are discussed in Ref. 36. The clear difficulty of this method is that the functions in question are not orthogonal in the physical region, but rather in the double spectral domain. Consequently, the number of terms needed by the expansion to reproduce the details of the amplitude is very large and not easy to control.

In Ref. 37 (and papers cited therein) the authors discuss a two-variable (energy and angle) expansion of the amplitude, up to a given energy, in a certain channel, in terms of basis functions of irreducible representations of $O(4)$. It is, however, not their concern to ensure the correct global analyticity properties of the amplitude.

The interest in the problem of constructing scattering amplitudes given the modulus in the three channels stems mainly from the fact that its statement is very simple, and the differences between the treatment in one complex variable and that in two variables are substantial. It appeared to the author as a surprise that such a construction can be performed explicitly. From a practical point of view, it is difficult to ascertain at present whether these methods can be applied in particle physics. However, related procedures are certainly of interest in optics, in the problem of phase retrieval from a single, two-dimensional intensity distribution.[38,39]

APPENDIX A: WEIERSTRASS' PREPARATION THEOREM

We use Weierstrass' preparation theorem many times in this paper. In this Appendix, we recall it and make some comments on its use and on the implicit function theorem in complex function theory.

One possible statement of Weierstrass' theorem for the case of two variables is (Ref. 16, p. 86):

Let $F(s,t)$ be holomorphic in s,t in a domain D and let $F(s_0,t_0) = 0$, $(s_0,t_0) \in D$, but $F(s_0,t) \not\equiv 0$. Then, there exists a neighborhood $U \times V$ of (s_0,t_0), where $F(s,t)$ can be written in the form

$$F(s,t) = ((t-t_0)^n + a_1(s)(t-t_0)^{n-1} + \cdots + a_n(s))\Omega(s,t)$$
$$= W_{s_0}(s,t)\Omega(s,t) \tag{A1}$$

where n is an integer $\geqslant 1$, $\Omega(s,t)$ is a holomorphic and nonvanishing function of s,t in $U \times V$, $a_1(s)$, $a_2(s)$, $\cdots a_n(s)$ are holomorphic functions of s, for s in U and

$$a_1(s_0) = a_2(s_0) = \cdots = a_n(s_0) = 0. \tag{A2}$$

Thus the zeros of $F(s,t)$ can be described in the neighborhood $U \times V$ of (s_0,t_0) by a "Weierstrass pseudopolynomial" $W_{s_0}(s,t)$ "with peak at s_0."

If one drops the assumption $F(s_0,t) \not\equiv 0$, then it can be shown (Ref. 16, p. 89) that, for two complex variables the decomposition of $F(s,t)$ in $U \times V$ reads

$$F(s,t) = (s-s_0)^p W_{s_0}(s,t)\Omega(s,t) \tag{A3}$$

for some integer p.

a. We first point out that Weierstrass' theorem contains the usual implicit function theorem for holomorphic functions.

To this end, we verify first that $\partial F/\partial t\,(s_0,t_0) \neq 0$ implies that $p=0$ and $n=1$ in (A3). Then we see from (A1) that if $F(s_0,t_0)=0$ and $\partial F/\partial t\,(s_0,t_0)\neq 0$, F holomorphic in a domain $D_1 \otimes D_2(\ni (s_0,t_0))$, there exists a neighborhood U of s_0 and a function $t(s)=t_0-a_1(s)$, holomorphic in U, with values in D_2, such that $t(s_0)=t_0$ and $F(s,t(s))\equiv 0$, for $s\in U$, as stated by the implicit function theorem.

b. Weierstrass' theorem makes precise the behavior of the solutions $t(s)$ of $F(s,t)=0$ at points where $\partial F/\partial t\,(s,t)=0$.

Namely, the solutions $t(s)$ of $F(s,t)=0$ can have at most algebraic branch points of order $(2\leqslant)p\leqslant n$ at points s_0,t_0 where $\partial F/\partial t\,(s_0,t_0)=0$. In the neighborhood of these points a solution $t(s)$ can be expanded as

$$t(s)=t_0+a_1(s-s_0)^{1/p}+a_2(s-s_0)^{2/p}+\cdots \tag{A4}$$

and the expansion has a finite radius of convergence.

To show how these statements follow from Weierstrass' theorem, we shall quote some intermediate steps of the proof, to be found in Ref. 16, p. 92 ff. These statements may also clarify the meaning of Weierstrass' theorem:

1. Each pseudopolynomial $W_{s_0}(s,t)$ can be written in a unique manner as a product of irreducible pseudopolynomials $W_{s_0}^{(i)}(s,t)$, with the same peak

$$W_{s_0}(s,t)=\prod_i W_{s_0}^{(i)}(s,t) \tag{A5}$$

(see Ref. 16, p. 94).

2. For an irreducible pseudopolynomial $W_{s_0}(s,t)$ all the roots of the equation in t: $W_{s_0}(s,t)=0$ are distinct, if s lies in a sufficiently small neighborhood U_0 of s_0 $(s\neq s_0)$ (see Ref. 16, p. 94).

Consequently, we can apply the implicit function theorem to the equation $W_{s_0}(s,t)=0$ at all points in $U_0\times\mathscr{C}$, except for (s_0,t_0) to find a solution $t(s)$ and then its continuation as s moves in U_0 around s_0.

3. If $W_{s_0}(s,t)$ is irreducible, then any root $t(s)$ of $W_{s_0}(s,t)=0$ at $s\neq s_0$, $s\in U_0$, can be obtained from any other one $t(s_1), s_1\in U_0'\subset U_0, s_0\notin U_0'$, by analytic continuation along a suitable path in U_0, avoiding s_0 (cf. Ref. 16, p. 109).

As a consequence, if $W_{s_0}(s,t)$ is irreducible and has degree p with respect to t, if we continue analytically a root $t(s_1)$, defined on $U_0'\subset U_0$, $s_0\notin U_0'$, along a path surrounding s_0, we recover the function with which we started after precisely p turns. In other words, $W_{s_0}(t)$ defines a function $t(s)$ which has a branch point of order p at s_0.

We now introduce the variable u by $s=s_0+u^p$. The function $t=t(s_0+u^p)=t(u)$ is then uniform with respect to u in a neighborhood $\mathscr{U}$ of $u=0$ and, since it is bounded, it is holomorphic in $\mathscr{U}$. We can thus write

$$t(u)=\sum_n c_n u^n=\sum_n c_n(s-s_0)^{n/p} \tag{A6}$$

and the expansion converges in a disk contained in $\mathscr{U}$.

c. Thus, a solution $t(s)$ of $F(s,t)=0$ with $F(s,t)$ holomorphic in a domain $D_s\times D_t$, will be a holomorphic function of s in D_s except for isolated algebraic branch points.

There will exist, in general, several solutions $t(s)$ with t in D_t for s in D_s. Assume that for each s in D_s there are n solutions with t in D_t. Then, although each of these functions have in general branch points in D_s, the symmetric sums

$$S_k(s)=\sum_{i=1}^{n}(t_i(s))^k,\quad k=1,2,\ldots,n \tag{A7}$$

are holomorphic functions for s in D_s.

This can be seen in two ways: the first is applying Weierstrass' theorem to those points where one of the t_i's has branch points. If we continue $S_k(s)$ around such a point, we get a function which is uniform (and thus, since bounded, holomorphic even at s_0) because we just permute the indices of the branches of the algebraic function $t(s)$. The other method used in fact to prove Weierstrass' theorem (see Ref. 16, p. 86), is to write $S_k(s)$ as

$$S_k(s)=\frac{1}{2\pi i}\oint_{\partial D_t} t^k\,\frac{\partial F/\partial t\,(s,t)}{F(s,t)}, \tag{A8}$$

where the integral is performed counterclockwise along the boundary of D_t, where $F(s,t)\neq 0$ for s in D_s (and we assumed for simplicity that D_t is bounded and the length of ∂D_t is finite). The functions $S_k(s)$ in (A7) are holomorphic in D_s, since the integrand is holomorphic in D_s and the integral is done on a finite interval.

If the $S_k(s)$ are known, one can construct algebraically the "symmetric combinations" of the roots $t_i(s)$, defined as

$$\begin{aligned}\beta_1(s)&=-\sum_{i=1}^{n}t_i(s),\\ \beta_2(s)&=\sum_{i>j}t_i(s)t_j(s)\\ &\ \,\vdots\\ \beta_n(s)&=\prod_{i=1}^{n}t_i(s)(-1)^n\ .\end{aligned} \tag{A9}$$

These functions are holomorphic functions of s in D_s and are the coefficients of a polynomial $P(s,t)$ which vanishes in $D_s\otimes D_t$ precisely where $F(s,t)$ vanishes

$$P(s,t)=\sum_{k=0}^{n}t^k\beta_{n-k}(s). \tag{A10}$$

Thus, in $D_s\otimes D_t$, we can write

$$F(s,t)=P(s,t)\Omega(s,t), \tag{A11}$$

with $\Omega(s,t)$ nonvanishing in $D_s\otimes D_t$.

APPENDIX B: THE CURVES γ_i: $\tau=\tau_i(\sigma)$

We prove in this Appendix properties (a)–(f) of the curves γ_i, defined in Sec. II. We denote by $\bar{\gamma}_i$ the image set in the τ plane of the trajectory function $\gamma_i:\tau=\tau_i(\sigma)$, $|\sigma|=1$.

The curve γ_i has clearly a continuous tangent (i.e., $d\,\mathrm{Re}\,\tau/d\,\mathrm{Im}\,\tau$ or the reciprocal are continuous functions of the parameter σ) at all points τ_0 where $d\tau_i/d\sigma\neq 0$ and finite. At such points, we can define an inverse function $\sigma_i(\tau)$, $d\sigma_i/d\tau\neq 0$, holomorphic in the neighborhood of τ_0. The function $\sigma_i(\tau)$ has unit modulus on γ_i.

If $d\tau_i/d\sigma$ is not finite at $\sigma=\sigma_0$, $\tau_i(\sigma_0)=\tau_0$, we know from Eq. (A4) that the change of variable $\tilde{\sigma}=(\sigma-\sigma_0)^{1/p}$, $p>1$, integer, renders $d\tau_i/d\tilde{\sigma}$ finite or zero. Thus, $|d\tau_i/d\sigma|$ is integrable at σ_0. Several curves $\gamma_j:\tau=\tau_j(\sigma)$ meet at $\tau_{0j}=\tau_j(\sigma_0)$; we have a choice as to the definition of γ_i on both sides of τ_0. As stated in the text, we choose as continuation the image through $\tau_i(\tilde{\sigma})$ of a line making an angle π/p in the $\tilde{\sigma}$ plane with the original direction (in a clockwise sense). We obtain thus, in general, a cusp of γ_i. The inverse function $\sigma_i(\tau)$ maybe holomorphic in a neighborhood of $\tau_0 = \tau_i(\sigma_0)$, or have a branch point at τ_0.

If $d\tau_i/d\sigma = 0$ at $\sigma = \sigma_0$, a cusp of γ_i might appear, although not necessarily [compare the situations of $\tau - (\sigma - 1)^2 = 0$ which has a cusp at $\tau = 0$ with that of $\tau - (\sigma - 1)^3 = 0$ which has no cusp]. It is not necessary that other γ_l meet γ_i at τ_0, but this may certainly be so [see $\tau^2 - (\sigma - 1)^5 = 0$]. The inverse function $\sigma_i(\tau)$ has a branch point at τ_0 and is Hölder continuous along γ_i, with some rational index.

We next show that there are only a finite number of points along $|\sigma| = 1$, where $d\tau_i/d\sigma$ is not finite or zero. To this end, assume that an infinite sequence of such points existed; then it would accumulate at (σ_0,τ_0) in $\bar{\mathscr{B}}$, $|\sigma_0| = 1$ and $f(\sigma_0,\tau_0) = 0$. From Weierstrass' preparation theorem, the zeros of $f(\sigma,\tau)$ are described in a neighborhood $U_{\sigma_0} \times U_{\tau_0}$ of (σ_0,τ_0) by a finite product of irreducible pseudopolynomials $W_{\sigma_0}^{(k)}(\sigma,\tau)$, with coefficients holomorphic in U_{σ_0}. Each $W_{\sigma_0}^{(k)}(\sigma,\tau)$ describes locally a function $\tau_k(\sigma)$ and for at least one of these, say $\tau_1(\sigma)$, and infinity of points $\sigma_n \to \sigma_0$ should exist where $d\tau_1/d\sigma = \infty$ or zero. However, (see Appendix A), there exists a neighborhood U_0 of σ_0, so that all functions $\tau_k(\sigma)$, described by $W_{\sigma_0}^{(k)}(\sigma,\tau)$, $k = 1,2, \ldots$, are holomorphic in U_0, except at σ_0. Consequently, there cannot be an accumulation to σ_0 of points σ_n, with $d\tau_1/d\sigma$ not finite or zero, except if $\tau_1(\sigma)$ is holomorphic at all σ_n, and $d\tau_1/d\sigma(\sigma_n) = 0$. This implies, however, that $\tau_1(\sigma) = \text{const} = \tau_0$, a situation which we can exclude for simplicity.

The argument shows that, even if all zero surfaces of $f(\sigma,\tau)$ are included, the number of points where $d\tau_i/d\sigma = 0$ or infinite in $\bar{\mathscr{B}}$ is finite.

We now consider property (c): if a curve γ_1 stays entirely contained in $|\tau| < 1$, then we must show that it has to be closed. To this end we consider the images $\mathscr{I}_k = (\tau_i^{(k)}(\sigma_1), \tau_i^{(k)}(\sigma_2))$ of an interval (σ_1,σ_2) along $|\sigma| = 1$ ($|\sigma_1| = 1$, $|\sigma_2| = 1$), under the successive branches $\tau_i^{(k)}(\sigma)$, reached by counterclockwise continuation along $|\sigma| = 1$ (with the usual convention). According to property (a), we are free to assume that there are no branch points of $\tau_i(\sigma)$ on (σ_1,σ_2). Also, if N is the number of roots in τ of $f(\sigma,\tau) = 0$ in $|\tau| < 1$ at $\sigma \in (\sigma_1,\sigma_2)$, it stays constant along (σ_1,σ_2) if the latter is sufficiently small; thus the zero trajectories of $f(\sigma,\tau)$ consist on (σ_1,σ_2) of N functions of σ, $\tilde{\tau}_j(\sigma)$, $j = 1,2, \ldots, N$, holomorphic on (σ_1,σ_2) and whose images are the segments $I_j = (\tilde{\tau}_j(\sigma_1), \tilde{\tau}_j(\sigma_2))$. It follows that the images $\mathscr{I}_k$ must coincide with a subset of the I_j's. Thus, by analytic continuation, we can obtain only a finite number of images $\mathscr{I}_k$. Let the sequence of distinct images obtained by continuation from $\mathscr{I}_1$ be $\mathscr{I}_1, \mathscr{I}_2, \ldots, \mathscr{I}_m$. We still have to show that, if we continue $\tau_1(\sigma)$ further (with the usual convention at branch points), we have to reach $\mathscr{I}_1$, so that, indeed, the curve will be closed. Assume we obtained, however, $\mathscr{I}_k$, $k \neq 1$, $(1 < k < m)$. If we continue counterclockwise further, we must obtain a loop consisting of $(\mathscr{I}_k, \mathscr{I}_{k+1}, \ldots, \mathscr{I}_m)$ by the uniqueness of analytic continuation. But continuing around $|\sigma| = 1$ in the reverse sense, starting from (σ_1,σ_2) and $\mathscr{I}_k$ we obtain either $\mathscr{I}_{k-1}$ or $\mathscr{I}_m$, and $\mathscr{I}_{k-1} \neq \mathscr{I}_m$, if $k \neq 1$. This is, however, impossible and proves our statement.

Consider now the situation of a curve γ_1 that "enters" $|\tau| < 1$; this means that, at some $\sigma = \sigma_0$ on $|\sigma| = 1$, $|\tau_1(\sigma_0)| = 1$ and that, for σ sufficiently close to σ_0, $|\tau_1(\sigma)| < 1$ for σ on one side ("forward") of σ_0, $|\tau_1(\sigma)| > 1$ on the other side. We claim that there must exist σ_0', $|\sigma_0'| = 1$, such that (i) the analytic continuation of $\tau_1(\sigma)$, counterclockwise along $|\sigma| = 1$ reaches again $|\tau| = 1$ at $\sigma = \sigma_0'$;(ii) $\tau_1(\sigma)$ takes on values larger than unity in a neighborhood of σ_0' on $|\sigma| = 1$ (γ_1 "leaves" $|\tau| \leqslant 1$).

To this end, we consider again a segment (σ_1,σ_2) in a sufficiently small neighborhood of σ_0, on the unit circle; we can assume that all the roots $\tau_i(\sigma)$ of $f(\sigma,\tau) = 0$ in $|\tau| < 1 + \epsilon_0 < 1 + \epsilon$ are regular functions on (σ_1,σ_2) and consider again the images $\mathscr{I}_k$ of (σ_1,σ_2) through the continuations $\tau_1^{(k)}(\sigma)$ of $\tau_1(\sigma)$ counterclockwise around $|\sigma| = 1$. By definition, there is one of them, $\mathscr{I}_1$, which contains points with $|\tau| = 1 + \epsilon_1$, $\epsilon_0 > \epsilon_1 \geqslant 0$ and is traversed from $|\tau| > 1 + \epsilon_1$ to $|\tau| < 1 + \epsilon_1$. Now, if with increasing k, $\tau_1^{(k)}(\sigma)$ leaves $|\tau| \leqslant 1 + \epsilon_0$, it must cross $|\tau| = 1$ at some σ_0' and the proof is finished. So assume $\tau_1^{(k)}(\sigma)$ stays inside $|\tau| < 1 + \epsilon_0$ for all k; then, we have seen that we must reach $\mathscr{I}_1$ again, for sufficiently large k [$k > N, N =$ the number of roots in $|\tau| < 1 + \epsilon_0$ of $f(\sigma,\tau) = 0$, $\sigma \in (\sigma_1,\sigma_2), |\sigma| = 1$]. Moreover, we must traverse $\mathscr{I}_1$ from large $|\tau|$ to small $|\tau|$ on (σ_1,σ_2). But, since $\tau(\sigma)$ is a continuous function of σ, there must then exist σ_1', $|\sigma_1'| = 1$, so that $\tau(\sigma)$ crosses $|\tau| = 1 + \epsilon_1$ from $|\tau| < 1 + \epsilon_1$ to $|\tau| > 1 + \epsilon_1$. There must then exist also a σ_0' with the same property with respect to $|\tau| = 1$. Property (c) is thus justified.

We now go over to property (d): if there existed an infinite number of γ_i's, they can at most be of the type with "two ends": otherwise, they would lead to an infinite number of roots of $f(\sigma,\tau) = 0$ at fixed $\sigma, |\sigma| = 1$. Assume that a sequence of "endpoints" $\tilde{\tau}_j$ of curves $\gamma_j : \tau = \tau_j(\sigma)$ existed, $|\tilde{\tau}_j| = 1$, $\tilde{\tau}_j = \tau(\tilde{\sigma}_j)$, $|\tilde{\sigma}_j| = 1$ and accumulated at some (σ_0,τ_0), $|\sigma_0| = 1$, $|\tau_0| = 1$. By continuity, $f(\sigma_0,\tau_0) = 0$. Applying Weierstrass' theorem to $f(\sigma,\tau)$ in a neighborhood $U_{\sigma_0} \times U_{\tau_0}$ of (σ_0,τ_0), we conclude there must exist an irreducible pseudopolynomial $W_{\sigma_0}(\sigma,\tau)$ vanishing on an infinite subsequence $\{\tilde{\sigma}_{j_k}, \tilde{\tau}_{j_k}\}$ of $\{\tilde{\sigma}_j, \tilde{\tau}_j\}$. As a consequence, the function $\tau(\sigma)$ defined by $W_{\sigma_0}(\sigma,\tau) = 0$ must obey: $\tau(\sigma)\tau^*(1/\sigma^*) = 1$ [and $\tau(\tilde{\sigma}_{j_k}) = \tilde{\tau}_{j_k}$]. But on the other hand, by hypothesis, in any neighborhood of each $\tilde{\sigma}_j$, there exist points σ_j', $|\sigma_j'| = 1$, with $|\tau_j(\sigma_j')| \neq 1$, and such that $f(\sigma_j', \tau_j(\sigma_j')) \equiv 0$; as $\sigma_j' \to \tilde{\sigma}_j$, $\tau_j(\sigma_j') \to \tilde{\tau}_j$. Thus, for σ in a neighborhood of $\tilde{\sigma}_{j_k}$, $|\sigma| = 1$, $f(\sigma,\tau)$ vanishes both at a point on $|\tau| = 1$ and at $\tau_{j_k}(\sigma_j')$; at $\sigma = \tilde{\sigma}_{j_k}, f(\sigma,\tau)$ has thus at least a double zero, i.e., $\partial f/\partial \tau$ $(\tilde{\sigma}_{j_k}, \tilde{\tau}_{j_k}) = 0$. It follows by continuity that $\partial f/\partial \tau(\sigma_0,\tau_0) = 0$. Further, from Weierstrass' preparation theorem applied to $\partial f/\partial \tau$ at (σ_0,τ_0) we conclude as above that part of its zeros must be described by the same pseudopolynomial $W_{\sigma_0}(\sigma,\tau)$ as for $f(\sigma,\tau)$. Thus, $f(\sigma,\tau)$ has at $\tau = \tau(\sigma)$ a zero of the second order for each σ in a neighborhood of σ_0. But then $f(\tilde{\sigma}_{j_k},\tau)$ must have a zero of the third order (at least) at $\tau = \tilde{\tau}_{j_k}$, etc. Since this can be repeated indefinitely, $\partial^l f/\partial\tau^l(\tilde{\sigma}_{j_k},\tau) = 0$ at $\tau = \tilde{\tau}_{j_k}$ for all l; thus $f(\tilde{\sigma}_{j_k},\tau) \equiv 0$, for all j_k, or $f \equiv 0$, a situation which we can *a priori* exclude. This concludes the proof of property (d).

We now discuss property (e): assume two trajectories γ_1, γ_2 are such that they intersect an infinite number of times. By this we mean there exist two sequences of points $\{\sigma_n\},\{\sigma'_n\}$, so that $\tau_1(\sigma_n) = \tau_2(\sigma'_n) = \tau_n$. The sequence τ_n accumulates at τ_0, $|\tau_0| \leqslant 1$. There exists a neighborhood of τ_0, so that the inverse functions $\sigma_1(\tau)$, $\sigma_2(\tau)$ are holomorphic in it apart from a possible branch point at τ_0. Both $\sigma_1(\tau)$, $\sigma_2(\tau)$ have unit modulus on γ_1, γ_2, respectively. Consider then $\sigma_2(\tau_1(\sigma))$. At the intersection points τ_n, $|\sigma_n| = |\sigma_1(\tau_n)| = 1$, and $\sigma_n \to \sigma_0$, as $n \to \infty$. Since $|\sigma_2(\tau_n)| = 1$, it follows that $|\sigma_2(\sigma)| = 1$, if $|\sigma| = 1$, i.e., $|\sigma_2(\tau)| = 1$ for τ on γ_1. Assume now that at $\tau = \tau_n$, $d\sigma_2/d\tau \neq 0$. Then a sufficiently small neighborhood U_{τ_n} of τ_n is mapped one-to-one by $\sigma_2(\tau)$ onto a domain U_{σ_n} containing σ_n and part of $|\sigma| = 1$. Both sets $\sigma_2(\gamma_1 \cap U_{\tau_n})$, $\sigma_2(\gamma_2 \cap U_{\tau_n})$ are contained in U_{σ_n}, are segments of the unit circle, and contain σ_n. Thus the mapping $\sigma_2(\tau)$ assumes the same value at points on γ_1 and γ_2, as close as one wishes to τ_n and this is impossible, unless γ_1,γ_2 coincide. If, however, $d\sigma_2/d\tau(\tau = \tau_n) = 0$, for all n, it follows that $\sigma_2(\tau) = \sigma_{20} = \text{const}$. This latter situation we can exclude by dividing first $f(\sigma,\tau)$ by $\sigma - \sigma_{20}$ (the corresponding curve γ_2 consists of the whole plane). This proves property (e). The same reasoning shows that no γ_i can intersect itself an infinite number of times.

The fact that $\mathrm{C}(U_i\gamma_i)$ consists of a finite number of domains [property (f)] follows directly from (e) and (d). Property (6) is evident from the proof of (a) and (c).

APPENDIX C: COMPLEMENT TO SEC. II

We discuss the modifications needed in the definitions of $\mathscr{N}(\tau)$, Eq. (2.4) and $\Omega^{(k)}(\tau)$, Eq. (2.10) if assumption (A), Sec. II, is dropped. Thus, we assume that there exists at least a root $\sigma = \sigma(\tau)$ of $f(\sigma,\tau) = 0$ so that $|\sigma(\tau)| = 1$, at least for some interval (τ_1,τ_2) of values of τ on the circle $|\tau| = 1$.

We notice first that, if $\epsilon > 0$ is sufficiently small, assumption (A) is valid for τ on any circle of radius $1 > R > 1 - \epsilon$. To verify this, suppose it were not true: there would exist then a sequence $\epsilon_n \to 0$ with the property that, on each circle $|\tau| = 1 - \epsilon_n$, an interval (τ'_n,τ''_n) exists, so that a root $\sigma_n(\tau)$ of $f(\sigma,\tau) = 0$ has unit modulus, $|\sigma_n(\tau)| = 1$, for $\tau \in (\tau'_n,\tau''_n)$. Apply then Weierstrass' preparation theorem to a neighborhood $U_{\sigma_0} \times U_{\tau_0}$ of the point (σ_0,τ_0), $|\sigma_0| = 1$, $|\tau_0| = 1$, which is an accumulation point for the intervals (τ'_n,τ''_n) and their images $(\sigma_n(\tau'_n), \sigma_n(\tau''_n))$, $|\sigma_n(\tau)| = 1$, $\tau \in (\tau'_n,\tau''_n)$. A root $\sigma(\tau)$ of at least an irreducible Weierstrass pseudopolynomial (with peak at s_0) must then obey $\sigma(\tau)\sigma^*((1-\epsilon_n)^2/\tau^*) = 1$, $\tau \in U_{\tau_0}$, since for n sufficiently large, $U_{\tau_0} \cap (\tau'_n, \tau''_n) \neq 0$. However, this entails $\sigma((1-\epsilon_n)^2/\tau) = \sigma((1-\epsilon_m)^2/\tau)$, for an infinite sequence of integers m, which is impossible, unless $\sigma(\tau)$ is a constant.

As a consequence, if we use the definitions (2.4) of $\mathscr{N}(\tau)$ and (2.10) of $\Omega^{(k)}(\tau)$, for some circle of radius $|\tau| = 1 - \epsilon, \epsilon > 0$ and sufficiently small, we can expect both Theorems 2.1 and 2.2 to stay true, only provided we know values of n_i and of $\sigma(\tau)$, $|\tau| = 1 - \epsilon$, that we should put in. We now show that this information can be obtained in a simple manner from the one on $|\tau| = 1$ for $|\sigma| \leqslant 1$: on each interval (τ_i,τ_{i+1}) of $|\tau| = 1$ where the number of zeros of $f(\sigma,\tau)$ in $|\sigma| \leqslant 1$ is constant, divide these into three classes: (i) n_{1i} zeros with $|\sigma(\tau)| < 1$; (ii) n_{2i} zeros with $|\sigma(\tau)| = 1$, moving counterclockwise as τ moves counterclockwise on $|\tau| = 1$; *(iii)* n_{3i} zeros moving clockwise at a counterclockwise motion of τ on $|\tau| = 1$; then $n_i = n_{1i} + n_{2i}$ in Eq. (2.4) and only $\sigma_i(\tau)$ in classes (i) and (ii) should be used in Eq. (2.10).

To see this, it is enough to check that, in a domain $\mathscr{D}_1$ adjacent to $|\tau| = 1$, $n(\tau)$ is indeed $n_{1i} + n_{2i}$. Since $n(\tau)$ is constant in $\mathscr{D}_1$, it suffices to find it in the neighborhood $U_1 \cap (|\tau| < 1)$ of a certain point τ_1, $|\tau_1| = 1$, where $\sigma_i(\tau)$ are regular functions, with $d\sigma_i/d\tau \neq 0$. Clearly, the n_{1i} zeros stay inside $|\sigma| < 1$ for τ in $U_1 \cap (|\tau| < 1)$. But $\sigma_i(\tau)$ maps $U_1 \cap (|\tau| < 1)$ onto the outside of $|\sigma| < 1$ in case (iii) and on the inside of $|\sigma| < 1$ in case (ii) (as one sees from the Cauchy–Riemann equations). This proves our assertion.

As an example, if $f(\sigma,\tau) = (1 - \sigma^3\tau)(\sigma^2 - \tau)$ on $|\sigma| = 1$, there are five roots of the equation $f(\sigma,\tau) = 0$, all of them with unit modulus. However, for $|\tau| = 1 - \epsilon$, only two roots are inside the unit disk $|\sigma| \leqslant 1$. Thus $n = 2$ and only $\sigma = \pm\sqrt{\tau}$ should be used in Eqs. (2.4) and (2.10).

The construction of Sec. II was performed ignoring the possible isolated zeros of $f(\sigma,\tau)$ present on $|\sigma| = 1 \otimes |\tau| = 1$, i.e., which are not end points of γ_i. Since the constructions of Theorems 2.1 and 2.2 require integrals over γ_i in order to find the Cousin data in $\overline{\mathscr{B}}$, this procedure seems justified. We show now directly that, indeed, if a zero of $f(\sigma,\tau)$ on $|\sigma| = 1 \otimes |\tau| = 1$ is isolated in $\partial\mathscr{B}$ (it is not an end point of γ_i), it is also isolated in $\overline{\mathscr{B}}$ [i.e., it is not a limit point of zeros of $f(\sigma,\tau)$ in $\mathscr{B}$]. Consequently, it will not affect the counting of zeros in the domains $\mathscr{D}_i$, as done in Theorem 2.1.

Indeed, assume (σ_0,τ_0) were such an isolated zero, $|\sigma_0| = |\tau_0| = 1$, and let U_{σ_0}, U_{τ_0} be sufficiently small neighborhoods of σ_0,τ_0. By Weierstrass' theorem, there exists (at least) a function $\tau(\sigma)$, holomorphic in U_{σ_0}, possibly with a branch point at σ_0, so that $f(\sigma,\tau(\sigma)) \equiv 0$, $\tau(\sigma_0) = \tau_0$. By hypothesis, (i) if $\sigma \in U_{\sigma_0}$, $|\sigma| = 1$, $|\tau(\sigma)| > 1$, and we may assume $\tau(\sigma) \in U_{\tau_0}$; also, (ii) if $|\sigma| < 1$, $\sigma \in U_{\sigma_0}$, $|\tau(\sigma)| \neq 1$, $\tau(\sigma) \in U_{\tau_0}$. Assume nevertheless that, for $|\sigma| < 1$, (a determination of) $\tau(\sigma)$ is such that $|\tau(\sigma)| < 1$. Join then $\sigma \in U_{\sigma_0}$, $|\sigma| < 1$ to $\bar{\sigma}_1$ on $|\sigma| = 1$ by a small line contained in U_{σ_0} and continue $\tau(\sigma)$ along it. Since, for $|\sigma| = 1$, $|\tau(\sigma)| > 1$, there must exist $\bar{\sigma}$, $\bar{\sigma} \in U_{\sigma_0}$, $|\bar{\sigma}| < 1$, so that $|\tau(\bar{\sigma})| = 1$. This, however, is impossible by hypothesis (ii) above about $\tau(\sigma)$. Thus, our assertion is proved. As an example, $f(\sigma,\tau) = \sigma(1 + \tau) - 2$ has an isolated zero at $\sigma = \tau = 1$.

APPENDIX D: THE CHOICE OF BRANCHES OF THE LOGARITHM IN EQ. (3.1)

We show in this Appendix that the determinations of the logarithm in the integrand of Eq. (3.1) can be chosen so that $\mathscr{K}_1(\sigma,\tau)$ is real analytic. We treat separately open curves γ_i (with ends on $|\tau| = 1$) and closed curves (contained in $|\tau| \leqslant 1$).

(i) if γ_1 ends on $|\tau| = 1$ at $\tau = \tau_a$, take for $|\sigma| < 1$, $\operatorname{Im}\sigma = 0$, any branch of $\ln(\sigma - \sigma_1(\tau_a))$ and then continue it along γ_1. Since $f(\sigma,\tau)$ is real analytic, to γ_1: $\tau = \tau_1(\sigma)$ there corresponds a curve $\gamma_2 \equiv \gamma_1^*$: $\tau = \tau_2(\sigma) = \tau_1^*(\sigma^*)$, which ends on $|\tau| = 1$ at $\tau = \tau_a^*$. Assume first $\gamma_1^* \neq \gamma_1$. We choose the logarithm for $\tau \in \gamma_2$, [along which the function $\sigma_2(\tau) = \sigma_1^*(\tau^*)$ is defined] so that $\ln(\sigma - \sigma_2(\tau)) = \ln(\sigma - \sigma_1(\tau^*))^*$. It is enough to choose this at $\tau = \tau_a^*(\text{Im}\,\sigma = 0)$ and follow by continuity along γ_2 in the sense opposite to the one assigned in Sec. II [a possible example is $f(\sigma,\tau) = (\sigma/2 - \tau)^2 + 1$]. Once a choice of branches is made for σ real, it follows by continuity for all $|\sigma| < 1$.

Assume now $\gamma_1^* \equiv \gamma_1$. This means that to each function element $\tilde{\tau}(\sigma)$ defining γ_1 locally, there corresponds an element $\tilde{\tau}'(\sigma) = \tilde{\tau}^*(\sigma^*)$, which can be obtained from the former by analytic continuation (with the usual convention at branch points) along $|\sigma| = 1$, and such that $|\tilde{\tau}(\sigma)| < 1$ in the course of the process. Especially consider the function element $\tilde{\tau}_a(\sigma)$ in the neighborhood of the point σ_a where γ_1 "enters" the unit disk $|\tau| \leqslant 1$. Then $\tilde{\tau}_e(\sigma) = \tilde{\tau}_a^*(\sigma^*)$ is a function element describing γ_1 as it "goes out" of $|\tau| \leqslant 1$. Now, $\tilde{\tau}_e(\sigma)$ can be obtained from $\tilde{\tau}_a(\sigma)$ by analytic continuation (in the counterclockwise sense) along a path $\mathscr{P}$ consisting of none or several copies of the unit circle plus the arc (σ_a, σ_a^*) along $|\sigma| = 1$ (with possible indentations). It is easy to verify that the path $\mathscr{P}/2$ has an end at $\sigma_1 = +1$ or $\sigma_1 = -1$. We claim that the function element $\tilde{\tau}'(\sigma)$ obtained in a neighborhood of σ_1, by continuation along $\mathscr{P}/2$ from $\tilde{\tau}_a(\sigma)$ is real for σ real. Indeed, $\tilde{\tau}'(\sigma)$ can also be obtained by continuation along $\mathscr{P}^*/2$ starting from $\tilde{\tau}_e(\sigma) = \tilde{\tau}_a^*(\sigma^*)$. The two results must be identical, so that $\tilde{\tau}'(\sigma) = \tilde{\tau}'^*(\sigma)$ for σ real. This shows that $\tau_1 \equiv \tilde{\tau}(\sigma_1)$ is real.

If $\sigma_1 = -1$, we choose the logarithm in (3.1) so that $\text{Im}\ln(\sigma - \sigma_1(\tau_1)) = \text{Im}\ln(\sigma + 1) = 0$ for σ real, $-1 < \sigma < 1$, and then follow by continuity along γ_1. Here, $\sigma_1(\tau)$ is the inverse function associated to γ_1. In particular, it will follow that $\ln(\sigma - \sigma_1(\tau_a)) = \ln(\sigma - \sigma_1(\tau_e))^*$.

If $\sigma_1 = +1$, consider σ_1' near σ_1 on the path $\mathscr{P}/2$ joining σ_a to σ_1, and σ_1'' near σ_1 on the path $\mathscr{P}^*/2$ *joining* σ_a^* to σ_1. We choose, for $-1 < \sigma < 1$,
$\text{Im}\ln(\sigma - \sigma_1(\tau_{1+})) \equiv \text{Im}\ln(\sigma - \sigma_1') = \pi$,
$\text{Im}\ln(\sigma - \sigma_1(\tau_{1-})) \equiv \text{Im}\ln(\sigma - \sigma_1'') = -\pi$, and then follow by continuity along $\mathscr{P}/2$ and $\mathscr{P}^*/2$. There is a discontinuity of $2\pi i$ in the logarithm of Eq. (3.1) at $\tau = \tau_1$. It always causes a pole in $\mathscr{K}_1(\sigma,\tau)$ so that in Eq. (3.2), $k_1 = 1$. If $\sigma_1 = 1$, the point τ_1 will be among the points τ_{oi} [Eq. (3.2)] [the case $\sigma_1 = +1$ occurs in $f(\sigma,\tau) = \sigma/2 - \tau - 1$, the case $\sigma_1 = -1$ in $f(\sigma,\tau) = \sigma/2 - \tau + 1$].

(ii) Assume now γ_1 is closed. It can happen that, if $\tau_1(\sigma)$ is any function element locally defining it, its conjugate $\tau_1'(\sigma^*) = \tau_1^*(\sigma)$ is obtained from it by an analytic continuation along $|\sigma| = 1$ or not. In the former case, $\gamma_1 \equiv \gamma_1^*$; in the latter case, another closed curve (denoted by $\gamma_2 \equiv \gamma_1^*$) exists, generated by continuation of $\tau_2(\sigma) = \tau_1^*(\sigma^*)$ along $|\sigma| = 1$.

Consider this latter case first [as an example, consider $f(\sigma,\tau) = (\sigma/8 - \tau)^2 + 1/4$]. Since both γ_1, γ_1^* are closed they contain (in turn) points $\tau_{1,2}(\sigma = -1) = \tau_0$, τ_0^*. If $\tilde{\tau}_1(\sigma)$ is a function element of γ_1, in the neighborhood of $\sigma = -1$, we choose $\tilde{\tau}_2(\sigma) = \tau_1^*(\sigma^*)$ and define (for $\text{Im}\,\sigma = 0$)
$\text{Im}\ln(\sigma - \sigma_1(\tau_0)) = \text{Im}\ln(\sigma - \sigma_2(\tau_0^*))$
$= \text{Im}\ln(\sigma + 1) = 0$, $[\sigma_1(\tau), \sigma_2(\tau)$ represent the local inverses of $\tilde{\tau}_1(\sigma), \tilde{\tau}_2(\sigma)]$, and then continue along γ_1 in the counterclockwise sense, and along γ_1^* in the clockwise sense. If $\sigma_1(\tau)$ rotates p times around $|\sigma| = 1$, until it reaches its original value by continuation along γ_1, there will be a discontinuity of the logarithm of $2\pi p i$ at τ_0 leading to a pole of $\mathscr{K}_1(\sigma,\tau)$. Thus, τ_0 belongs to the collection of τ_{oi}, of Eq. (2.3), with $k_1 = p$. One can verify that for γ_1^*, the discontinuity leads again to a pole of $\mathscr{K}_1(\sigma,\tau)$ at $\tau = \tau_0^*$, with the same value of $k_2 = k_1 = p$.

Consider now the case $\gamma_1 \equiv \gamma_1^*$. Among the k distinct function elements $\tau_1'(\sigma), \ldots, \tau_1^{(k)}(\sigma)$ defining γ_1: $\tau = \tau_1(\sigma)$ in the neighborhood of $\sigma = -1$ (assuming for simplicity $d\tau_1^{(k)}/d\sigma \neq 0$ and finite at $\sigma = -1$), there exists either one whose inverse satisfies $\sigma_1'(\tau) = \sigma_1'^*(\tau^*)$ or a pair with complex conjugate inverses, $\sigma_1'(\tau)$, $\sigma_1''(\tau)$, $\sigma_1'(\tau) = \sigma_1''^*(\tau^*)$.

In the former case, let $\tau_1'(\sigma = -1) = \tau_a$, τ_a real; in the latter $\tau_1'(\sigma = -1)^* = \tau_1''(\sigma = -1)^* = \tau_a$ complex. Choose in both situations, for σ real, $\text{Im}\ln(\sigma - \sigma_1'(\tau_a))$
$= \text{Im}\ln(\sigma + 1) = 0 = \text{Im}\ln(\sigma - \sigma_1''(\tau_a^*))$.

Consider now the first case. Let $\mathscr{P}$ be the path described along $|\sigma| = 1$ by σ_1' as we return it to its original value along γ_1. The point on $|\sigma| = 1$ reached after $\mathscr{P}/2$ is either $\sigma_0 = +1$ or -1. The value $\tau_1'(\sigma_0) = \tau_0$ obtained from $\tau_1'(\sigma)$ by continuation to that point is real as one checks by obtaining it along $\mathscr{P}^*/2$. Let r be the number (integer or half-integer) of turns of $|\sigma| = 1$ contained in $\mathscr{P}/2$. Defining the logarithm along γ_1 by continuation along $\mathscr{P}/2$ and $\mathscr{P}^*/2$ away from τ_a, we obtain a discontinuity of $2r \times 2\pi i$ at τ_0 (counted in the integration sense), thus, $\tau_0 \in \{\tau_{oi}\}$ of Eq. (3.2) and the corresponding $k_1 = +2r$ [for an example, see Appendix F Example 1].

We now turn to the second case. Let $\mathscr{P}$ be the path along $|\sigma| = 1$ needed to obtain $\tau_1''(\sigma)$ from $\tau_1'(\sigma)$ by counterclockwise continuation. Again, the value obtained at $\mathscr{P}/2$ (at $\sigma_e = +1$) is a real number τ_e. The discontinuity of the logarithm at τ_e is again $2r \times 2\pi i$, if we define it by continuity from $\tau_1'(\sigma = -1) = \tau_a$, and $\tau_1''(\sigma = -1) = \tau_a^*$ along $\mathscr{P}/2$ and $\mathscr{P}^*/2$. If $\mathscr{P}_1$ is the path along $|\sigma| = 1$ needed to obtain $\tau_1''(\sigma)$ from $\tau_1'(\sigma)$ by clockwise continuation, we obtain again a real value $\tau_e'(\sigma_e')$, $\sigma_e' = +1$ being reached after $\mathscr{P}_1/2$. The discontinuity at τ_e', in the integration sense is given by $2r_1 \times 2\pi i$, where r_1 is defined in the same way as r above, with respect to $\mathscr{P}_1$. Thus γ_1 is parted into two halves $\gamma_1', \gamma_1'^*$, with ends at τ_e, τ_e'; τ_e, τ_e' belong to the set $\{\tau_{oi}\}$ of Eq. (3.2) with values $k_e = r, k_e' = r_1$. An example is afforded by $f(\sigma,\tau) = \tau^2 - \sigma/2$.

We now verify that, indeed, this choice of branches makes $\mathscr{K}(\sigma,\tau)$, Eq. (3.2) real analytic, i.e., real for σ,τ real. It follows from the above discussion that the points $\{\tau_{oi}\}$ in Eq. (3.2) are either real or fall into complex conjugate pairs. Consider next the factors of $\mathscr{K}_1(\sigma,\tau)$, Eq. (3.1), corresponding to the various γ_i. If $\gamma_i \neq \gamma_i^* \equiv \gamma_{i+1}$ (closed or "open") then, as $\sigma_i(\tau)$ moves counterclockwise in the integral along γ_i, $\sigma_{i+1}(\tau^*) = \sigma_i^*(\tau)$ moves clockwise, and thus against the integration sense, $\tau^* \in \gamma_{i+1}$. Reversing the sense of the integral on γ_{i+1} we obtain, according to the choice above for the determinations of the logarithm, the difference of two complex conjugate expressions in the exponent of $\mathscr{K}_1(\sigma,\tau)$, if σ,τ are real. The factor $1/(2\pi i)$ ensures then that the result is purely real. If $\gamma_1 \equiv \gamma_1^*$, we have seen that, in both the "open" and "closed" case, we can part the curve into two halves, γ_1', $\gamma_1'^*$, so that if $\tau_1(\sigma)$ generates γ_1', $\tau_1^*(\sigma)$ generates $\gamma_1'^*$. The logarithms are constructed so as to be complex conjugate to each other, for σ real, in correspondent points $\tau_1(\sigma')$, $\tau_1^*(\sigma')$ on γ_1', $\gamma_1'^*$. Thus, reversing the sense of integration on $\gamma_1'^*$ and taking the factor $1/(2\pi i)$ into account, we again obtain a real contribution to $\mathscr{K}_1(\sigma,\tau)$.

APPENDIX E: BEHAVIOR OF $C(\sigma,\tau)$ AT SOME SPECIAL POINTS

(a) Consider a cusp of a curve γ_1: $\tau = \tau_1(\sigma')$ at $\sigma' = \sigma_0$, $\tau_1(\sigma_0) = \tau_0$ and let σ in Eq. (3.4) have $|\sigma| < 1$. The angle α of the cusp is measured clockwise on the left of γ_1, with respect to the sense described in Sec. II. In general, there might be several curves γ_l, $l = 2, 3, \ldots, k$ meeting at τ_0 and having there a cusp. Consider a domain D around $\tau_0, D \subset (|\tau| < 1)$ and such that it intersects no other curves γ_1 but those meeting at τ_0, and contains no other cusps of these latter. Every curve γ_l, $l = 1, 2, \ldots, k$ divides D into two parts D_{l+}, D_{l-}, D_{l+} lying to the left of γ_1. Plemelj's formulas can still be applied (see Ref. 18, p. 428) to obtain the value at τ_0 of the factor $\mathscr{k}_l(\sigma,\tau)$ corresponding to γ_l:

$$\lim_{\substack{\tau\to\tau_0\\ \tau\in D_{l+}}} \mathscr{k}_l(\sigma,\tau) = (\sigma - \sigma_l(\tau_0))^{\alpha/2\pi - 1} \times \exp\left[-\frac{1}{2\pi i} P \int_{\gamma_1} \frac{\ln(\sigma - \sigma_l(\tau'))}{\tau' - \tau} d\tau' \right], \tag{E1}$$

$$\lim_{\substack{\tau\to\tau_0\\ \tau\in D_{l-}}} \mathscr{k}_l(\sigma,\tau) = (\sigma - \sigma_l(\tau_0))^{\alpha/2\pi} \times \exp\left[-\frac{1}{2\pi i} P \int_{\gamma_l} \frac{\ln(\sigma - \sigma_l(\tau'))}{\tau' - \tau} d\tau' \right].$$

If there is just one curve γ_1 having a cusp at τ_0, use of Eq. (3.5) establishes as in Sec. III the holomorphy of $C(\sigma,\tau)$ at τ_0. In general, there are $2k$ domains $\mathscr{D}_i$, whose boundaries have τ_0 as a common point (see Appendix F for an example with $k = 2$). Let $F_i(\sigma,\tau)$ be the determinations of $\mathscr{F}(\sigma,\tau)$ in $\mathscr{D}_i$, $i = 1, \ldots, 2k$. From Sec. III, we know that the $2k$ functions $F_i(\sigma,\tau)\ \Pi_{l=1}^{k}\ \mathscr{k}_l(\sigma,\tau), i = 1, \ldots, 2k$ are the analytic continuations of each other through the curves γ_l (for $\tau \in D$, $\tau \neq \tau_0$). To prove that $C(\sigma,\tau)$ is holomorphic at τ_0, we only have to show that the $2k$ functions above are bounded there. But this is manifest for $F_i(\sigma,\tau_0)$ and for each of the $\mathscr{k}_l(\sigma.\tau)$ from (E1) (for $|\sigma| < 1$) [see Fig. 6 for a diagram in the case of $f(\sigma,\tau) = \sigma - 1 - \tau^3$].

(b) We discuss now the behavior of $C(\sigma_0,\tau)$, $|\sigma_0| = 1$ in a neighborhood of those points τ_0 of γ_1: $\tau = \tau_1(\sigma)$, for which $\tau_1(\sigma_0) = \tau_0$. For τ near τ_0, the relevant factor of $\mathscr{K}(\sigma_0,\tau)$ is defined by Eq. (3.4), with the branch of $\ln(\sigma_0 - \sigma_1(\tau'))$ chosen such that, as $\sigma_1(\tau')$ moves past σ_0, in the counterclockwise sense, Eq. (3.8) holds. We noticed in Sec. III that $C(\sigma_0,\tau)$ is one valued and holomorphic in a neighborhood of τ_0, except

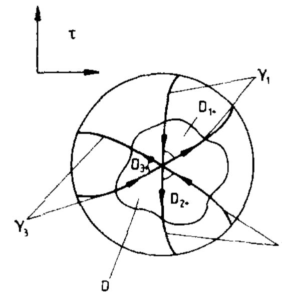


FIG. 6. To the discussion of the behavior of $C(\sigma,\tau)$ at cusps of γ_i [Appendix E, Subsection (a)].

maybe at τ_0 itself. We wish to show that in fact $\lim_{\tau\to\tau_0} C(\sigma_0,\tau) = 0$. This establishes both the holomorphy of $C(\sigma_0,\tau)$ at $\tau = \tau_0$ and the fact that $C(\sigma_0,\tau)$ vanishes precisely at those points where $f(\sigma_0,\tau)$ vanishes.

Assume first that τ_0 is a "regular" point of γ_1: this means that $d\sigma_1/d\tau(\tau = \tau_0) \neq 0$ and finite. For τ near τ_0, we write [restricting γ_1 to a neighborhood (τ_a,τ_b) of τ_0]

$$\begin{aligned}\mathscr{k}_1(\sigma_0,\tau) &= \exp\left[-\frac{1}{2\pi i} \int_{\tau_a}^{\tau_b} \left[\ln\left(\frac{\sigma_0 - \sigma_1(\tau')}{\tau_0 - \tau'} \right) \right] / (\tau' - \tau) d\tau' \right] \\ &\quad \times \exp\left[-\frac{1}{2\pi i} \int_{\tau_a}^{\tau_b} \frac{\ln(\tau_0 - \tau')}{\tau' - \tau} d\tau' \right] \\ &\equiv \mathscr{k}_{11}(\sigma_0,\tau) \mathscr{k}_{21}(\tau_0,\tau).\end{aligned} \tag{E2}$$

In $\mathscr{k}_{21}(\tau_0,\tau)$ we choose a branch of the logarithm with a cut in the τ' plane, starting at τ_0 and running along γ_1 in the sense of integration, and we let $\lim_{\tau'\to\tau_0} \Delta \arg(\tau_0 - \tau') = -\pi$ for τ' above the cut. The integration in (E2) is performed along the lower lip of the cut. The branch of the logarithm in $\mathscr{k}_{11}(\sigma_0,\tau)$ is chosen so that Eq. (E2) holds [recall that the value of $\ln(\sigma_0 - \sigma_1(\tau'))$ is already defined by the convention of Appendix D]. As in Sec. III, Eq. (3.6), one can verify that ($\mathscr{D}_2$ lies to the left of γ_1)

$$\begin{aligned}&\lim_{\substack{\tau\to\tau_0\\ \tau\in\mathscr{D}_2}} \mathscr{k}_{11}(\sigma_0,\tau) = (d\sigma_1/d\tau)_{\tau=\tau_0}^{-1/2}\, \psi(\sigma_0,\tau_0) \\ &\lim_{\substack{\tau\to\tau_0\\ \tau\in\mathscr{D}_1}} \mathscr{k}_{11}(\sigma_0,\tau) = (d\sigma_1/d\tau)_{\tau=\tau_0}^{1/2}\, \psi(\sigma_0,\tau_0)\end{aligned} \tag{E3}$$

with $\psi(\sigma_0,\tau_0)$ finite. Writing $\sigma_0 - \sigma_1(\tau) = (\sigma_0 - \sigma_1(\tau))/(\tau_0 - \tau) \times (\tau_0 - \tau) = F'(\sigma_0,\tau) F''(\tau_0,\tau)$, one verifies that the function $C_{11}(\sigma_0,\tau)$ defined as $F'(\sigma_0,\tau)\mathscr{k}_{11}(\sigma_0,\tau)$ $\tau \in \mathscr{D}_2$, and $\mathscr{k}_{11}(\sigma_0,\tau)$, $\tau \in \mathscr{D}_1$ is holomorphic at $\tau = \tau_0$. To evaluate $\mathscr{k}_{21}(\tau_0,\tau) \equiv \exp - (Q_1(\tau_0,\tau))$ we consider the following integral, performed clockwise on a path $\mathscr{C}$ consisting of the segment (τ_a,τ_b) of γ_1, however, along the upper lip of the cut for $\tau' \in (\tau_0,\tau_b)$ and an arc $\mathscr{C}_1$ ending at τ_a,τ_b which avoids γ_1 and is entirely contained in $\mathscr{D}_2$ (see Fig. 7),

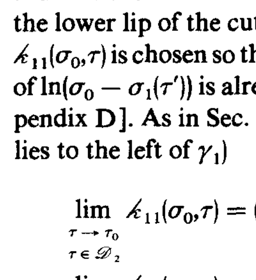


FIG. 7. To the discussion of the behavior of $C(\sigma,\tau)$ at points $(\sigma,\tau = \tau_i(\sigma)), |\sigma| = 1$ (Appendix E, Subsection (b)].

$$I(\tau_0,\tau)=\frac{1}{2\pi i}\int_{\mathscr{C}}\frac{\ln(\tau_0-\tau')}{\tau'-\tau}d\tau'$$
$$=\begin{cases}-\ln(\tau_0-\tau), & \tau\in\mathscr{D}_2\\ 0\ , & \tau\in\mathscr{D}_1\end{cases}. \tag{E4}$$

Adding $I(\tau_0,\tau)$ to $Q_1(\tau_0,\tau)$, we obtain

$$Q_1(\tau_0,\tau)=\int_{\tau_0}^{\tau_b}\frac{d\tau'}{\tau'-\tau}+\frac{1}{2\pi i}\int_{\mathscr{C}_1}\frac{\ln(\tau_0-\tau')}{\tau'-\tau}d\tau'-I(\tau_0,\tau). \tag{E5}$$

The second integral is holomorphic at $\tau=\tau_0$, and the first one has a singular contribution canceled by $I(\tau_0,\tau)$ if $\tau\in\mathscr{D}_2$. Thus the remaining factor $C_{21}(\tau_0,\tau)$ defined as $(\tau_0-\tau)\mathscr{k}_{21}(\tau_0,\tau)$ for $\tau\in\mathscr{D}_2$ and $\mathscr{k}_{21}(\tau_0,\tau)$ for $\tau\in\mathscr{D}_1$ is holomorphic at τ_0 and has there a simple zero as announced.

If τ_0 is not a regular point of γ_1, more care is required: it can in particular happen (though not necessarily) that several curves γ_l meet at such a point [if $d\sigma_1/d\tau(\tau_0)=0$], each of them having a cusp (see Appendix B). This possibility corresponds to a multiple zero of $C(\sigma_0,\tau)$ at $\tau=\tau_0$. However, a cusp of γ_1 does not necessarily occur at an "irregular" point of γ_1 (see Fig. 1). We call nevertheless α_0, the "total angle of the cusp" at τ_0 the total angular variation measured clockwise of $\tau(\sigma)$ around τ_0 when σ moves clockwise on the half circle $|\sigma-\sigma_0|=\epsilon$ contained in $|\sigma|<1$ (cf. with the curve c' in the τ plane in Fig. 1). For a regular point τ_0, the angle is π. "Irregular" points of γ_1 are characterized by $\alpha_0\neq\pi$. If τ_0 is not regular, γ_1 might still have a continuous tangent at τ_0 (see Fig. 1). The total angle α_0 is related to the power μ—a rational number—of $(\tau-\tau_0)$ needed to render $\lim(\sigma_1(\tau)-\sigma_1(\tau_0))/(\tau-\tau_0)^\mu$ finite and $\neq 0$. One can verify that $\alpha_0=\pi/\mu$.

The "geometrical" angle α of the cusp, as defined at the beginning of this Appendix and occuring in Eq. (E1) is $\alpha=\alpha_0-2k\pi$, where k is chosen so that $0<\alpha<2\pi$.

We shall show that each factor of $\mathscr{K}(\sigma,\tau)$ corresponding to a (simple) curve γ_1 passing through τ_0 produces a simple zero of $C(\sigma_0,\tau)$, so that a zero of the desired order is obtained. The proof is analogous to some extent to the one for the situation with $d\sigma_1/d\tau(\tau_0)\neq 0$ and finite.

To this end, we consider again a domain D surrounding τ_0 and contained in $|\tau|<1$, so that no other curves γ_k intersect it, apart from those passing through τ_0, and so that the latter curves have no other irregular points in it. A curve γ_1 passing through τ_0 divides it into two parts, D_{1+}, D_{1-}, D_{1+} lying to the left of γ_1. It is convenient to change variables to $(\tau_0-\tau)^\mu=v$, where $\mu=\pi/\alpha_0$. This maps the two sides of γ_1 contained in D, "before" and "beyond" τ_0, onto a smooth curve $\tilde\gamma$ in the v plane; the change of the imaginary part of $\ln v$ as one moves past $v=0$ in the sense induced by that of γ_1, is $+\pi$. The domain D_{1+} is mapped inside the half plane lying left of $\tilde\gamma$, if $\mu\geqslant\frac{1}{2}$. If $\mu<\frac{1}{2}$, this half plane contains $k+1$ images of D_{1+} and k of D_{1-}, where k is the integer above relating the geometrical and the total angle of the cusp.

We point out that, if $\mu<\frac{1}{2}$, although the multiplicity of γ_1 is unity, there are at least k roots, $\sigma'(\tau)$, $\sigma''(\tau)$, ..., of $f(\sigma,\tau)=0$, approaching σ_0 as $\tau\to\tau_0$, both from D_{1+} and D_{1-}. To make this clear, recall the discussion of the "total angle of the cusp" α_0 above and consider the continuation of a root $\sigma_1(\tau)\cong\sigma_0+C_0(\tau-\tau_0)^\mu$ clockwise away from a boundary L_1 of D_{1+} (the side of γ_1 "before" τ_0) on a circular path around τ_0, starting in D_{1+}. As we reach the other side L_2 of γ_1 ("beyond" τ_0), it is not yet true that $|\sigma_1(\tau)|\to 1$, as $\tau\to L_2$. We have to perform k more turns around τ_0 in order to reach $|\sigma_1(\tau)|=1$, as $\tau\to L_2$. We obtain this way $k+1$ different values of $\sigma_1(\tau)$ in D_{1+} and k different values in D_{1-}. All tend to σ_0 as $\tau\to\tau_0$ [cf. $f(\sigma,\tau)=\tau-(\sigma-1)^3$, Fig. 1].

Similarly to (E3) we consider now

$$Q_\mu(\tau_0,\tau)=\frac{1}{2\pi i\mu}\int_{\tilde\gamma}\frac{(\ln v)v^{1/\mu-1}}{v^{1/\mu}-(\tau_0-\tau)}dv. \tag{E6}$$

For definiteness, we choose the cut of the logarithm in the v plane to lie along $\tilde\gamma$ in the sense of integration starting at $v=0$; the integral is performed along the lower side of the cut. The cut of $v^{1/\mu}$ lies in the domain to the right of $\tilde\gamma$. As before, we add the clockwise integral along $\mathscr{C}$, a contour running along $\tilde\gamma$, but on the upper side of the logarithmic cut and closing by a semicircle to the left of $\tilde\gamma$

$$I_\mu(\tau_0,\tau)=\frac{1}{2\pi i\mu}\int_{\mathscr{C}}\frac{(\ln v)v^{1/\mu-1}}{v^{1/\mu}-(\tau_0-\tau)}dv=\begin{cases}-\sum_{n=1}^{k+1}[\ln(\tau_0-\tau)_n^\mu],\ \tau\in D_{1+}\\ -\sum_{n=1}^{k}[\ln(\tau_0-\tau)_n^\mu],\ \tau\in D_{1-}\\ 0,\ \text{if}\ k=0\end{cases} \tag{E7}$$

where the lower index on $(\tau_0-\tau)$ denotes different determinations enclosed by $\mathscr{C}$ (lying in the half plane left of $\tilde\gamma$). Adding (E6) and (E7) and using the fact that the result contains, as in (E5), an integral over the discontinuity of $\ln v$, we obtain the behavior of $\mathscr{k}_{21}(\tau_0,\tau)=\exp(-Q_\mu(\tau_0,\tau))$ for τ near τ_0 in D_{1+},D_{1-}

$$\exp[-Q_\mu(\tau_0,\tau)]=\begin{cases}\dfrac{\tau_0-\tau}{(\tau_0-\tau)_1^\mu\cdots(\tau_0-\tau)_{k+1}^\mu}\psi(\tau_0,\tau),\ \tau\in D_{1+}\\ \left[\dfrac{\tau_0-\tau}{(\tau_0-\tau)_1^\mu\cdots(\tau_0-\tau)_k^\mu}\psi(\tau_0,\tau)\right.\quad \tau\in D_1\\ (\tau_0-\tau)\psi(\tau_0,\tau),\ \text{if}\ k=0\end{cases} \tag{E8}$$

with $\psi(\tau_0,\tau)$ continuous at $\tau=\tau_0$, from both sides of $\tilde\gamma$.

On the other hand, consider the exponents of the contributions analogous to $k_{11}(\sigma_1,\tau)$ in Eq. (E3): the only difference is that the usual "derivative quotients" are replaced by Hölder quotients

$$R_\mu(\tau_0,\tau) = \frac{1}{2\pi i}\int_{L_1} \frac{\ln[\sigma_0 - \sigma_{1,1}(\tau')]/(\tau_0 - \tau')_1^\mu}{\tau' - \tau}\,d\tau' + \frac{1}{2\pi i}\int_{L_2} \frac{\ln[\sigma_0 - \sigma_{1,k+1}(\tau')]/(\tau_0 - \tau')_{k+1}^\mu}{\tau' - \tau}\,d\tau', \tag{E9}$$

where L_1, L_2 are the two sides of the cusp of γ_1 and the lower indices on $(\tau_0 - \tau')^\mu$ and $\sigma_1(\tau')$ indicate, e.g., that we perform k turns around τ_0 before substituting in the second term of (E9). We can now evaluate the limiting values as $\tau \to \tau_0$ of $k_{11}(\sigma_0,\tau) \equiv \exp(-R_\mu(\sigma_0,\tau))$ using the formulas (E1), which take into account the existence of a cusp of opening $\pi/\mu - 2k\pi$

$$\lim_{\tau\to\tau_0} k_{11}(\sigma_0,\tau) = \begin{cases} (C_0)^{1/2\mu - k - 1}\phi(\tau_0), & \tau \to \tau_0, \quad \tau \in D_{1+} \\ (C_0)^{1/2\mu - k}\phi(\tau_0), & \tau \to \tau_0, \quad \tau \in D_{1-}. \end{cases} \tag{E10}$$

We can now show that the factor of $C(\sigma_0,\tau)$ defined as

$$C_1(\sigma_0,\tau) = \begin{cases} \prod_{i=1}^{k+1} (\sigma_0 - \sigma_1^{(i)}(\tau)) k_{11}(\sigma_0,\tau) k_{21}(\tau_0,\tau), & \tau \in D_{1+} \\[2ex] \prod_{i=1}^{k} (\sigma_0 - \sigma_1^{(i)}(\tau)) k_{11}(\sigma_0,\tau) k_{21}(\tau_0,\tau), & \tau \in D_{1-} \end{cases} \tag{E11}$$

has indeed a simple zero at $\tau = \tau_0$. Indeed, using (E8)

$$\begin{aligned} \lim_{\substack{\tau\to\tau_0\\ \tau\in D_{1+}}} \frac{C_1(\sigma_0,\tau)}{\tau_0 - \tau} &= \lim_{\substack{\tau\to\tau_0\\ \tau\in D_{1+}}} \prod_{i=1}^{k+1} (\sigma_0 - \sigma_1^{(i)}(\tau)) C_0^{1/2\mu - k - 1} \frac{\psi(\tau_0,\tau)\phi(\tau_0)}{\Pi_{i=1}^{k+1}(\tau_0 - \tau)_i^\mu} \\ &= \psi(\tau_0,\tau_0)\phi(\tau_0) C_0^{1/2\mu} = \lim_{\substack{\tau\to\tau_0\\ \tau\in D_{1-}}} \prod_{i=1}^{k} (\sigma_0 - \sigma_1^{(i)}(\tau)) C_0^{1/2\mu - k} \frac{\psi(\tau_0,\tau)\phi(\tau_0)}{\Pi_{i=1}^{k}(\tau_0 - \tau)_i^\mu} = \lim_{\substack{\tau\to\tau_0\\ \tau\in D_{1-}}} \frac{C_1(\sigma_0,\tau)}{\tau_0 - \tau} = \text{finite}. \end{aligned} \tag{E12}$$

By the reasoning of Sec. III, $C_1(\sigma_0,\tau)$ is holomorphic in the neighborhood of τ_0, so that (E12) establishes that it is holomorphic even at τ_0 and has there a simple zero.

We can do this reasoning for each curve γ_1 passing through τ_0, such that $\lim_{\tau\to\tau_0}\sigma_1(\tau) = \sigma_0$ and having there a regular or irregular point. This might not be immediately obvious, since one might worry that the zeros $\sigma_1^{(i)}(\tau)$ in (E11) could be required in several factors $C_l(\sigma_0,\tau)$, similar to (E11) and associated to the γ_l's, $l \neq 1$. It is, however, true that the zeros generated by clockwise continuation around τ_0 of the inverse function of γ_1, $\sigma_1(\tau)$ (starting from the side of γ_1 "before" τ_0), are different from those generated by the other curves. One can verify that, in fact, in order to obtain the roots "belonging" to the other curves, one must continue $\sigma_1(\tau)$ on a path which rotates so long around τ_0, until $\sigma_1(\tau) - \sigma_0$ rotates by $2m\pi$ around the origin ($m = 1, 2, \ldots$) [cf. the example of $f(\sigma,\tau) = \tau^2 - (\sigma - 1)^9$].

Thus each γ_l generates a simple zero of $C_l(\sigma_0,\tau)$ at τ_0 and $C(\sigma_0,\tau)$ vanishes at τ_0 with an order equal to the number of γ_l's passing through τ_0 and verifying $\lim_{\tau\to\tau_l}\sigma_l(\tau) = \sigma_0$.

(c) We next wish to prove the continuity of $C(\sigma,\tau)$ at σ_0,τ_0, $|\sigma_0| = 1$, $\tau_0 = \tau_i(\sigma_0)$, for some i, (say $i = 1$) and $|\tau_0| < 1$, as (σ,τ) approaches (σ_0,τ_0) from $|\sigma| < 1$, $|\tau| < 1$. The reason why we need this is to show that all values $C(\sigma,\tau)$ for $|\sigma| = 1$, $|\tau| < 1$ are properly defined, independently of the special procedure we adopted. We must thus evaluate

$$\begin{aligned} |C(\sigma,\tau) - C(\sigma_0,\tau_0)| &\leqslant |C(\sigma,\tau) - C(\sigma_0,\tau)| \\ &\quad + |C(\sigma_0,\tau) - C(\sigma_0,\tau_0)|. \end{aligned} \tag{E13}$$

The last term can be made as small as we wish, by choosing τ correspondingly close to τ_0, since $C(\sigma_0,\tau)$ is holomorphic at τ_0. We need to show only that the first term can be made arbitrarily small, if $|\sigma - \sigma_0|$ is chosen appropriately, but uniformly with respect to τ, for τ in some disk $|\tau - \tau_0| \leqslant r_1$, r_1 a fixed number. To this end, it is enough to prove that the family of functions $C(\sigma_n,\tau)$, for a sequence $\sigma_n \to \sigma_0$, holomorphic with respect to τ in the unit disk, is uniformly bounded with respect to n, for τ in a disk $|\tau - \tau_0| \leqslant r$. Since, by definition, for τ not on γ_i, $\lim_{n\to\infty} C(\sigma_n,\tau) = C(\sigma_0,\tau)$, it follows from the principle of uniform boundedness (Vitali's theorem), that this limit is also uniform with respect to τ, in $|\tau - \tau_0| \leqslant r - \epsilon$, for any $\epsilon > 0$, and this will prove our statement. By the maximum modulus theorem, it is enough to prove the uniform boundedness on the circle $|\tau - \tau_0| = r$, with r appropriately chosen.

Further, it is enough to consider one factor $k_1(\sigma,\tau)$ belonging to one curve, say γ_1, passing through τ_0, and obeying $\lim_{\tau\to\tau_0} \sigma_1(\tau) = \sigma_0$, and the purpose is to find a majorization of the absolute value of the exponent of $k_1(\sigma,\tau)$, uniform in τ, for $|\sigma - \sigma_0| \leqslant d$, $|\sigma| < 1$, d appropriately chosen. This is conveniently done as follows: choose r so small that γ_1 intersects $|\tau - \tau_0| = r$ only twice and part the integral into two [for τ on $|\tau - \tau_0| = r$], one part contained inside a circle of radius kr, $0 < k < 1$ and another one outside it. Choose then d so small that the logarithmic singularities of the integrand for σ, $|\sigma| = 1$, in $|\sigma - \sigma_0| \leqslant d$ are contained in a circle of radius, say, $kr/2$. This is possible due to the Hölder continuity of both $\tau_1(\sigma)$, defining γ_1, and its inverse $\sigma_1(\tau)$. Then, the part of the integral contained in the circle of radius kr gives a bounded contribution on $|\tau - \tau_0| = r$, uniformly in σ. The remaining part is the integral of a Hölder continuous function with the Cauchy kernel $1/(\tau' - \tau)$ and is clearly bounded, e.g., by using the estimates of Ref. 18, p. 38.

(d) We now show that, if $\tau_0 = \tau_1(\sigma_0)$, $|\sigma_0| = 1$, $\tau_0 \in \gamma_1$, then for $|\sigma| < 1$ near σ_0

$$|I| = \Big| \int_{\tau_a}^{\tau_e} \frac{\ln(\sigma - \sigma_1(\tau'))}{\tau' - \tau_0} d\tau' \Big| \leqslant C \ln^2 |\sigma - \sigma_0|, \tag{E14}$$

where the integral is understood as a limit $\tau \to \tau_0$ from $\mathscr{D}_2$ or $\mathscr{D}_1$ (domains lying left or right of γ_1), τ_a, $\tau_e \in \gamma_1$, $\tau_0 \in (\tau_a, \tau_e)$ and C is a constant. Equation (E14) establishes a uniform bound on the integrals of $|\ln C(re^{i\theta}, \tau)|$ on $0 \leqslant \theta \leqslant 2\pi$, as required at the end of Sec. III. To show this, we write $(\sigma = re^{i\theta},\ r < 1,\ \tau_a = \tau_1(\sigma_a),\ \tau_e = \tau_1(\sigma_e))$:

$$|I| \leqslant I_1 + I_2 = \int_{\sigma_a}^{\sigma_e} \frac{|\ln(\sigma - \sigma') - \ln(\sigma - \sigma_0)|}{|\sigma' - \sigma_0|} \times \frac{|\sigma' - \sigma_0|}{|\tau_1(\sigma') - \tau_0|} \frac{|d\tau_1(\sigma')|}{|d\sigma'|} |d\sigma'| + |\ln(\sigma - \sigma_0)| \Big| \int_{\tau_a}^{\tau_e} \frac{d\tau'}{\tau' - \tau_0} \Big|. \tag{E15}$$

The second integral satisfies already the bound (E14); in the first one, the last two factors make up a continuous function on (σ_a, σ_e) and are thus bounded by a constant there, so that one is left with (for $k_1 > 1$)

$$I_1 \leqslant C \Big(\int_{|u| < k_1 |\sigma - \sigma_0|} + \int_{|u| > k_1 |\sigma - \sigma_0|} \Big) \times \frac{|\ln(1 - u/(\sigma - \sigma_0))|}{|u|} |du| \leqslant C_1 + C_2 \int_{k_1}^{b} \frac{\ln|v|}{|v|} d\,|v| \leqslant C \ln^2 |\sigma - \sigma_0| \tag{E16}$$

with $b = \max(|\sigma_{a1} - \sigma_0|/|\sigma - \sigma_0|,\ |\sigma_l - \sigma_0|/|\sigma - \sigma_0|)$ and we have used $v = u/(\sigma - \sigma_0)$, $|du| \leqslant \text{const}\, d\,|u|$ if the arc around σ_0 is sufficiently small, $|\ln(1 - v)| \leqslant C \ln|v|$, for $|v| > k_1$ and C, C_1, C_2 are constants.

(e) We discuss shortly the behavior of $C(\sigma, \tau)$ at points (σ_i, τ_0), $|\sigma_i| < 1$, τ_0 an end point of a curve $\gamma_1: \tau = \tau_1(\sigma)$. Following Ref. 18, p. 74, the behavior of the corresponding factor $\mathscr{k}_1(\sigma,\tau)$, Eq. (3.4), near τ_0 is given by

$$\mathscr{k}_1(\sigma,\tau) = \exp\Big[-\frac{1}{2\pi i} \ln(\sigma - \sigma_1(\tau_0)) \ln(\tau_0 - \tau) - \phi(\sigma,\tau) \Big] \equiv \phi_1(\sigma,\tau) \exp(-\phi(\sigma,\tau)), \tag{E17}$$

where $\phi(\sigma,\tau)$ is a function holomorphic in the τ plane cut along γ_1 and continuous at $\tau = \tau_0$. Clearly, as $\tau \to \tau_0$, the phase of $\phi_1(\sigma,\tau)$ increases without bounds. However, $\ln \mathscr{k}_1(\sigma,\tau)$ has obviously an integrable modulus at τ_0, for fixed σ, $|\sigma| < 1$.

APPENDIX F: TWO EXAMPLES TO SEC. III

We wish to illustrate the formulas in Secs. II and III by means of two examples; the first one is very simple, and could have been treated as shown in the Introduction [Eqs. (1.8)–(1.11)]. The second one would have been presumably

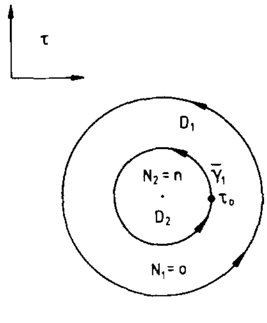


FIG. 8. Example 1 of Appendix F.

difficult to treat without the Cousin method.

Example 1: $f(\sigma,\tau) = \sigma^n/2 - \tau$, $n \geqslant 1$, integer. For all σ on $|\sigma| = 1$, there is one root in $|\tau| \leqslant 1$ and for $|\tau| = 1$, there is no root in $|\sigma| \leqslant 1$. We can apply the method of the Introduction, and obtain immediately that $\beta_1(\sigma) = -\sigma^n/2$ [Eq. (1.9)]. Thus, $C_0(\sigma,\tau)$, Eq. (1.10), is $\tau - \sigma^n/2$, and $E(\sigma,\tau) = +1$, Eq. (1.11). The function $f(\sigma,\tau)$ is reconstructed from its modulus up to a sign ambiguity.

We now treat the same example by the "complicated" way of Secs. II and III. There exists in the τ plane one curve γ_1: $\tau = \sigma^n/2$, $|\sigma| = 1$; this curve is the circle $\bar{\gamma}_1: |\tau| = \frac{1}{2}$ covered n times. The inverse function $\sigma_1(\tau') = (2\tau')^{1/n}$ is many valued on $\bar{\gamma}_1$. It should be clear from the text that the integrations in Eqs. (2.4) and (2.10) are performed over $\bar{\gamma}_1$ covered n times. The determinations of $\mathscr{N}(\tau)$, Eq. (2.4), in the two domains $\mathscr{D}_1$, $\mathscr{D}_2$ are shown in Fig. 8. To compute $\Omega^{(1)}(\tau)$, we write

$$\Omega^{(1)}(\tau) = \frac{1}{2\pi i} \int_{\gamma_1} \frac{(2\tau')^{1/n}}{\tau' - \tau} d\tau' = \frac{1}{2\pi i} \sum_{k=1}^{n} \oint_{|\tau'| = \frac{1}{2}} \frac{(2\tau')_k^{1/n}}{\tau' - \tau} d\tau', \tag{F1}$$

where the sum runs over the n determinations of $(2\tau')^{1/n}$. The integral is converted to the unit circle in the σ' plane, by $\sigma' = (2\tau')^{1/n}$.

$$\Omega^{(1)}(\tau) = \frac{n}{2\pi i} \oint_{|\sigma'| = 1} \frac{\sigma'^n d\sigma'}{\sigma'^n - 2\tau} = \begin{cases} 0, & \text{if } \tau \in \mathscr{D}_1 \\ \sum_{k=1}^{n} (2\tau)_k^{1/n} = 0, & \text{if } \tau \in \mathscr{D}_2 \end{cases}. \tag{F2}$$

Similarly,

$$\Omega^{(l)}(\tau) = \begin{cases} 0 & \text{if } \tau \in \mathscr{D}_1 \\ \sum_{k=1}^{n} (2\tau)_k^{l/n} & \text{if } \tau \in \mathscr{D}_2 \end{cases} \quad l = 2, 3, \ldots. \tag{F3}$$

But the sum occurring for $\tau \in \mathscr{D}_2$ vanishes if $l < n$ and is $2\tau n$ if $l = n$. The Newton formulas, Eq. (2.18) give then

$$\mathscr{F}(\sigma,\tau) = \begin{cases} 1 & \text{if } \tau \in \mathscr{D}_1 \\ \sigma^n - 2\tau & \text{if } \tau \in \mathscr{D}_2 \end{cases}. \tag{F4}$$

With the same change of variables as above, [cf. Eq. (3.1)]

$$\mathscr{K}_1(\sigma,\tau) = \exp\Big[-\frac{1}{2\pi i} \int_{\gamma_1} \frac{\ln(\sigma - (2\tau')^{1/n})}{\tau' - \tau} d\tau' \Big] = \exp\Big[-\frac{n}{2\pi i} \oint_{|\sigma'| = 1} \sigma'^{n-1} \frac{\ln(\sigma - \sigma')}{\sigma' - 2\tau} d\sigma' \Big]. \tag{F5}$$

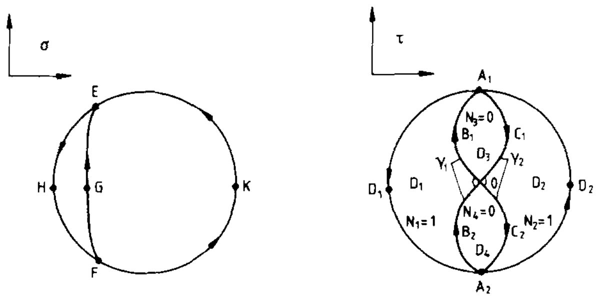


FIG. 9. Example 2 of Appendix F.

The logarithm is chosen as explained in Appendix D: for σ real, this amounts to choosing a cut of the logarithm in the σ' plane from σ to the right, $\operatorname{Im}\ln(\sigma-\sigma') = -\pi i$, if σ' is above the cut ($\tau_a = (-1)^n/2$, $\tau_0 = \frac{1}{2}$). There exists a discontinuity of $2\pi i$ at $\tau = \tau_0$, producing a pole in $\mathscr{K}_1(\sigma,\tau)$. We define thus

$$\mathscr{K}(\sigma,\tau) = (\tau - \tfrac{1}{2})\mathscr{K}_1(\sigma,\tau). \tag{F6}$$

The function $\mathscr{K}_1(\sigma,\tau)$ can be in this case explicitly computed by picking out the contributions of the cut of the logarithm and the possible pole; we obtain

$$\mathscr{K}_1(\sigma,\tau) = \begin{cases} \dfrac{\sigma^n - 2\tau}{1-2\tau} & \text{if } |\tau| > \frac{1}{2} \quad (\tau \in \mathscr{D}_1) \\[2ex] \dfrac{1}{1-2\tau} & \text{if } |\tau| < \frac{1}{2} \quad (\tau \in \mathscr{D}_2) \end{cases}. \tag{F7}$$

Thus,

$$C(\sigma,\tau) = \mathscr{F}(\sigma,\tau)\mathscr{K}(\sigma,\tau) = -(\sigma^n - 2\tau). \tag{F8}$$

One checks easily that $E(\sigma,\tau) = \frac{1}{2}$.

Example 2: $f(\sigma,\tau) = \sigma^2\tau^2 - \sigma - 1$. The zero trajectories in the σ and τ planes are shown in Fig. 9. For $|\sigma| = 1$, there are two zeros in $|\tau| \leqslant 1$ only along the arc EHF. For σ on the arc FKE, there are no zeros in $|\tau| \leqslant 1$. For σ on EHF, the two zeros describe the curves $A_2B_2OB_1A_1$ (γ_1) and $A_1C_1OC_2A_2(\gamma_2)$. There is a branch point of the trajectories $\tau(\sigma)$ lying at $\sigma = -1$. Thus, the curves γ_1, γ_2 have a cusp at $\tau(-1) = 0$, and they are connected as described in Sec. II. Along the semicircles $A_2D_2A_1$ and $A_1D_1A_2$, the zeros of $f(\sigma,\tau)$ are given by

$$\sigma_k(\tau) = \frac{1 - (\sqrt{1+4\tau^2})_k}{2\tau^2}, \quad k = 1,2, \tag{F9}$$

where the index 1 refers to $A_2D_2A_1$ and 2 to the other semicircle. For τ on $A_2D_2A_1$, the root is chosen to be positive at $+1$, and for τ on $A_1D_1A_2$ the root is chosen positive at -1. The images $\sigma_{1,2}(\tau)$ describe for $|\tau| = 1$, the curve FGE, once for τ on $A_2D_2A_1$ and once for τ on $A_1D_1A_2$. Thus (reversing the roles of σ and τ in the text), the curve FGE has multiplicity two. There is nothing special about the points $A_1,A_2(\gamma_1,\gamma_2$ cross there). The path Γ_1 needed to return $\sigma_1(\tau)$ to its original value is $A_2D_2A_1C_1OC_2A_2$. Indeed, as we continue $\sigma_1(\tau)$ from D_2 to A_1 we reach the value $\sigma_0 = -\frac{1}{2} + i(\frac{3}{4})^{1/2}$ and only $\gamma_2{:}\tau = -(\sigma+1)^{1/2}/\sigma, |\sigma| = 1$ has the property $\lim_{\sigma\to\sigma_0}\tau(\sigma) = +i$. The path Γ_2 is $A_1D_1A_2B_2OB_1A_1$. It is easy to verify that the assignments of values to $\mathscr{N}(\tau)$, Eq. (2.4) as done in Fig. 9 are correct.

By writing Cauchy integrals, one verifies that

$$\Omega^{(1)}(\tau) = \begin{cases} 0, & \text{for } \tau \text{ in } \mathscr{D}_3,\mathscr{D}_4, \\ (1 - (\sqrt{1+4\tau^2})_k)/2\tau^2, & \\ & \text{for } \tau \text{ in } \mathscr{D}_k, k = 1,2. \end{cases} \tag{F10}$$

The function $\mathscr{K}_1(\sigma,\tau)$, eq. (3.1) is given by (after using on $\gamma_{1,2}$ the parameter $\sigma' \in EHF$)

$$\mathscr{K}_1(\sigma,\tau) = \exp\left[-\frac{1}{2\pi i}\int_{EHF} \frac{\ln(\sigma-\sigma')(\sigma'+2)}{\sigma'(\sigma'+1-\tau^2\sigma'^2)}\,d\sigma'\right]. \tag{F11}$$

This expression may appear obscure at first sight. We illuminate it as follows: let $\operatorname{Im}\sigma = 0$, $\sigma > 0$ and enlarge the integral to one along the line $KEHFK$; the integral is not done on a closed path, because of the cut of the logarithm running from σ to the right [we choose, as in Appendix D, $\operatorname{Im}\ln(\sigma+1) = 0$]. We must subtract two contributions: the integrals along KE and FK. However, it is important to notice that the denominator of the integrands does not vanish in $|\tau| \leqslant 1$, if σ' is confined to EKF. Thus, the corresponding factor in $\mathscr{K}_1(\sigma,\tau)$ is a function holomorphic and free of zeros in $|\sigma| < 1 \otimes |\tau| < 1$ and which we denote by $C_E(\sigma,\tau)$. The integral along the path $KEHFK$ is evaluated as in Example 1, by writing it as a sum of the contributions of the cut of the logarithm and the poles in $|\sigma'| < 1$.

After a careful calculation, one obtains

$$\begin{aligned}\mathscr{K}(\sigma,\tau) &= \mathscr{K}_1(\sigma,\tau) \\ &= \begin{cases} C_E(\sigma,\tau)(\tau^2\sigma^2 - \sigma - 1)/(\tau^2 - 2), \ \tau \in \mathscr{D}_{3,4} \\ C_E(\sigma,\tau)(2\tau^2\sigma - 1 - (\sqrt{1+4\tau^2})_k)/2(\tau^2 - 2)' \end{cases} \\ &\quad \tau \in \mathscr{D}_k, \\ &\quad k = 1,2.\end{aligned} \tag{F12}$$

The Cousin function is then the manifestly holomorphic expression

$$\begin{aligned} C(\sigma,\tau) &= C_E(\sigma,\tau)(\tau^2\sigma^2 - \sigma - 1)/(\tau^2 - 2) \\ &= \overline{C}_E(\sigma,\tau)(\tau^2\sigma^2 - \sigma - 1). \end{aligned} \tag{F13}$$

It is easy to verify that

$$E(\sigma,\tau) = -1/\overline{C}_E(\sigma,\tau). \tag{F14}$$

It is instructive, although not very easy, to perform a similar calculation starting from the pattern of curves in the σ plane.

APPENDIX G: THE PROOF OF LEMMA 4

To prove Lemma 4 of Sec. IV, we notice first that if, for another sequence s_n', $\lim_{n\to\infty} t(s_n') = t_1 \neq t_0$, then $t_1 \in C_t$. Indeed, if t_1 belonged to Q_s, it would follow by continuity that $A(s_0,t_1) = 0$ and t_1 would be a simple or multiple zero of $A(s_0,t)$. Let d_r be a disk $|t - t_1| < r$, so that $A(s_0,t) \neq 0$, for $t \in \overline{d}_r$, if $t \neq t_1$. Consider then the function $h(s) = |t(s) - t_1|$; by Lemma 1 (or 3) it is continuous on (s_1,s_0), $s_1 < s_0$. Since, for $s_n \to s_0$, $\lim_{n\to\infty} h(s_n) = |t_1 - t_0| = a \neq 0$, it follows that $h(s)$ assumes on (s_1,s_0) all values between ϵ and $a - \epsilon$, for any $\epsilon > 0$, an infinite number of times. This implies that a sequence $s_n'' \to s_0$ exists so that $\lim|t(s_n'') - t_1| = r_1 < r$. This means that $A(s_0,t) = 0$ for t in $d_r, t \neq t_0$, which is impossible. Thus, $t_1 \in C_t$.

Consider then the continuous function $l(s) = \operatorname{Re}(t(s) - t_1)$. As before, $l(s)$ assumes any value on (t_0,t_1) as $s \to s_0$. By the continuity of $A(s,t)$ in $\overline{\mathscr{M}}$, one concludes that $A(s_0,t)=0$, for t in $(t_0,t_1) \subset C_t$. Thus, if $t_0 \neq t_1, A(s_0,t) \equiv 0$, a situation which we can discard.

APPENDIX H: BEHAVIOR OF C(s,t) AT SOME SPECIAL POINTS (SEC. V)

(i) Consider first a point $t_0, t_0 < -4m^2$, where a curve $\Gamma_1: t = t_1(s)$ joins a curve $\Gamma_2': t = t_2(u)$. In general, there is a cusp between Γ_1 and Γ_2' at t_0 (see Fig. 4); this follows from the fact that the line $s = s_2(u)$, $u \in C_u$, up to which the continuation of the s channel trajectory $u = u_1(s)$ can proceed (see Lemma 6) intersects in general the line $\operatorname{Im} s = 0$ under an angle different from 0 or π, and the mapping $t(s) = 4m^2 - s - u_2(s)$ might have a nonvanishing derivative at $s = s_0$. Further, t_0 is also a joining point of the curves $\Gamma_1^*: t = t_1^*(s^*)$, $\Gamma_2^{*\prime}: t = t_2^*(u^*)$ (see Fig. 4).

As we argued in Sec. IV, the analytic continuation around t_0 of the inverse function $s_1(t)$ corresponding to Γ_1 is $\bar{s}_2(t)$, corresponding to Γ_2'. The continuation of $s_1(t)$ to points $t, t \notin \Gamma_1$, $t \notin \Gamma_2'$, leads to a root of $A(s,t) = 0$ lying in Q_t and is accounted for by one[40] factor $s - s_1(t)$ in the determination of $\mathscr{F}_s(s,t)$ in the domain D_+ lying left of Γ_1, Γ_2'. We can then evaluate using (E1)

$$\lim_{\substack{t \to t_0 \\ t \in D_+}} (s - s_1(t))\bar{k}_1(s,t) = (s - s_1(t_0))^{\beta/2\pi} \times \exp\left(-\frac{1}{2\pi i} P \int_{\Gamma_1 \cup \Gamma_2'} \frac{\varphi(s,t_0,t')}{t' - t_0} dt'\right) = \lim_{\substack{t \to t_0 \\ t \in D_-}} \bar{k}_1(s,t), \tag{H1}$$

where $\bar{k}_1(s,t)$ is the factor of $K(s,t)$ corresponding to Γ_1 and Γ_2', β is the opening of the cusp and $\varphi(s,t_0,t') = \ln(s - s_1(t'))$ for $t' \in \Gamma_1$, $= \ln(s + t_0 - \bar{s}_2(t') - t'$ for $t' \in \Gamma_2'$; $\varphi(s,t_0,t')$ is Hölder continuous at $t' = t_0$, for $s \in D_s$ [discontinuities of the argument of the logarithm are removed by factors $(t - t_{oi})$ in (5.5)]. A similar argument holds for $\Gamma_1^*, \Gamma_2'^*$; this proves analyticity of $C(s,t)$ at fixed s at t_0. Equation (H1) is a holomorphic expression in s in the open cut plane Q_{t_0}, and thus $C(s,t)$ is holomorphic in s and t at all points (s,t_0) in $\mathscr{M}$.

(ii) Consider now the case when t_0 is an end point of $\Gamma_i: t = t_i(s)$ such that $t_i(s) \to t_0$ as $s \to \infty$, e.g., along C_s^+; then, by (A5) (i), t_0 is also an end point of $\Gamma_j'^{(*)}: t = t_j(u)$, $u \in C_u^-$, $t \to t_0$ as $u \to \infty$. We have seen in sec. IV that there is a cusp of $\Gamma_i \cup \Gamma_j'$ at t_0 with opening $\pi\alpha - 2k\pi$, $\alpha > 0$, with $k \geqslant 0$ chosen so that $0 < \pi\alpha - 2k\pi < 2\pi$. If $k \geqslant 1$, it means that, as s moves on a large circle an angle π away from C_s^+, $(t_i(s) - t_0)$ performs k complete rotations around the origin. Consequently, as in Appendix E, there are $k+1$ roots $s_n(t)$ of $A(s,t) = 0$ for t in D_+ and k such roots for t in D_-, with the property that $|s_n(t)| \to \infty$ as $t \to t_0$ (D_+, D_- are the domains lying to the left and right of $\Gamma_i \cup \Gamma_j'$). By (A5) (i), they have the property that $s_n(t)\ (t - t_0)_n^{1/\alpha} \to$ const, as $t \to t_0$; the lower index indicates the branch (obtained after n rotations around t_0).

We shall show that the contribution to $C(s,t)$ of the pair of curves Γ_i, Γ_j' and of the zeros $s_i(t)$, $s_j'(t)$ generated by them is a holomorphic function of t at t_0 for s fixed, if one takes into account a factor $(t - t_0)^2$ as shown in Eq.(5.5).

To this end, we write for $\Gamma_i \cup \Gamma_j'^{(*)}$, using (A5) (i)

$$I(s,t) = -\frac{1}{2\pi i} \int_{\Gamma_i} \frac{\ln(s - C_0(t')/(t' - t_0)_1^{1/\alpha})}{t' - t} \frac{w(t')}{w(t)} dt' - \frac{1}{2\pi i} \int_{\Gamma_j'^{(*)}} \frac{\ln'(s + t - t' - C_0(t')/(t' - t_0)_{k+1}^{1/\alpha})}{t' - t} \times \frac{w(t')}{w(t)} dt' + \int_{\Gamma_j'^{(*)}} \frac{w(t')}{w(t)} \frac{dt'}{t' - t}. \tag{H2}$$

In (H2), we have understood the imaginary part of the logarithm in the integral along $\Gamma_j'^{(*)}$ to be 2π rather than zero, as it should be according to the definition of $K_{ou}(u,t)$. We corrected this by the third term, which diverges logarithmically for $t \to t_0$. The choice of the logarithm in (H2) can be understood from the fact that the argument of $s - s_i(t')$ increases by π as $s' = s_i(t')$ moves from C_s to C_u. Multiplying by $(t' - t_0)_{1,k+1}^{1/\alpha}$ in the logarithms in (H2), we can separate a contribution $I_1(s,t)$ with a Hölder continuous integrand at $t = t_0$, and evaluate it by (E1), Appendix E and an s independent term

$$I_2(t,t_0) = \frac{1}{2\pi i} \int_{\Gamma_i} \frac{\ln(t' - t_0)_1^{1/\alpha}}{t' - t} \frac{w(t')}{w(t)} dt' + \frac{1}{2\pi i} \int_{\Gamma_j'^{(*)}} \frac{\ln(t' - t_0)_{k+1}^{1/\alpha}}{t' - t} \frac{w(t')}{w(t)} dt'. \tag{H3}$$

This integral, analogous to (E6) can be evaluated as in (E8):

$$\exp[I_2(t,t_0)] = \begin{cases} \left(\prod_{n=1}^{k+1} (t - t_0)_n^{1/\alpha}\right) & \psi(t,t_0)/(t - t_0),\ t \in D_+ \\ \left(\prod_{n=1}^{k} (t - t_0)_n^{1/\alpha}\right) & \psi(t,t_0)/(t - t_0),\ t \in D_-,\ k \neq 0 \\ & \psi(t,t_0)/(t - t_0),\ t \in D_-,\ k = 0 \end{cases} \tag{H4}$$

and $\psi(t,t_0)$ is continuous from all sides as $t \to t_0$. We can now estimate the relevant factor $\overline{C}(s,t)$ of $C(s,t)$ in the neighborhood of these points:

$$\lim_{\substack{t \to t_0 \\ t \in D_+}} \overline{C}(s,t) = \lim_{\substack{t \to t_0 \\ t \in D_+}} (t - t_0)^2 \prod_{n=1}^{k+1} (s - C_0(t) \div (t - t_0)_n^{1/\alpha}) \times \frac{\Pi_{n=1}^{k+1} (t - t_0)_n^{1/\alpha}}{t - t_0} \times \frac{k(t)}{t - t_0} (-C_0(t_0))^{\alpha/2 - k - 1} \psi(t,t_0) \Phi(s,t_0)) = (-C_0(t_0))^{\alpha/2} k(t_0) \psi(t_0,t_0) \Phi(s,t_0) = \lim_{\substack{t \to t_0 \\ t \in D_-}} \overline{C}(s,t). \tag{H5}$$

We have used the factor $(t - t_0)^2$ present in (5.5), $\Phi(s,t_0)$ corresponds to the exponential factor in (E1) and $k(t)/(t - t_0)$ is the exponential of the third term in (H2). Thus, $\overline{C}(s,t)$ is finite at t_0 and, since it is one valued and holomorphic in a neighborhood of t_0, it is holomorphic even at t_0. The value of $\overline{C}(s,t)$ at t_0 is holomorphic in $s, s \in D_s$, and thus, $C(s,t)$ is holomorphic at the points $(s,t_0) \in \mathscr{M}$, as asserted.

APPENDIX I: COMPLEMENT TO SEC. V

We wish to refine the estimate (5.25) of $I(s,t)$ so that it stays valid even if t lies on one of the curves Γ_i. Consider to this end, for $t \in \Gamma_i$, a disk D_r of radius r centered at t, which contains an interval (t_a, t_b) of Γ_i. The bounds for the integrals on Γ_i on regions outside D_r are the same as in Sec. V. We establish first an inequality like (5.25) for $|s|$ high enough, $s \in D_s \cup C_s^+ \cup C_s^-$ and thus verify condition (5.23). The behavior of $I(s,t)$ for finite $s \in C_s$ is discussed at the end.

Let (t_{0a}, t_{0b}) be an interval of Γ_i containing (t_a, t_b) and assume first that $s_i(t')$ stays finite inside (t_{0a}, t_{0b}). We shall place an upper bound on

$$|\widetilde{I}(s,t)| = \left| \frac{1}{2\pi i} \int_{t_{0a}}^{t_{0b}} \frac{\ln(s - s_i(t'))}{t' - t} \frac{w(t')}{w(t)} dt' \right|. \tag{I1}$$

To this end, let $\psi(t)$ be a real positive differentiable function defined along Γ_i, with compact support contained in (t_{0a}, t_{0b}) equal to unity on (t_a, t_b). We write then

$$\begin{aligned} \widetilde{I}(s,t) &= \frac{1}{2\pi i} \int_{t_{0a}}^{t_{0b}} \frac{\ln(s - s_i(t'))}{t' - t} \psi(t') \frac{w(t')}{w(t)} dt' \\ &\quad + \frac{1}{2\pi i} \int_{t_{0a}}^{t_{0b}} \frac{\ln(s - s_i(t'))(1 - \psi(t'))}{t' - t} \frac{w(t')}{w(t)} dt' \\ &= I_\psi(s,t) + I_{1\psi}(s,t). \end{aligned} \tag{I2}$$

The integrand of $I_{1\psi}(s,t)$ has support outside (t_a, t_b) and can thus be bounded by $K(t)\ln|s|$, as shown in Sec. V. It is then enough to consider $I_\psi(s,t)$ which has, for $|s|$ high enough, Hölder continuous limiting values for $t \in (t_{0a}, t_{0b})$. We notice that the function $\Delta(s,t,t') \equiv \ln(s - s_i(t)) - \ln(s - s_i(t'))$ obeys for $t, t' \in (t_{0a}, t_{0b})$:

$$\begin{aligned} |\Delta(s,t,t')| &\leqslant |s_i(t) - s_i(t')| \frac{1}{|s| - \max_{t_{0a} < t < t_{0b}} |s_i(t)|} \\ &\leqslant \frac{K_1}{|s|} |t - t'|^\beta, \end{aligned} \tag{I3}$$

where β is the Hölder index of $s_i(t)$ and K_1 a constant (we have used $|\ln(1 - x)/(1 - y))| \leqslant |x - y|/(1 - \max(|x|, |y|))$ for $|x|, |y| < 1$). Then,

$$\begin{aligned} |I_\psi(s,t)| &\leqslant \frac{1}{2\pi} \int_{t_{0a}}^{t_{0b}} \frac{|\Delta(s,t,t')| |\psi(t')| |w(t')|}{|t' - t| \, |w(t)|} |dt'| \\ &\quad + \frac{1}{2\pi} \frac{|\ln(s - s_i(t))|}{|w(t)|} \left| P \int_{t_{0a}}^{t_{0b}} \frac{\psi(t') w(t')}{t' - t} dt' \right| \\ &\quad + \frac{1}{2} |\ln(s - s_i(t))| |\psi(t)| \leqslant K \frac{\ln|s|}{|w(t)|}. \end{aligned} \tag{I4}$$

In (I4) we have used (I3) and the fact that the principal-value integral is finite for all $t \in (t_{0a}, t_{0b})$.

Consider now the situation when $s_i(t)$ gets infinite on $\Gamma_i, |s_i(t)| \to \infty$ as $t \to t_0$. The integral whose modulus must be bounded is (H2). We separate out of it (H3) which is s independent, as well as the last term in (H2). The remainder is further parted into two contributions, from Γ_i and Γ_i'. It is sufficient to consider the one from Γ_i; we use again a weight function $\psi(t)$, equal to unity on (t_a, t_0) and to zero outside (t_{0a}, t_0) and let $w_1(t) = w(t)\psi(t)$. After separating an s independent bounded term, we are left with the task of bounding

$$|I_0(s)| = \left| \frac{1}{2\pi i w(t)} \int_{\substack{t_{0a} \\ t \in \Gamma_i}}^{t_0} \frac{w_1(t') \ln(1 - s/s_i(t'))}{t' - t_0} dt' \right| \tag{I5}$$

and a similar integral on Γ_i'. Using (A5) (ii), we transfer the integral to s' space:

$$\begin{aligned} |I_0(s)| &\leqslant \frac{1}{2\pi |w(t)|} \left| \int_{s_i(t_{0a})}^{|s|/2} + \int_{|s|/2}^{2|s|} + \int_{2|s|}^{\infty} w_1(t(s')) \right. \\ &\quad \left. \times \ln(1 - s/s') F(s', t_0) ds' \right| = |I_{01}(s) + I_{02}(s) + I_{03}(s)| \end{aligned} \tag{I6}$$

with

$$F(s', t_0) = (dt_i/ds')/(t_i(s') - t_0) = 1/s' + 0(1/s'^{1+\epsilon}) \tag{I7}$$

and continuous for $s' \in (s_i(t_{0a}), \infty)$. The following bounds are staighforward:

$$\begin{aligned} &|I_{01}(ds)| \leqslant \frac{\text{const}}{|w(t)|} \ln^2|s|, \quad |I_{02}(s)| \leqslant \frac{\text{const}}{|w(t)|}, \\ &|I_{03}(s)| \leqslant \frac{\text{const}}{|w(t)|}. \end{aligned} \tag{I8}$$

They establish the bound

$$|I_0(s)| \leqslant \frac{\text{const}}{|w(t)|} \ln^2|s| \tag{I9}$$

which is sufficient for the reasoning of Sec. V.

We will have to verify that, even if t lies on Γ_i, the integrals in the dispersion relation for $l(s,t)$, Eq. (5.14) exist and define a function which has the same imaginary part as $l(s,t)$. To this end, we must replace $K(s,t)$ in Eq. (5.21) by the limiting value of the Cousin function on Γ_i, Eq. (5.7), divided by the polynomial containing the zeros of $C(s,t)$ in Q_t. We denote by $\widetilde{I}_i(s,t)$ the logarithm of the function of s, for t on Γ_i, defined by the right-hand side of eq. (5.7), divided by $F_i(s,t)$. Clearly, $\widetilde{I}_i(s,t)$ is continuous at all $s \in C_s$, except for those s for which $s = s_i(t)$. At these points, however, $\widetilde{I}_i(s,t)$ can be written as $(s - s_i(t))^{-\alpha} R(s,t)$, for any $0 < \alpha < 1$, and $R(s,t)$ a continuous function at $s = s_i(t)$ (vanishing there). The difference Δ_I between $I_i(s,t)$ and its dispersion relation may then have isolated singularities at $s = s_i(t)$, but these are such that $\lim_{s \to s_i(t)} (s - s_i(t))^{\alpha + \epsilon} \Delta_I(s) = 0$, so that, in fact, $\Delta_I(s)$ is holomorphic there. Then the argument of Sec. V, following Eq. (5.23) can be repeated, and this finishes the proof.

APPENDIX J: EXAMPLES TO SEC. V

We discuss two examples: the first is artificial, but leads to a Cousin function in algebraic form: the second one is more realistic and we bring the Cousin function to a form which is more illuminating than the starting point. We set[41]: $4m^2 = 1$, $\sigma = (1 - s)^{1/2}$, $\tau = (1 - t)^{1/2}$, $\eta = (1 - u)^{1/2}$. The Mandelstam domain is then the set of points $(\sigma, \tau, \eta) \in \mathbb{C}^3$ lying on the complex surface

$$\sigma^2 + \tau^2 + \eta^2 = 2 \tag{J1}$$

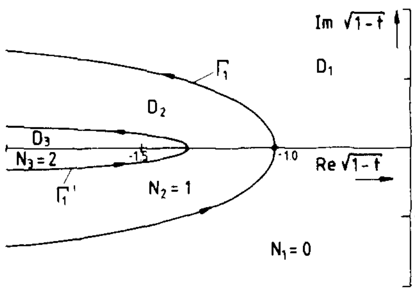


FIG. 10. Example 1 of Appendix J.

and obeying

$$\operatorname{Re}\sigma<0,\ \operatorname{Re}\tau<0,\ \operatorname{Re}\eta<0. \tag{J2}$$

I. Consider

$$A(\sigma,\tau,\eta)=2\sigma+\eta+1. \tag{J3}$$

We may ignore the fact that A has no t cut, start from the pattern of zero trajectories in $D_t (D_t=\operatorname{Re}\tau<0)$ and apply the method of Sec. V. The situation is the following (see Fig. 10): for τ on $\operatorname{Re}\tau=0$, there are no solutions in $\overline{\mathscr{M}}$; for η on $\operatorname{Re}\eta=0$, there is one solution in $\mathscr{M}$ (generating Γ_1')

$$(\sigma(\eta),\tau_-(\eta),\eta)=\left(-\frac{(\eta+1)}{2},\ -\frac{1}{2}(7-2\eta-5\eta^2)^{1/2},\eta\right) \tag{J4}$$

(the root has a positive real part); for σ on $\operatorname{Re}\sigma=0$, there is also one solution in $\overline{\mathscr{M}}$, generating Γ_1:

$$(\sigma,\tau_-(\sigma),\eta(\sigma))=(\sigma,-(1-4\sigma-5\sigma^2)^{1/2},\ -2\sigma-1). \tag{J5}$$

The number of zeros in $\overline{\mathscr{M}}$ for τ in the regions of D_t delimited by Γ_1, Γ_1' and $\operatorname{Re}\tau=0$ is shown in Fig. 10. There is a discontinuity at $\tau=-1$ in the logarithm appearing in $K_{0s}(s,t)$. The exponent of $K_{0s}(s,t)$ is

$$I=-\frac{1}{2\pi i}\int_{\Gamma_1}\frac{\ln(\sigma_1^2(\tau')-\sigma^2)}{\tau'^2-\tau^2}\frac{w(\tau)}{w(\tau')}2\tau'd\tau'. \tag{J6}$$

We choose $w(\tau)=(\tau-1)(\tau-2)$, $\sigma_1(\tau)$ is the inverse function defined on Γ_1 and write $I=I_1+I_2$ with

$$I_2=-\frac{1}{2\pi i}\int_{\Gamma_1}\frac{\ln(\sigma_1^2(\tau')-\sigma^2)}{\tau'+\tau}\frac{w(\tau)}{w(\tau')}d\tau'. \tag{J7}$$

$I_2(\sigma,\tau)$ is a holomorphic function of σ,τ in $\operatorname{Re}\sigma<0\otimes\operatorname{Re}\tau<0$ and generates thus a factor $C_E'(\sigma,\tau)$ of $C(\sigma,\tau)$ which is nonvanishing in $\operatorname{Re}\sigma<0\otimes\operatorname{Re}\tau<0$ and consequently in $\mathscr{M}$. The remaining integral can be computed as in Appendix F by going over to the σ'-plane. The choice of branches of the logarithm indicated in the text is realized if, for σ real, one lays the cut from $|\sigma|$ to the left. In complete analogy to Appendix F one obtains

$$K_{0s}(s,t)=\begin{cases} C_E(\sigma,\tau)\dfrac{5\sigma^2+4\sigma+\tau^2-1}{\sigma_-^2(\tau)-\sigma^2} & \tau\in\mathscr{D}_1\\ C_E(\sigma,\tau)\dfrac{5\sigma^2+4\sigma+\tau^2-1}{(\sigma_-^2(\tau)-\sigma^2)(\sigma_+^2(\tau)-\sigma^2)} & \tau\in\mathscr{D}_2,\mathscr{D}_3\end{cases} \tag{J8}$$

where

$$\sigma_\pm(\tau)=\frac{-2\pm(9-5\tau^2)^{1/2}}{5} \tag{J9}$$

are the roots of the function in the numerator, the square root has a positive real part and $C_E(\sigma,\tau)$ is the product of $C_E'(\sigma,\tau)$ with other factors, nonvanishing in $\operatorname{Re}\sigma<0\otimes\operatorname{Re}\tau<0$, coming from the weight $w(\tau')$. We can regard $K_{0s}(s,t)$ as the function $\mathscr{K}$, Eq. (3.1), for the amplitude $5\sigma^2+4\sigma+\tau^2-1$ (holomorphic in a direct product domain) except for the replacement in the denominator of $\sigma_-(\tau)-\sigma$ by the difference of squares. Similarly,

$$K_{0u}(u,t)=\begin{cases}\overline{C}_E(\eta,\tau)\dfrac{5\eta^2+2\eta+4\tau^2-7}{\eta^2-\eta_-^2(\tau)} & \tau\in\mathscr{D}_1,\mathscr{D}_2\\ \overline{C}_E(\eta,\tau)\dfrac{5\eta^2+2\eta+4\tau^2-7}{(\eta^2-\eta_-^2(\tau))(\eta^2-\eta_+^2(\tau))} & \tau\in\mathscr{D}_3\end{cases}, \tag{J10}$$

where η_-, η_+ are the roots of the polynomial in the numerator. Notice

$$\eta_+^2(\tau)+\sigma_-^2(\tau)+\tau^2=\eta_-^2(\tau)+\sigma_+^2(\tau)+\tau^2=2. \tag{J11}$$

One obtains

$$C_1(\sigma,\tau)=C(\sigma,\tau)C_E(\sigma,\tau)^{-1}\overline{C}_E(\eta,\tau)^{-1}$$
$$=\begin{cases}-\dfrac{(\sigma_+(\tau)-\sigma)(\eta-\eta_+(\tau))}{(\sigma_-(\tau)+\sigma)(\eta+\eta_-(\tau))} & \tau\in\mathscr{D}_1,\mathscr{D}_2\\ \dfrac{(\sigma_+(\tau)-\sigma)(\sigma_-(\tau)-\sigma)}{(\eta+\eta_+(\tau))(\eta+\eta_-(\tau))} & \tau\in\mathscr{D}_3\end{cases}. \tag{J12}$$

Using (J11), we verify that $C_1(\sigma,\tau)$ is actually one and the same function, namely

$$C_1(\sigma,\tau)=\frac{5\sigma^2+4\sigma-1+\tau^2}{5\eta^2-2\eta+4\tau^2-7}. \tag{J13}$$

Further, using (J1) we obtain

$$C_1(\sigma,\tau)=-\frac{2\sigma+\eta+1}{2\sigma+\eta-1}, \tag{J14}$$

where the denominator is nonvanishing in $\operatorname{Re}\sigma<0\otimes\operatorname{Re}\eta<0$.

II. Consider

$$A(\sigma,\tau,\eta)=2\sigma+\tau+\eta+2. \tag{J15}$$

The pattern of trajectories in the t plane is shown in Fig. 4. The situation is the following: for τ on $\operatorname{Re}\tau=0$, there is a trajectory

$$(\sigma_+(\tau),\tau,\eta_-(\tau))=\left(\frac{-2(\tau+2)+(6-6\tau^2-4\tau)^{1/2}}{5},\ \tau,\ \frac{-(\tau+2)-2(6-6\tau^2-4\tau)^{1/2}}{5}\right) \tag{J16}$$

on the physical sheet for small $|\tau|$; it disappears through the s cut at $\tau=\pm 2i(2/5)^{1/2}$. For σ on $\operatorname{Re}\sigma=0$, there are two trajectories near $\sigma=0$, crossing at this point

$$(\sigma,\tau_\mp(\sigma),\eta_\pm(\sigma))=\left(\sigma,\frac{-2(\sigma+1)\pm(-6\sigma^2-8\sigma)^{1/2}}{2},\ \frac{-2(\sigma+1))\mp(-6\sigma^2-8\,\sigma)^{1/2}}{2}\right) \tag{J17}$$

One of the trajectories disappears through the t cut, the other through the u cut, at the same value of $\sigma = \pm i(2/5)^{1/2}$; their images are the curves Γ_1, Γ_2 in Fig. 4; the functions $\tau_{\pm}(\sigma),\eta_{\mp}(\sigma)$ satisfy a relation analogous to (J11). For η on $\operatorname{Re}\eta = 0$, there is one trajectory obtained from (J16) by interchanging η with τ; it generates Γ'_3 in Fig. 4 and disappears through the s cut at $\eta = \pm 2i(2/5)^{1/2}$. The number of zeros of $A(\sigma,\tau,\eta)$ in $\mathscr{M}$, for τ in the various domains of D_t is shown in Fig. 4. There are two closed circuits $\tilde{\Gamma}_1,\tilde{\Gamma}_2$ required to return the inverse functions to their original values: $\tilde{\Gamma}_1 \equiv \Gamma_1 \cup (\frac{8}{5} - i\epsilon, \frac{8}{5} + i\epsilon), \tilde{\Gamma}_2 = \Gamma_2 \cup \Gamma'_3$.

Using the experience of Example 1, we build $C(s,t,u)$ (up to factors which do not vanish in $\operatorname{Re}\sigma<0 \otimes \operatorname{Re}\tau<0$ or $\operatorname{Re}\eta<0 \otimes \operatorname{Re}\tau<0$) from the functions $\mathscr{K}$, Eq. (3.1) needed in the Cousin functions of the amplitudes $F_1(\sigma,\tau),F_2(\eta,\tau)$ obtained by eliminating in turn η and σ between Eqs. (J1) and (J15). There is a difference with respect to example 1: the pattern of trajectories of $A(s,t,u)$ in D_t is only a part of the superposition of the patterns obtained from $F_1(\sigma,\tau),F_2(\eta,\tau)$: the dotted lines Γ_4, Γ'_5 in Fig. 4 are missing in the former.

$$\begin{aligned} F_1(\sigma,\tau) &= 5\sigma^2 + 4\sigma(2+\tau) + 2(\tau+1)^2 \\ &= 5(\sigma - \sigma_+(\tau))(\sigma - \sigma_-(\tau)) \\ F_2(\eta,\tau) &= 5\eta^2 + 2\eta(2+\tau) + 5\tau^2 + 4\tau - 4 \\ &= 5(\eta - \eta_-(\tau))(\eta - \eta_+(\tau)) \end{aligned} \tag{J18}$$

Up to factors that do not vanish in $\operatorname{Re}\sigma<0 \otimes \operatorname{Re}\tau<0$, $\operatorname{Re}\eta<0 \otimes \operatorname{Re}\tau<0$ and using

$$\psi_1(\sigma,\tau)\psi_2(\eta,\tau) = \exp\left[\frac{1}{2\pi i}\int_{\Gamma_4}\frac{\ln(\sigma_-^2(\tau') - \sigma^2)}{\tau'-\tau}d\tau'\right]\exp\left[\frac{1}{2\pi i}\int_{\Gamma'_5}\frac{\ln(\eta^2 - \eta_+^2(\tau'))}{\tau'-\tau}d\tau'\right] \tag{J19}$$

we obtain [in analogy to (J12)]

$$C(\sigma,\tau) = \begin{cases} -\dfrac{(\sigma-\sigma_+(\tau))(\eta-\eta_+(\tau))}{(\sigma+\sigma_-(\tau)(\eta+\eta_-(\tau))}(\tau+2)\psi_1(\sigma,\tau)\psi_2(\eta,\tau), & \tau\in\mathscr{D}_1,\mathscr{D}_2,\mathscr{D}_{3+} \\ \dfrac{(\sigma-\sigma_+(\tau))(\sigma-\sigma_-(\tau))}{(\eta+\eta_-(\tau))(\eta+\eta_+(\tau))}(\tau+2)\psi_1(\sigma,\tau)\psi_2(\eta,\tau), & \tau\in\mathscr{D}_4 \end{cases} \tag{J20}$$

In the last equation, we can replace the fraction by the ratio $F_1(\sigma,\tau)/F_2(-\eta,\tau)$. Using Eq. (J11) one can verify that the two lines in Eq. (J20) are in fact the same function. The notation $\mathscr{D}_{3+}$ means the part of $\mathscr{D}_3$ lying on the left of the curve Γ_4. The expressions (J20) have discontinuities along Γ_4, Γ'_5, but are otherwise holomorphic in $\mathscr{D}_{3+} \otimes \operatorname{Re}\sigma<0 \cap \mathscr{D}_{3+} \otimes \operatorname{Re}\eta<0$. The analytic continuation of (J20) to all of D_t gives

$$C(\sigma,\tau) = \begin{cases} \dfrac{F_1(\sigma,\tau)(\eta-\eta_+(\tau))}{\eta+\eta_-(\tau)}(\tau+2)\psi_1(\sigma,\tau)\psi_2(\eta,\tau), & \tau\in\mathscr{D}'_{3+} \\ F_1(\sigma,\tau)F_2(\eta,\tau)(\tau+2)\psi_1(\sigma,\tau)\psi_2(\eta,\tau), & \tau\in\mathscr{D}_{3-} \end{cases} \tag{J21}$$

where the notation $\mathscr{D}'_{3+}$, $\mathscr{D}_{3-}$ is given in Fig. 4. The factor $(\tau+2)$ removes a pole at $\tau = -2$, due to the discontinuity of the logarithm.

[1]A. Martin, Nuovo Cimento A **59**, 131 (1969).
[2]D. Atkinson, P. W. Johnson, and R. L. Warnock, Commun. Math Phys. **28**, 133 (1972); **33** 221 (1973).
[3]G. R. Bart, P. W. Johnson, and R. L. Warnock, J. Math. Phys. **14**, 1558 (1973); Nucl. Phys. B **72**, 329 (1974).
[4]D. Atkinson, G. Mahoux, and F. J. Yndurain, Nucl. Phys. B **54**, 263 (1973); B **66**, 429 (1973).
[5]C. Itzykson and A. Martin, Nuovo Cimento A **17**, 245 (1973).
[6]D. Atkinson, L. P. Kok, M. de Roo, and P. W. Johnson, Nucl. Phys. B **77**, 109 (1974).
[7]H. Burkhardt and A. Martin, Nuovo Cimento A **29**, 141 (1975).
[8]D. Atkinson, L. P. Kok, and M. de Roo, Phys. Rev. D **17**, 2492 (1978).
[9]J. E. Bowcock and H. Burkhardt, Rev. Mod. Phys. **38**, 1099 (1975).
[10]J. H. Crichton, Nuovo Cimento **45** A, 256 (1966).
[11]I. S. Stefanescu, J. Math. Phys. **23**, 1190 (1982).
[12]I. S. Stefanescu, Phys. Rev. D **21**, 3225 (1980).

[13]At most denumerable, since the set of zero surfaces of a function holomorphic in a domain is at most denumerable.

[14]G. Höhler and I. S. Stefanescu, Z. Physik C **2**, 253 (1979).
[15]C. B. Lang and I. S. Stefanescu, Phys. Lett. B **79**, 479 (1978).
[16]W. F. Osgood, *Lehrbuch der Funktionentheorie*, Bd. II, (Chelsea, New York, 1965), p. 259ff.
[17]R. C. Gunning and M. Rossi, *Analytic Functions of Several Complex Variables* (Prentice-Hall, Englewood Cliffs, NJ, 1965), pp. 31ff; F. Hörmander, *An Introduction to Complex Analysis in Several Variables* (van Nostrand, Princeton, NJ, 1966), p. 132ff.
[18]N. I. Muskhelishvili, *Singular Integral Equations* (Noordhoff, Groningen, The Netherlands, 1953).

[19]No distinction between γ_1 and $\bar{\gamma}_1$ need be made if $\bar{\gamma}_1$ is simple and has a single-valued inverse.
[20]Using the reasoning of Lemma 1, Sec. IV.

[21]H. W. Turnbull, *Theory of Equations* (Oliver and Boyd, Edinburgh, 1957).
[22]J. E. Marsden, *Basic Complex Analysis* (Freeman, San Francisco, 1973).

[23]A function $f(s,t)$ can be defined to be holomorphic in two variables in a domain $D_s \otimes D_t$ if it is holomorphic in D_t at each fixed s in D_s and holomorphic in D_s at fixed t in D_t; thus a citation of Hartogs' theorem might seem superfluous. However, some textbooks (Ref. 16) add to this definition the requirement that $f(s,t)$ be continuous in $D_s \otimes D_t$. Hartogs' theorem, which is quite difficult to prove (see Ref. 16, p. 227), states that if a function is holomorphic in each variable separately (in the sense above), it is actually continuous [in contrast to the situation for real variables, e.g., $f(x,y) = xy/(x^2+y^2)$, $x\neq 0$, $y\neq 0$, $=0$ if $x=0$ or $y=0$].

[24]P. L. Duren, *Theory of H^p Spaces* (Academic, New York and London, 1970).

[25]Lemma 1 is clearly true also for $s\in C_s^-$, if $\operatorname{Im} s>0$ is replaced by $\operatorname{Im} s<0$.

[26]W. Rudin, *Real and Complex Analysis* (McGraw-Hill, New York 1974).

[27]In the following, we drop for simplicity the indices i,k on ζ and on t and refer only to $s_{0,1} = s_1$.
[28]We apply Lemma 5 to $u(s)$, rather than $t(s)$, but this makes no difference to its result.
[29]The path $\mathscr{P}$ may consist of several turns around t_0.
[30]It is convenient to place an index (*) on those curves Γ_i,Γ'_i for which the argument runs along C_s^-,C_u^-. If the symbol * is present without brackets, it simply means complex conjugation (for curves Γ_i)

[31]H. Cornille and A. Martin, Nucl. Phys. B **48**, 104 (1972).

[32]G. Auberson, T. Kinoshita, and A. Martin, Phys. Rev. D **3**, 3185 (1971).
[33]H. Cornille and F. R. A. Simao, Nuovo Cimento **A5**, 138 (1971).
[34]G. Auberson and M. Roy, Nucl. Phys. B **117**, 322 (1976).
[35]P. Presnajder and J. Pisut, Nucl. Phys. B **82**, 313 (1974).
[36]P. Presnajder, in *Hadron Constituents and Symmetries*, Proceedings of the Conference Smolenice 1976, edited by J. Lanik and M. Noga (Slovak Academy of Sciences, Bratislava, 1978).
[37]M. Daumens and P. Winternitz, Phys. Rev. D **21**, 1919 (1980).
[38]I. Manolitsakis, J. Math. Phys. **23**, 2291 (1982).
[39]I. S. Stefanescu, work in progress
[40]There could be several zeros, lying in both D_+ and D_-; we consider just one for simplicity; the general treatment is analogous to that of Appendix E, Subsection b.
[41]No confusion should arise between σ,τ in this Appendix and the variables of Secs. II and III.

Exact diagonalization of the Hamiltonian of a Dirac particle with anomalous moment, interacting with an external magnetic field

Jean-François Roux
Service de Physique, Faculté des Sciences, Parc de Grandmont, 37200 Tours, France

(Received 11 August 1983; accepted for publication 6 January 1984)

A unitary transformation is performed on a Hamiltonian which is built from two independent Dirac matrices with noncommuting coefficient operators. A simple symmetric and exactly diagonalizable expression is obtained for the Hamiltonian of a Dirac particle, including the effect of an external magnetic field on an anomalous moment; a nonrelativistic approximation comes readily, which allows high magnetic fields.

PACS numbers: 12.20. – m

I. INTRODUCTION

The transformation of Foldy and Wouthuysen (FW)[1] is a well-known method for reducing the free Dirac particle Hamiltonian to an even form; the consequence is a diagonal form and the possible definition of particle states with two-component spinors. When the particle is interacting with an external field, the FW transformation does not generally enable an exact diagonalization; several approximate methods have been used,[1–5] with equivalent results up to a given order.

As first shown by Case,[6] a closed form does exist for the diagonalizing transformation in the case of an odd interacting term, provided some conditions of commutation are satisfied; the method runs on an electron in a homogeneous magnetic field. In some other cases of interaction, exact diagonal expressions have been given for the transformed Hamiltonian; an extension of the theorem of Case[6] is used, which has been stated by Cohen[7] and by Johnson and Chang,[8] and which involves the same conditions of commutation.

Eriksen[2] uses its method on a neutral Dirac particle with anomalous electric moment, in a constant electric field.

Tsai[9] works on a charged Dirac particle with anomalous magnetic moment, in a constant homogeneous magnetic field. He first carries out a Melosh-type transformation, and then gets the eigenvalues of the energy after squaring the transformed Hamiltonian.

On the same case, Weaver[10] has shown that the two steps can be reversed. He also uses other ways, consisting either of a projection (giving a two-component equation) followed with a Melosh-type transformation or of two successive Melosh-type transformations; the diagonalization is bound to a diagonal realization of the Dirac matrix α_3.

We elsewhere[11] gave a method, derived from the transformation of Baktavatsalou,[12] which brings a given Dirac Hamiltonian to a parametered form; a correct choice of the parameters gives diagonalizable expressions.

We first briefly develop the method. An application is then performed on the case of a Dirac charged particle with anomalous moment, interacting with a constant magnetic field. It is a one-step and closed form transformation, the result is different from those of Tsai[9] and Weaver[10]; it agrees with Case's result[6] for a vanishing anomalous moment. Different approximations can be derived, whatever the intensity of the magnetic field may be, showing particularly the interest of the method in high constant inhomogeneous field.

II. AN OUTLINE OF THE METHOD

Just like Baktavatsalou,[12] we define a unitary transformation U as the Cayley transform of a Hermitian operator Λ:

$$U = (\Lambda - iI)(\Lambda + iI)^{-1}. \tag{1}$$

The transformation is applied on a time-independent Dirac Hamiltonian H:

$$H' = UHU^{-1}. \tag{2}$$

From this expression and definition (1), the following equation holds:

$$\Lambda\,(H - H')\Lambda + i[\Lambda, (H + H')]_- + (H - H') = 0. \tag{3}$$

It can be cast into an equivalent form:

$$[(H - H')\Lambda + i(H + H')]^2 - 2i(H^2 - H'^2)\Lambda + 2(H^2 + H'^2) = 0. \tag{4}$$

With given expressions for H and H', the Hermitian operator Λ is theoretically derived as a solution of Eq. (4). Such a general solution, however, looks out of reach.

A solution has been derived[11] with the following constraint assumed on the expression of the transformed Hamiltonian H':

$$H'^2 = H^2. \tag{5}$$

Let us notice here that the most general expression for a Dirac particle Hamiltonian reads

$$H = \beta M + \gamma_5 P + \beta\gamma_5 N + Q. \tag{6}$$

Moreover, we want the transformed Hamiltonian to be diagonalizable, and so it has to include only one of the Dirac matrices.

Thus, as it can be verified, expression (5) only holds being given the two following conditions:

(i) the Hamiltonian H is homogeneous ($Q = 0$);

(ii) one term among the three ones, M, N, P, must commute with the two other ones.

Then, Eq. (4) can be simplified:

$$[(H - H')\Lambda + i(H + H')]^2 = -4H^2. \tag{7}$$

Let $\mathscr{A}$ be a Hermitian operator such that

$$H^2 = \mathscr{A}^2 = H'^2. \tag{8}$$

 0022-2488/84/062087-03$02.50

Equation (7) gives

$$(H-H')\Lambda + i(H+H') = \pm 2i\mathscr{A}. \tag{9}$$

One then gets

$$\Lambda = i(H-H')^{-1}[\pm 2\mathscr{A} - (H+H')], \tag{10}$$

with the following Hamiltonian:

$$H' = \mathscr{A}^{-1}H\mathscr{A}. \tag{11}$$

A solution is thus obtained for the Hermitian operator Λ and the unitary transformation:

$$\Lambda = i[\mathscr{A},H]_{-}^{-1}\{\pm 2H^2 - [\mathscr{A},H]_{+}\}, \tag{12}$$

$$U = (H-\mathscr{A})^{-1}(\pm H - H^{-1}\mathscr{A}H). \tag{13}$$

III. APPLICATION ON A GIVEN FORM OF THE HAMILTONIAN

Let H be with only two terms:

$$H = \beta M + \gamma_5 P. \tag{14}$$

Operators P and M generally do not commute with each other.
Operator $\mathscr{A}$ reads, as the most general expression,

$$\mathscr{A} = A + iB\beta\gamma_5 + C\beta - iD\gamma_5. \tag{15}$$

It follows from condition (8) that A, B, C are Hermitian operators and D is an anti-Hermitian one.

To proceed with the solution, let

$$\begin{aligned} &A+B=X, \quad A-B=\overline{X},\\ &C+D=Y, \quad C-D=\overline{Y},\\ &R=P^2+M^2+i[P,M]_{-},\\ &\overline{R}=P^2+M^2-i[P,M]_{-}. \end{aligned} \tag{16}$$

Using (16) in (15), and (14) and (15) in (11), the transformed Hamiltonian H' then reads

$$\begin{aligned} H' = {}&\tfrac{1}{2}\{R^{-1}[X(M+iP)\overline{Y}+Y(M-iP)X]+\overline{R}^{-1}[\overline{X}(M-iP)Y+\overline{Y}(M+iP)\overline{X}]\}\\ &+\tfrac{1}{2}i\beta\gamma_5\{R^{-1}[X(M+iP)\overline{Y}+Y(M-iP)X]-\overline{R}^{-1}[\overline{X}(M-iP)Y+\overline{Y}(M+iP)\overline{X}]\}\\ &+\tfrac{1}{2}\beta\{R^{-1}[X(M+iP)\overline{X}+Y(M-iP)Y]+\overline{R}^{-1}[\overline{X}(M-iP)X+\overline{Y}(M+iP)\overline{Y}]\}\\ &+\tfrac{1}{2}i\gamma_5\{R^{-1}[X(M+iP)\overline{X}+Y(M-iP)Y]-\overline{R}^{-1}[\overline{X}(M-iP)X+\overline{Y}(M+iP)\overline{Y}]\}. \end{aligned} \tag{17}$$

The particular case with $M=m$ (rest mass of the particle) can be solved with

$$Y = kP,$$

where k is a Hermitian operator, acting as a function of m and P.

In the general case, let us suppose, as an extension,

$$Y = h(M+iP). \tag{18}$$

h is a Hermitian operator, a function of M and P, which commutes with the product $(M+iP)(M-iP)$. The transformed Hamiltonian (17) can be diagonalized only provided that it is axial[13]; this means that, among the three terms including the Dirac matrices β, γ_5, and $\beta\gamma_5$, two ones are vanishing. From the symmetry of the expression, the β and γ_5 terms are to vanish. Operator h in (18) is then derived, and H' has a simple expression:

$$H' = \pm\left[\tfrac{1}{2}(R^{1/2}-\overline{R}^{1/2})+\tfrac{1}{2}i\beta\gamma_5(R^{1/2}+\overline{R}^{1/2})\right]. \tag{19}$$

IV. THE CASE OF A DIRAC PARTICLE WITH ANOMALOUS MOMENT, INTERACTING WITH A CONSTANT MAGNETIC FIELD

Now, operators P and M respectively read

$$\begin{aligned} &P = \boldsymbol{\sigma}\cdot(\mathbf{p}-e\mathbf{A}),\\ &M = m+\mu_a\boldsymbol{\sigma}\cdot\mathbf{B}; \end{aligned} \tag{20}$$

μ_a is the anomalous moment. Substitution of these expressions in (16) gives

$$\begin{aligned} R = {}& m^2+\mu_a^2\mathbf{B}^2+(\mathbf{p}-e\mathbf{A})^2-e\boldsymbol{\sigma}\cdot\mathbf{B}\\ &+2\mu_a\boldsymbol{\sigma}\cdot[m\mathbf{B}+\mathbf{B}\times(\mathbf{p}-e\mathbf{A})],\\ \overline{R} = {}& m^2+\mu_a^2\mathbf{B}^2+(\mathbf{p}-e\mathbf{A})^2-e\boldsymbol{\sigma}\cdot\mathbf{B}\\ &+2\mu_a\boldsymbol{\sigma}\cdot[m\mathbf{B}-\mathbf{B}\times(\mathbf{p}-e\mathbf{A})]. \end{aligned} \tag{21}$$

Positive signs have been chosen in expression (19). A nonrelativistic approximation is readily derived up to first order in (1/m):

$$\begin{aligned} H'_{\rm nr} \simeq {}& (\mu_a/m)\boldsymbol{\sigma}\cdot[\mathbf{B}\times(\mathbf{p}-e\mathbf{A})]\\ &+i\beta\gamma_5\{m+(1/2m)[(\mathbf{p}-e\mathbf{A})^2-e\boldsymbol{\sigma}\cdot\mathbf{B}+\mu_a^2\mathbf{B}^2]\}. \end{aligned} \tag{22}$$

In the subspace of two-component spinors representing particle states, with a diagonal realization for matrix $i\beta\gamma_5$, the approximate expression of the transformed Hamiltonian is

$$\begin{aligned} H''_{\rm nr} \simeq {}& m+(1/2m)(\mathbf{p}-e\mathbf{A})^2\\ &-\boldsymbol{\sigma}\cdot[e\mathbf{B}-(\mu_a/m)\mathbf{B}\times(\mathbf{p}-e\mathbf{A})]+\mu_a^2\mathbf{B}^2. \end{aligned} \tag{23}$$

The classical Pauli Hamiltonian is recovered with a vanishing anomalous moment.

V. DISCUSSION AND CONCLUSION

The method here described differs from those of Tsai and Weaver, since we get a diagonalizable Hamiltonian for a spin-$\frac{1}{2}$ particle by means of a one-step transformation. It is more general, since the case of a nonhomogeneous magnetic field can be considered.

The transformed Hamiltonian (19) has a simple symmetric expression; with the matrix $i\beta\gamma_5$ diagonal, the particle (resp. antiparticle) states are described by means of the Hamiltonian:

$$H'' = R^{1/2} \quad (\text{resp. } H'' = -\overline{R}^{1/2}).$$

Although the expressions are simple, the occurrence of square roots involves approximations in effective developments.

As a counterpart to the simplicity of the Hamiltonian, the expression for the unitary transformation U is not explicitly known, and we thus cannot give expressions for other transformed operators and observables, which is a lack in view of interpretation.

The method should run on a spin-1 particle with anomalous moment, since the Sakata–Takentani[14] equation includes two matrices which follow the same algebra as the Dirac matrices β, γ_5, and $i\beta\gamma_5$. The general case of higher spins has not yet been considered.

The results are of interest as they allow a description of quasiuncoupled particle/antiparticle states in high magnetic fields, such as those occurring in the vicinity of fast rotating stars.

ACKNOWLEDGMENT

The author would like to thank the referee for helpful remarks on the first version of this paper and for calling attention on the possible extension of the method to higher spins.

[1]L. L. Foldy and S. A. Wouthuysen, Phys. Rev. **78**, 29 (1950).
[2]E. Eriksen, Phys. Rev. **111**, 1011 (1958); E. Eriksen and M. Kolsrud, Suppl. Nuovo Cimento **18**, 1 (1960).
[3]G. Morpurgo, Nuovo Cimento **15**, 624 (1960).
[4]E. I. Blount, Phys. Rev. **128**, 2454 (1962).
[5]J. D. Morrison and R. E. Moss, Mol. Phys. **41**, 491 (1980).
[6]K. M. Case, Phys. Rev. **95**, 1323 (1954).
[7]A. Cohen, Am. J. Phys. **39**, 409 (1971).
[8]J. E. Johnson and K. K. Chang, Phys. Rev. D **10**, 2421 (1974).
[9]W. Tsai, Phys. Rev. D **7**, 1945 (1973).
[10]D. L. Weaver, Phys. Rev. D **12**, 4001 (1975); **14**, 2824 (1976); J. Math. Phys. **18**, 306 (1977).
[11]J. F. Roux, Nuovo Cimento B **29**, 225 (1975).
[12]M. Baktavatsalou, Nuovo Cimento **25**, 964 (1962).
[13]J. F. Roux, Lett. Nuovo Cimento **20**, 39 (1977).
[14]M. Taketani and S. Sakata, Proc. Phys. Math. Soc. Jpn. **22**, 757 (1939).

Combined symmetries in curved space-times

M. D. Maia
International Centre for Theoretical Physics, Trieste, Italy and Universidade de Brasília, Departamento de Matemática, 70.910 Brasília D. F., Brazil

(Received 22 March 1983; accepted for publication 18 October 1983)

The exact combination of internal and of geometrical symmetries in curved space-times is discussed. It is seen that it is possible to define locally a Lie group S containing two noninvariant subgroups E and G, such that in the flat limit of the space-time, S decomposes in the product of the Poincaré group resulting from a contraction of E and an internal group with point dependent parameters resulting fom G. Furthermore, E does not have necessarily a nilpotent action on S except in the flat limit. An explicit squared mass-splitting expression for S is derived and its behavior in a local gravitational field and in the flat limit of the space-time is analyzed.

PACS numbers: 12.25. + e, 11.30.Ly, 02.20.Rt

1. INTRODUCTION

It has been known for some time that an exact combination of the Poincaré group with an internal symmetry group produces a mass spectra for particle multiplets which is either continuous or it consists of a single point. An analysis of the various pertinent theorems shows that this happens because the Lie algebra of the Poincaré group acts nilpotently on that of the combined symmetry.[1–5] This nilpotent action can be avoided if the combined symmetry is not a Lie group as, for example, in some infinite Lie group combinations and in some supersymmetric combinations.[6] On the other hand the mentioned theorems also indicate that the nilpotent action is a consequence of the peculiar structure of the Poincaré group, which has a mass operator constructed with operators belonging to an Abelian invariant subgroup. In this context Segal proposed that the mass-splitting problem could also be solved if the Poincaré group P is replaced by a more general space-time symmetry E such that it contracts into P and that the nilpotent action of E on a combined symmetry occurs only after that contraction.[7]

The replacement of the Poincaré group by another space-time symmetry, notably by the de Sitter cosmological symmetry, has been considered by several authors.[8–12] Because the de Sitter group SO(4,1) [or the anti de Sitter group SO(3,2)] is semisimple and produces the Poincaré group by contraction, the nilpotent action of this group on a combined symmetry group would occur only in the flat limit of the de Sitter universe. Interestingly enough, the presence of a gravitational field even at the weak cosmological scale of strength is sufficient to produce a nonzero mass splitting. The resulting masses and mass splitting are admittedly small [of the order of 10^{-50} (MeV)^2] and although these masses could in principle be rescaled, the result suggests that a local gravitational field, like the one existing at the surface of the Earth, could produce a more pronounced mass difference.

The purpose of this note is to investigate a possible implementation of Segal's proposal for a curved space-time V of general relativity so that a local gravitational field is considered in the evaluation of the masses of particles.

In Sec. 2 a local space-time group E is constructed in a natural way such that, in the flat limit of V, E contracts into the Poincaré group. In general a space-time V does not admit Killing vector fields. Therefore, E cannot be identified with the group of isometries of V. However, at any point of V a local space can be constructed such that the group of isometries of that space generate the group of isometries of V if any. Here this local space is taken to be the local isometric minimal-embedding space $M(m,n)$ of V. The isometric-embedding condition says that the metric of V is induced by that of $M(m,n)$. This means that instead of the group of isometries of V, the group of isometries of $M(m,n)$ can be considered in the definition of the group E.

Since the concept of internal symmetry is usually defined in Minkowski space, that is in conjunction with the Poincaré group, then in the presence of the gravitational field the internal group does not necessarily decouple from the space-time symmetry E. Therefore, a combined symmetry S containing E as a subgroup has to be considered. It happens that in some cases the group of isometries of the embedding space is sufficiently large to accommodate not only E but another subgroup G locally defined on V. G and E are noninvariant subgroups of S but in the flat limit of V, E contracts into P while G becomes completely disconnected from E (that is, P) so that G can be regarded as an internal group. In Sec. 3 the mass-splitting problem, is investigated for the combined symmetry S. The mass operator is taken to be the second-order Casimir operator of S, reproducing the Poincaré mass operator in the flat limit condition. A general squared mass-splitting expression is derived, showing its dependence on the local curvature radius of the space-time. In the flat limit of V the mass splitting vanishes as expected.

Some estimates of the order of magnitude of the masses for the Schwarzschild space-time at the surface of the Earth is given at the end of that section.

2. THE EMBEDDING SYMMETRY

In order to guarantee a maximal group of isometries, the isometric local-embedding space $M(m,n)$ is assumed to be flat (although in principle a constant curvature embedding space could also be considered). The dimension of $M(m,n)$ is p $(= m + n)$ and its metric signature is $m + (-n)$. Since $M(m,n)$ is flat, Cartesian coordinates Z^μ $(\mu = 1,\dots, p)$ may be used throughout. On the other hand, the use of a Gaussian coordinate system x^α $(\alpha = 1,\dots, p)$ based on V may be more convenient for dealing with objects which are space-time defined. This Gaussian system can be

 0022-2488/84/062090-05$02.50

constructed with the four space-time coordinates x^i ($i = 1,\ldots, 4$) and with $p-4$ coordinates x^A ($A = 5,\ldots, p$), defined on $p-4$ unit directions $\eta_A(x^i)$ orthogonal to V and to each other with respect to the metric of $M(m,n)$. Thus $\{x^\alpha\} = \{x^i, x^A\}$. The transformation between these two coordinate systems is given by

$$Z^\mu(x^i,x^A) = X^\mu(x^i) + x^A\eta^\mu_A(x^i), \tag{1}$$

where $X^\mu(x^i)$ are the Cartesian coordinates of a point in V and η^μ_A are the Cartesian components of η_A.

The index convention will be as follows: Greek indices run from 1 to p. Small case Latin indices run from 1 to 4 and capital Latin indices run from 5 to p. Summation applies automatically to all cases of repeated upper and lower indices.

From (1) it follows that the definition of V as a hypersurface of $M(m,n)$ is given simply by $x^A = 0$. If $\Omega(x^\alpha)$ is a geometrical object defined in $M(m,n)$, its space-time "projection" is defined by restriction to V

$$\Omega(x^\alpha)|_V = \Omega(x^\alpha)|_{x^A=0}.$$

For example, if $\eta_{\mu\nu}$ denotes the Cartesian components of the metric tensor of $M(m,n)$, its Gaussian components are

$$g_{\alpha\beta} = Z^\mu_{,\alpha} Z^\nu_{,\beta}\eta_{\mu\nu}. \tag{2}$$

Therefore, its space-time projections are

$$\begin{aligned} g_{ij}|_V &= X^\mu_{,i}X^\nu_{,j}\eta_{\mu\nu} = \bar g_{ij},\\ g_{iA}|_V &= 0, \\ g_{AB}|_V &= \epsilon^A\delta_{AB}, \end{aligned} \tag{3}$$

where $\bar g_{ij}$ denotes the metric tensor of V in the coordinates x^i. Notice that the last equation follows from the assumed orthogonality of the η_A fields

$$\eta^\mu_A\eta^\nu_B\eta_{\mu\nu} = \epsilon^A\delta_{AB},$$

where $\epsilon^A = \pm 1$ are the signature numbers of the subspace orthongonal to V.

The Christoffel symbols $\Gamma_{\alpha\beta\gamma}$ of the metric affine connection ∇M of $M(m,n)$ in the Gaussian system are, as usual

$$\Gamma_{\alpha\beta\gamma} = \tfrac{1}{2}(g_{\alpha\gamma,\beta} + g_{\beta\gamma,\alpha} - g_{\alpha\beta,\gamma}). \tag{4}$$

Then it follows that $\Gamma_{ijk}|_V = \bar\Gamma_{ijk}$ (the Christoffel symbols of $\bar g_{ij}$). The Gauss–Codazzi–Ricci equations which are the integrability conditions for the embedding are obtained from the independent equations resulting from $R^\alpha{}_{\beta\gamma\delta}|_V = 0$ which expresses the flatness of $M(m,n)$ as seen from V.[13,14]

Now consider an infinitesimal displacement dx^i in V and from (1) calculate the corresponding variation on Z^μ

$$\Delta Z^\mu = (X^\mu_{,i} + x^A\eta^\mu_{A,i})dx^i.$$

If the direction dx^i is such that the above variation vanishes then that point Z^μ is a center of curvature and dx^i defines a principal direction in V. Thus $\Delta Z^\mu = 0$ gives after contraction with $X^\nu_{,j}\eta_{\mu\nu}$

$$(\bar g_{ij} + x^A b_{ijA})dx^i = 0,$$

where

$$b_{ijA} = \eta^\mu_{A,i}X^\nu_{,j}\eta_{\mu\nu} = \Gamma_{ijA}|_V.$$

The above equation has a nontrivial solution dx^i when

$$\det(\bar g_{ij} + x^A b_{ijA}) = 0. \tag{5}$$

The solutions $x^A = \rho^A_{(i)}$ of this equation define the radius of curvature of V, corresponding to the principal direction dx^i and to the normal direction η^A. These quantities express the curvature of V in an extrinsic form. Considering all possible principal directions dx^i, each normal η^A has a curvature radius component defined by

$$\rho^A = \bar g^{ij}\rho^A_{(i)}\rho^A_{(j)}. \tag{6}$$

Then each $1/\rho^A$ tends to zero in the flat limit of S and they can be used as contracting factors for the group E with respect to the Lorentz group of the tangent space.

The group of symmetries of the local isometric embedding is $SO(m,n)$ whose Lie algebra in the Cartesian frame is given by

$$[L_{\mu\nu}, L_{\rho\sigma}] = \eta_{\mu\rho}L_{\nu\sigma} + \eta_{\nu\sigma}L_{\mu\rho} - \eta_{\mu\sigma}L_{\nu\rho} - \eta_{\nu\rho}L_{\mu\sigma}.$$

In the Gaussian frame the Lie algebra operators are

$$L_{\alpha\beta} = Z^\mu_{,\alpha}Z^\nu_{,\beta}L_{\mu\nu} \tag{7}$$

so that

$$[L_{\alpha\beta}, L_{\gamma\delta}] = g_{\alpha\gamma}L_{\beta\delta} + g_{\beta\delta}L_{\alpha\gamma} - g_{\alpha\delta}L_{\beta\gamma} - g_{\beta\gamma}L_{\alpha\delta}. \tag{8}$$

Projecting this Lie algebra on V and denoting $l_{ij} = L_{ij}|_V$, $l_{iA} = L_{iA}|_V$, $l_{AB} = L_{AB}|_V$ and using (3), the projected Lie algebra reads

$$\begin{aligned} [l_{ij}, l_{kl}] &= \bar g_{ik}l_{jl} + \bar g_{jl}l_{ik} - \bar g_{il}l_{jk} - \bar g_{jk}l_{il},\\ [l_{ij}, l_{kA}] &= \bar g_{ik}l_{jA} - \bar g_{jk}l_{iA},\\ [l_{ij}, l_{AB}] &= 0,\\ [l_{iA}, l_{jB}] &= \bar g_{ij}l_{AB} + g_{AB}l_{ij},\\ [l_{iA}, l_{BC}] &= g_{AC}l_{iB} - g_{AB}l_{iC},\\ [l_{AB}, l_{CD}] &= g_{AC}l_{BD} + g_{BD}l_{AC} - g_{AD}l_{BC} - g_{BC}l_{AD}. \end{aligned} \tag{9}$$

Now change the basis of this Lie algebra by introducing the operators (a linear combination of l_{iA} at each point)

$$\pi_i = \sum_A \frac{1}{\rho^A} l_{iA}. \tag{10}$$

The resulting Lie algebra is

$$\begin{aligned} [l_{ij}, l_{kl}] &= \bar g_{ik}l_{jl} + \bar g_{jl}l_{ik} - \bar g_{il}l_{jk} - \bar g_{ik}l_{il},\\ [l_{ij}, \pi_k] &= \bar g_{ik}\pi_j - \bar g_{jk}\pi_i,\\ [\pi_i, \pi_j] &= \sum g_{AB}\frac{1}{\rho^A}\frac{1}{\rho^B}l_{ij},\\ [\pi_i, l_{BC}] &= \sum\frac{1}{\rho^A}g_{AC}l_{iB} - \sum\frac{1}{\rho^A}g_{AB}l_{iC},\\ [l_{AB}, l_{CD}] &= g_{AC}l_{BD} + g_{BD}l_{AC} - g_{AD}l_{BC} - g_{BC}l_{AD}. \end{aligned} \tag{11}$$

The group S is the Lie group corresponding to the above Lie algebra. It is in essence the projected group $SO(m,n)$ in an appropriate basis. It contains two noninvariant subgroups: The 10-parameter subgroup E generated by l_{ij} and π_i and the $(P-4)(P-5)/2$-parameter subgroup G generated by l_{AB}

$$[E,E] = E, \quad [E,G] \in S, \quad [G,G] = G.$$

The action of E and G on V can be best analyzed from the associated infinitesimal transformations. The infinitesimal transformations of $SO(m,n)$ are

$$x'^{\gamma} = x^{\gamma} + \xi^{\gamma}$$

with

$$\xi^{\gamma} = \theta^{\alpha\beta} L_{\alpha\beta}(x^{\gamma}),$$

where $\theta^{\alpha\beta}$ are the infinitesimal parameters of SO(m,n) and since this is the homogeneous group of isometries of $M(m,n)$

$$\xi^{(\alpha;\beta)}(\nabla M) = 0 \quad \text{(fixed origin).} \tag{12}$$

The infinitesimal transformations of the projected group S are

$$x'^{\alpha} = x^{\alpha} + \xi^{\alpha}$$

with Killing's equation evaluated at V

$$\xi^{(\alpha;\beta)}(\nabla M)|_V = 0. \tag{13}$$

Separating the different indices it follows that

$$\xi^{(i;j)}(\nabla M)|_V = (g^{k(i}\xi^{j)}{}_{,k} + g^{k(i}\Gamma^{j)}_{mk}\xi^{m})|_V + g^{k(i}\Gamma^{j)}_{Ak}\xi^{A}|_V = 0$$

or, denoting $\xi^{i}|_V = \phi^{i}, \xi^{A}|_V = \bar{\xi}^{A}$ and using the definition $\Gamma^{\alpha}_{\beta\gamma} = g^{\alpha\delta}\Gamma_{\beta\gamma\delta}$ it follows that $\Gamma^{j}{}_{Ak}|_V = -\bar{g}^{jl} b_{klA}$ so that

$$\phi^{(i;j)} = \bar{g}^{k(i}\bar{g}^{j)l} b_{klA}\bar{\xi}^{A} \tag{14}$$

where now the covariant derivative is calculated with $\bar{g}_{ij}$.
The equation for mixed indices i,A gives

$$(\xi^{i}{}_{,A})|_V = -g_{AM}\bar{g}^{im}(\bar{\xi}^{M}{}_{,m} + \bar{g}^{MN}A_{mNB}\bar{\xi}^{B}), \tag{15}$$

where $A_{mNB} = \Gamma_{mNB}|_V$.
Finally the equation for the indices A,B gives

$$\xi^{(A;B)}(\nabla M)|_V = (g^{C(A}\xi^{B)}{}_{,C} + g^{C(A}\Gamma^{B)}_{CD}\xi^{D})|_V = 0.$$

However, $\Gamma^{B}_{CD}|_V = 0$ so that this equation reduces to

$$\xi^{(A,B)}|_V = 0. \tag{16}$$

As it can be seen these infinitesimal transformations are not independent and this is a consequence of the fact that the subgroups E,G are not invariant. The particular transformations which send space-time points to space-time points require that $\bar{\xi}^{A} = 0$ (so that $x^A = 0 \Longrightarrow x'^A = 0$). From (14) it follows that these transformations are the isometries of V. Therefore, the group of isometries of V is the subgroup of E characterized by the condition $\bar{\xi}^{A} = \xi^{A}|_V = 0$. Notice that in general this condition imposes constraints on the parameters of E so that the resulting group of isometries has in general less than 10 parameters, the exception occurring when V has constant curvature. In the general case (i.e., without the condition $\bar{\xi}^{A} = 0$) the equations (16) which correspond to the subgroup G can be easily solved giving

$$\xi^{A} = \theta^{A}{}_{j}(x^i)x^j + \theta^{A}{}_{B}(x^i)x^C, \tag{17}$$

where

$$\theta^{(A,B)}(x^i) = g^{B(\alpha}\theta^{A)}{}_{\alpha}|_V = 0.$$

Therefore, G consists of pseudorotations in the subspace of $M(m,n)$ orthogonal to V together with the transformations with parameters $\theta^{A}{}_{j}(x^i)$. These last transformations are again consequence of the fact that G and E are not invariant subgroups of S. Therefore, while the subgroup E has to do with space-time transformations, the subgroup G behaves like a local internal (or gauge) group. In the general situation they cannot be separated.

To see that E satisfies the first condition of Segal consider now the flat limit of V. Under this limit the Lie algebra of S,(11), becomes

$$\begin{aligned}
&[\mathring{l}_{ij},\mathring{l}_{kl}] = \mathring{g}_{ik}\mathring{l}_{jl} + \mathring{g}_{jl}\mathring{l}_{ik} - \mathring{g}_{il}\mathring{l}_{jk} - \mathring{g}_{jk}\mathring{l}_{il},\\
&[\mathring{l}_{ij},\mathring{\pi}_{k}] = \mathring{g}_{ik}\mathring{\pi}_{j} - \mathring{g}_{jk}\mathring{\pi}_{i},\\
&[\mathring{\pi}_{i},\mathring{\pi}_{j}] = 0,\\
&[\mathring{\pi}_{i},l_{BC}] = 0,\\
&[l_{AB},l_{CD}] = g_{AC}l_{BD} + g_{BD}l_{AC} - g_{AD}l_{BC} - g_{BC}l_{AD},
\end{aligned} \tag{18}$$

where "$\circ$" is used to indicate that these operators are calculated in Minkowski space-time. It follows that $E|_{\text{flat}}$ is an invariant subgroup of $S|_{\text{flat}}$ isomorphic to the full Poincaré group. The subgroup $G|_{\text{flat}} = \mathring{G}$ is now completely disconnected from $E|_{\text{flat}} \approx P$ and it contains $(P-4)(P-5)/2$ parameters. Therefore the vanishing of the gravitational field breaks the projected embedding symmetry S into $P \times \mathring{G}$,

$$[P,P] = P, \quad [P,\mathring{G}] = 0, \quad [\mathring{G},\mathring{G}] = \mathring{G}.$$

It is a known fact that the number of dimensions of the embedding space increases as the space-time geometry becomes less symmetric. The dimension $p = 10$ corresponds to the minimum theoretical limit for analytic embeddings (the Janet–Cartan theorem). On the other hand for a differentiable embedding this limit increases to 14 dimensions.[15] Confining the attention to the physically relevant cases, it follows that $\mathring{G}$ can be SO(6) for $p = 10$ and SO(10) for $p = 14$.

3. MASS AND MASS SPLITTING

The verification of the second Segal condition requires an appropriate definition of the mass operator. Such an operator must be capable of providing an adequate mass spectrum and it must reduce to the Poincaré mass operator when E contracts into P. Since E is not an invariant subgroup of S, the invariant operator to be considered as the mass operator should belong to the universal enveloping algebra of S (and not of E). The most obvious candidate is the second-order Casimir operator of S (as it is the case of the Poincaré mass operator). It has also been suggested that the mass formula for mesons can be derived from the second-order Casimir operator of some group containing a noninvariant subgroup contractible into P.[16] Therefore, the mass operator can be defined as

$$M^2 = \left(\frac{k(\rho)}{\rho}\right)^2 l_{\alpha\beta}l^{\alpha\beta}, \quad \rho^2 = \sum g_{AB}\rho^A\rho^B, \tag{19}$$

where $k(\rho)$ is a mass scaling factor such that

$\lim_{\rho\to\infty} k(\rho) = K_0 = \text{const.}$ To analyze the behavior of (19) under the flat limit it can be written as

$$\begin{aligned}
M^2 &= \left(\frac{k}{\rho}\right)^2 (l_{ij}l^{ij} + 2l_{iA}l^{iA} + l_{AB}l^{AB})\\
&= \left(\frac{k}{\rho}\right)^2 l_{ij}l^{ij} + k^2\pi_i\pi^i + k^2\sum\left(\frac{1}{\rho^A\rho_A}\left(\sum_{B\neq A} l_{iB}l^{iB}\right)\right)\\
&\quad - k^2\sum\left(\frac{1}{\rho^A}l_{iA}\left(\sum_{B\neq A}\frac{1}{\rho^B}l_{iB}\right)\right) + \left(\frac{k}{\rho}\right)^2 l_{AB}l^{AB}
\end{aligned}$$

so that in the flat limit of V

$$M^2|_{\text{flat}} = K_0^2\mathring{\pi}_i\mathring{\pi}^i$$

which is the Poincaré mass operator. The constant K_0 is required for expressing the appropriate values of mass in a given mass dimension.

Now consider a representation of S such that it is completely reducible with respect to E and such that M^2 is Hermitian as an operator acting on the representation space H. Then H is a direct sum of the representation spaces H_λ for E and under these conditions the spectrum of M^2 may have a mass-like structure [that is, continuous but with isolated points (Ref. 7)].

Let I be an operator of S such that it contains at least one eigenstate $|b\rangle$ of M^2 in its domain. To obtain a nonzero transition probability $\langle a|I|b\rangle$ between distinct representation spaces of E the operator I should belong to G. Then I is a linear combination of l_{AB},

$$I = a^{AB} l_{AB}.$$

Taking two distinct representations a,b of E, the difference between the eigenvalues of M^2 is

$$m_a^2 - m_b^2 = \langle a|M^2|a\rangle - \langle b|M^2|b\rangle = \frac{\langle a[M^2,I]|b\rangle}{\langle a|I|b\rangle}. \tag{20}$$

The mass operator (19) can also be written as

$$M^2 = \left(\frac{k}{\rho}\right)^2 (g^{\alpha\gamma} g^{\beta\delta} l_{\alpha\beta} l_{\gamma\delta})|_V$$
$$= \left(\frac{k}{\rho}\right)^2 [\bar{g}^{ik}\bar{g}^{jl} l_{ij} l_{kl} + 2\bar{g}^{ij} g^{AB} l_{iA} l_{jB} + g^{AB} g^{CD} l_{AC} l_{BD}]$$

so that

$$[M^2,I] = \left(\frac{k}{\rho}\right)^2 a^{EF} [\bar{g}^{ik}\bar{g}^{jl} C^{\gamma\delta}_{ijEF}\{l_{\gamma\delta},l_{kl}\} + 2\bar{g}^{ij} g^{AB} C^{\gamma\delta}_{iAEF}\{l_{\gamma\delta},l_{jB}\} + g^{AB} g^{CD} C^{\gamma\delta}_{ACEF}\{l_{\gamma\delta},l_{BD}\}]$$

where $\{\ ,\ \}$ denotes the anticommutator and $C^{\rho\sigma}_{\alpha\beta\gamma\delta}$ are the structure constants of the group S in the basis l_{ij},l_{iA},l_{AB} [so that these constants are obtainable from (9)]. Replacing in (20) and noting that $\langle a|E|b\rangle = 0$ for $a\neq b$

$$m_a^2 - m_b^2 = \left(\frac{k}{\rho}\right)^2 a^{EF} [\bar{g}^{ik}\bar{g}^{jl} C^{\gamma\delta}_{ijEF}(s^a_{kl} + s^b_{kl}) + 2\bar{g}^{ij} g^{AB} C^{\gamma\delta}_{iAEF} \times (p^a_{jB} + p^b_{jB}) + g^{AB} g^{CD} C^{\gamma\delta}_{ACEF} \times (I^a_{BD} + I^b_{BD})] \frac{\langle a|l_{\gamma\delta}|b\rangle}{\langle a|I|b\rangle}, \tag{21}$$

where

$$s^a_{kl} = \langle a|l_{kl}|a\rangle,$$
$$p^a_{jB} = \langle a|l_{jB}|a\rangle,$$
$$I^a_{BD} = \langle a|l_{BD}|a\rangle$$

are terms which correspond to spin, momentum, and internal parameters, respectively.

The expression (21) can be further simplified. Since $l_{\alpha\beta} = Z^{\mu}_{,\alpha} Z^{\nu}_{,\beta} L_{\mu\nu}|_V$ then $C^{\gamma\delta}_{\mu\nu EF}$ (where μ,ν are Cartesian indices) may be defined by

$$C^{\gamma\delta}_{\alpha\beta EF} = Z^{\mu}_{,\alpha} Z^{\nu}_{,\beta} C^{\gamma\delta}_{\mu\nu EF},$$

then

$$m_a^2 - m_b^2 = \left(\frac{k}{\rho}\right)^2 [(f^{\mu\rho} + h^{\mu\rho})(f^{\nu\sigma} + h^{\nu\sigma})]|_V (\langle a|l_{\rho\sigma}|a\rangle + \langle b|l_{\rho\sigma}|b\rangle) a^{EF} \frac{C^{\gamma\delta}_{\mu\nu EF}\langle a|l_{\gamma\delta}|b\rangle}{\langle a|I|b\rangle},$$

where

$$f^{\mu\nu} = g^{ik} Z^{\mu}_{,i} Z^{\nu}_{,k}, \quad h^{\mu\nu} = g^{AB} Z^{\mu}_{,A} Z^{\nu}_{,B}.$$

Now since $|a\rangle$, $|b\rangle$ belong to distinct representations of E, $\langle a|l_{ij}|b\rangle = 0$ and $\langle a|l_{iA}|b\rangle = 0$ for $a\neq b$. Thus

$$\frac{\langle a|a^{EF} C^{\gamma\delta}_{\mu\nu EF} l_{\gamma\delta}|b\rangle}{\langle a|I|b\rangle} = \frac{\langle a|[L_{\mu\nu},I]|b\rangle}{\langle a|I|b\rangle} = \langle a|L_{\mu\nu}|a\rangle - \langle b|L_{\mu\nu}|b\rangle.$$

Therefore, denoting

$$U^{+-}_{\mu\nu}(a,b) = \tfrac{1}{2}(\langle a|L_{\mu\nu}|a\rangle \pm \langle b|L_{\mu\nu}|b\rangle) \tag{22}$$

and noting that

$$(f^{\mu\nu} + h^{\mu\nu})|_V = \eta^{\mu\nu},$$

expression (21) reduces to

$$m_a^2 - m_b^2 = \left(\frac{k}{\rho}\right)^2 \eta^{\mu\rho}\eta^{\nu\sigma} U^+_{\rho\sigma}(a,b) U^-_{\mu\nu}(a,b). \tag{23}$$

Notice that by evaluating the product of $U^+_{\rho\sigma}(a,b) U^-_{\mu\nu}(a,b)$ the initial expression (20) is recovered.

The most interesting aspect of expression (23) is that its space-time dependence rests only on the curvature radius ρ. In fact the functions $U^{+-}_{\mu\nu}(a,b)$ are functions only of the representations of the group SO(m,n). It follows that in the flat limit of V ($\rho\to\infty$), the mass splitting vanishes at the same time that E contracts into P, in accordance with O'Raifeartaigh's theorem. On the other hand for large (but finite) values of ρ, the group contraction is not completed, that is, $[\pi_i,\pi_j]\approx 0$ but $[\pi_i,l_{BC}]\neq 0$ so that even for weak gravitational fields the mass splitting (21) does not vanish. In fact, the expression (21) generalizes a mass splitting previously derived for the de Sitter space-time with the group SO(4,1) regarded as a cosmological symmetry. In this case the internal group is postulated and for $\rho = R = 10^{-28}$ cm the mass splitting has been estimated to be of the order of 10^{-50} (MeV)^2.[8,9]

In the case of local gravitational fields with $p>5$, the internal group emerges from the embedding. It is possible to interpret the group SO(m,n) as a combination of various local de Sitter groups to produce a combined internal and space-time symmetry.[11,12] However, in the present context such geometrical interpretation is irrelevant.

The order of magnitude of the mass splitting depend on the function $k(\rho)$. The resulting values must be comparable to the experimental data. Let K_0 be the constant so that the eigenvalues of the Poincaré mass operator are expressed in (MeV)^2.

$$M_P^2 = K_0^2 \mathring{\pi}_i \mathring{\pi}^i.$$

Since the contraction of (19) must reproduce M_P^2, it follows that $k(\infty) = K_0$, suggesting that $k(\rho)$ may be represented by an asymptotic expansion

$$k(\rho) = \sum_0^\infty \frac{c_n}{\rho^n}, \tag{24}$$

truncated to some finite interger s. This representation of $k(\rho)$ is accurate for large values of ρ, for a relatively small s. The coefficients C_n in (24) may be determined from the experimental data. Assuming that the practically-constant particle masses as measured at the surface of the Earth ($\rho = \rho_\oplus$) is in accordance with (19) and (24), then the coefficients of the asymptotic expansion may be determined from the conditions

$$M(\rho_\oplus) = M_P, \quad \frac{d^r}{d\rho^r} M(\rho)|_{\rho_\oplus} = 0, \quad r = 1,\dots, s.$$

Observe that the identification $M(\rho_\oplus) = M_P$ is numerical only. In terms of $k(\rho)$ these conditions read

$$k(\rho_\oplus) = \rho_\oplus K_0, \quad k'(\rho_\oplus) = K_0, \quad k^{(r)}(\rho_\oplus) = 0 \quad r = 2,\dots, s,$$

so that the s coefficients $C_n, C_0 = K_0$, in (24) can be determined. For any value of s the mass of the particles at the surface of the Earth is

$$M^2(\rho_\oplus) = K_0^2 l_{\alpha\beta} l^{\alpha\beta}$$

so that all the masses are multiples of K_0, and K_0 is the mass corresponding to $l_{\alpha\beta} l^{\alpha\beta} = 1$. As a numerical example, take the spin-zero pion as the particle corresponding to $l_{\alpha\beta} l^{\alpha\beta} \approx 1$. Then

$$M(\rho_\oplus)_{\pi_0} \approx K_0 = m_{\pi_0}$$
$$= (0.134\,962\,59 \pm 0.000\,0039) 10^3 \text{ MeV}.$$

For other values of ρ the expression (24) has to be evaluated with the curvature radius for the Schwarzschild solution given by Ref. 17

$$\rho = \sqrt{\frac{r^3}{2m}},$$

where r is the radial coordinate and $2m$ the Schwarzschild radius. Thus taking the Earth's average radius $r_\oplus = 6.371 \times 10^8$ cm, it follows that $\rho_\oplus = 1.708\,6208 \times 10^{13}$ cm. It is interesting to notice that for $s = 3$, the value of $M(\rho)_{\pi_0}$ for different values of r remains within the limits of exact measurement of m_{π_0} in the near vicinity of the Earth. To find the exact values of the masses, specific representations of SO(m,n) have to be calculated. It is also possible to increase the number of terms in the expansion of $k(\rho)$ so as to meet the accuracy of the experimental verification. For small values of ρ, a high-order asymptotic expansion has to be considered, thus producing particles with large masses associated with strong gravitational fields.

In conclusion, by use of an asymptotic representation of $k(\rho)$ it is always possible to adjust the scale of the mass operator (19) so as to compare with the experimental results. Since in the flat limit of V the Poincaré masses are recovered but not the mass splitting, it appears that the gravitational field plays a significant role in the particle multiplet composition.

ACKNOWLEDGMENTS

The author is indebted to the International Atomic Energy Agency and Unesco for hospitality at the International Center for Theoretical Physics, Trieste, where part of this work was developed.

[1] L. O'Raifeartaigh, Phys. Lett. **14**, 757 (1965).
[2] M. Flato and D. Sternheimer, Phys. Rev. Lett. **15**, 934 (1965).
[3] I. Segal, Duke Math. J. **18**, 221 (1951).
[4] S. Coleman and I. Mandula, Phys. Rev. **159**, 1251 (1967).
[5] M. Flato and D. Sternheimer, C. R. Acad. Sci. Paris, **260**, 3532 (1965).
[6] R. Haag, T. Lopuszanski, and M. Sohnius, Nucl. Phys. B **88**, 257 (1975).
[7] I. Segal, J. Funct. Anal. **1**, 1 (1967).
[8] P. Roman, Nuovo Cimento A **45**, 268 (1966).
[9] P. Roman, *Proceedings of Conference on Noncompact Groups in Particle Physics*, edited by Y. Chow (Benjamin, New York, 1966), p. 89.
[10] W. Tait and J. F. Cornwell, J. Math. Phys. **12**, 1651 (1971).
[11] R. Prasad, Nuovo Cimento **44**, 1 (1966).
[12] A. Araki and S. Okubo, Lett. Nuovo Cimento **3**, 511 (1972).
[13] L. P. Eisenhart, *Riemannian Geometry*, 6th printing (Princeton University, Princeton, N. J., 1966), p. 143 ff.
[14] P. Szekeres, Nuovo Cimento **43**, 1062 (1966).
[15] R. E. Greene, Mem. Amer. Math. Soc. **97**, 1–63 (1970).
[16] A. O. Barut and A. Böhm, Phys. Rev. B **139**, 1107 (1965).
[17] P. G. Bergmann, *The Riddle of Gravitation* (John Murray, London, 1969).

Nonperturbative confinement in quantum chromodynamics. III. Improved gluon propagator

D. Atkinson,[a] H. Boelens,[a] S. J. Hiemstra,[a] P. W. Johnson,[b] W. J. Schoenmaker,[c] and K. Stam[a]
High Energy Physics Division, Argonne National Laboratory, Argonne, Illinois 60439

(Received 18 November 1983; accepted for publication 27 January 1984)

An ansatz is introduced for the three-gluon vertex that is consistent with the Slavnov–Taylor identity in Landau gauge. It is shown that the gluon has a confining infrared singularity; but there is also a tachyon, indicating an insufficiency either of quarkless QCD or at least of our approximation to it.

PACS numbers: 12.35.Cn

I. INTRODUCTION

It is an attractive hypothesis that a severe infrared singularity is a signal of confinement in quarkless quantum chromodynamics (QCD).[1,2] Although the gluon self-energy is gauge-dependent, the gauge-invariant Wilson loop can be constructed from it; and it has been shown that a sufficiently singular propagator leads to an area law in leading order.[3] A proper study of such singularities entails a nonperturbative approximation. As an alternative to the lattice approach, with attendant uncertainties concerning the continuum limit, one may study truncated continuum Dyson–Schwinger (DS) equations, using either covariant or axial gauges. The DS equation for the gluon propagator may be truncated through use of Slavnov–Taylor (ST) identities, parametrizing longitudinal parts of vertex-functions.

In an axial gauge, the gluon field decouples from the ghost field, so that the DS equations and the ST identities have a simple form. If one makes the *ad hoc* assumption that the full propagator has the same tensor structure as the bare propagator, then the scalar function that multiplies it may be obtained as the solution of the scalar equation obtained by projection of the DS equation onto the direction of the axial vector n. This equation does not contain the four-gluon vertex. A disadvantage is that the equation involves the unphysical, gauge-dependent parameter $(pn)^2$, where p is the momentum variable. Furthermore, since the ST identity involves projection onto p, rather than n, there is a certain arbitrariness in the projected DS equation.[4]

In the Landau gauge, the DS equation for the gluon propagator involves a single tensor structure, and hence reduces to a scalar equation. However, both the DS equations and the ST identities involve ghost couplings. We do not expect the ghost fields themselves to produce infrared singularities, so we replace the ghost propagator and ghost–ghost–gluon vertex by bare values. The four-gluon vertex does appear in the scalar equation, but we drop it in the interest of simplicity, not expecting cancellations between three- and four-gluon couplings.

The approximation of Mandelstam[5] in Landau gauge involved the replacement of one internal gluon line and of the three-gluon vertex by free values. Such a replacement is motivated by the form of the ST identity, but is not strictly consistent with it. In our analysis of Mandelstam's equation in Refs. 1 and 2, we confirm that the gluon propagator is of order p^{-4} at small spacelike momenta. However, there are first-sheet branch points in the variable p^2, which accumulate at $p^2 = 0$ in the timelike direction. These singularities are presumably unphysical, and in any event they invalidate the Wick rotation to Euclidean momenta.

In this paper we propose an ansatz for the vertex function that has the same tensor structure as the corresponding bare vertex. The multiplicative scalar function can be chosen so that the ST identity in the external leg is automatically satisfied. The resultant scalar equation, which is generally similar to that obtained in Mandelstam's approach, with, however, a somewhat more intricate structure, is derived in Sec. II. We analyze this equation in Sec. III, and show by methods similar to those of Refs. 1 and 2 that there is a solution which is free from complex branch points. This solution has the infrared asymptote p^{-4}, uniformly in the cut p^2-plane, and is therefore suggestive of confinement.

The resultant gluon propagator also has a simple pole at a real, spacelike momentum. Such a pole does not spoil the Wick rotation, but would imply the existence of an unstable tachyon, if it were taken seriously. Recall that in perturbative QED there is a tachyon (Landau ghost), which would not be expected in perturbative QCD because of the opposite sign in the self-energy. Our nonperturbative tachyon might conceivably be an indication of the insufficiency of quarkless QCD; unless it is merely a deficiency of our approximation scheme. It may be that neglect of the four-gluon vertex has produced an instability of the type familiar in a scalar theory with ϕ^3 interaction.

II. ANSATZ FOR VERTEX FUNCTION

We shall use a consistent notation, in which primes distinguish full from bare propagators and vertex functions. We suppress all color indices, since they only yield a trivial multiplicative factor in the final equation. The full propagator is

$$D'_{\mu\nu}(q) = F(-q^2)D_{\mu\nu}(q) = -q^{-2}F(-q^2)\Delta_{\mu\nu}(q), \tag{2.1}$$

where

$$\Delta_{\mu\nu}(q) = g_{\mu\nu} - q_\mu q_\nu/q^2. \tag{2.2}$$

[a] Permanent address: Institute for Theoretical Physics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands.
[b] Permanent address: Illinois Institute of Technology, Chicago, IL 60616.
[c] Permanent address: Institute for Theoretical Physics, University of Kaiserslautern, D-6750 Kaiserslautern, Germany.

 0022-2488/84/062095-04$02.50

The object is to obtain an equation for the scalar function $F(-q^2)$.

The Slavnov–Taylor identity relating the full three-gluon vertex, Γ', to the propagator[6] is

$$p^\lambda \Gamma'_{\lambda\mu\nu}(p,q,r) D'^{\mu\sigma}(q) D'^{\nu\tau}(r) = G(-p^2) D'^{\mu\sigma}(q) \Delta^{\nu\tau}(r) \tilde{\Gamma}'_{\mu\nu}(p,r;q) - G(-p^2) \Delta^{\mu\sigma}(q) D'^{\nu\tau}(r) \tilde{\Gamma}'_{\mu\nu}(p,q;r) . \quad (2.3)$$

Here $\tilde{\Gamma}'_{\mu\nu}$ is the full ghost–ghost–gluon vertex, and the ghost propagator is $-p^{-2}G(-p^2)$. As discussed in the Introduction, we replace the ghost functions by their bare values, $\tilde{\Gamma}'_{\mu\nu} \to g_{\mu\nu}$ and $G\to 1$. With use of (2.1), we then obtain

$$p^\lambda \Gamma'_{\lambda\mu\nu}(p,q,r) D'^{\mu\sigma}(q) D'^{\nu\tau}(r) = [F(-r^2)/r^2 - F(-q^2)/q^2]\, \Delta^{\mu\sigma}(q) \Delta^{\nu\tau}(r) g_{\mu\nu} . \quad (2.4)$$

Our basic ansatz[7] is to suppose that

$$\Gamma'_{\lambda\mu\nu}(p,q,r) D'^{\mu\sigma}(q) D'^{\nu\tau}(r) = f(p,q,r) \Gamma_{\lambda\mu\nu}(p,q,r) D^{\mu\sigma}(q) D^{\nu\tau}(r) , \quad (2.5)$$

where f is some scalar function that is to be related to F by requiring that the Slavnov–Taylor identity (2.4) is satisfied. Since the bare version of (2.4) is obtained by removing the primes and setting F equal to unity, we find, by contracting both sides of (2.5) against p^λ,

$$\frac{F(-r^2)}{r^2} - \frac{F(-q^2)}{q^2} = f(p,q,r)\left(\frac{1}{r^2} - \frac{1}{q^2}\right), \quad (2.6)$$

so that

$$f(p,q,r) = [r^2 F(-q^2) - q^2 F(-r^2)]/(r^2 - q^2) , \quad (2.7)$$

which can be inserted into (2.5). For comparison, the Mandelstam ansatz, in which the three-gluon vertex and one, but not both, of the gluon propagators are replaced by their bare values, is of the form (2.5), with the function $f(p,q,r)$ equal to $F(-q^2)$. It must be emphasized that this form, unlike (2.7), in inconsistent with the Slavnov–Taylor identity (2.4).

The Dyson–Schwinger equation for the gluon propagator, in which only the three-gluon vertex contribution is retained, with the ansatz (2.5), takes the form

$$D'_{\lambda\rho}(p) = D_{\lambda\rho}(p) = D_{\lambda\mu}(p) \Pi^{\mu\nu}(p) D'_{\nu\rho}(p) , \quad (2.8)$$

where the self-energy is

$$\Pi^{\mu\nu}(p) = \frac{ig^2}{(2\pi)^4}\int d^4q\, f(p,q,r) \Gamma^{\mu\sigma\tau}(p,q,r) \times \Gamma^{\nu\rho\omega}(p,q,r) D_{\sigma\rho}(q) D_{\tau\omega}(r) . \quad (2.9)$$

Contracting both sides of (2.8) by $g^{\lambda\rho}$ and dividing throughout by $F(-p^2)$, we find

$$1/F(-p^2) = 1 + (1/p^2)\Pi(-p^2) , \quad (2.10)$$

where

$$\Pi(-p^2) = \Delta_{\mu\nu}(p) \Pi^{\mu\nu}(p) = \frac{ig^2}{(2\pi)^4}\int \frac{d^4q}{q^2 r^2} f(p,q,r) \Omega(p,q,r) , \quad (2.11)$$

with

$$\Omega(p,q,r) = \Gamma^{\mu\sigma\tau}(p,q,r) \Delta_{\mu\nu}(p) \Delta_{\sigma\rho}(q) \times \Delta_{\tau\omega}(r) \Gamma^{\nu\rho\omega}(p,q,r) . \quad (2.12)$$

We next make a Wick rotation to Euclidean space: $-p^2 \to p^2_{\text{eucl}} \equiv x$. The angular integrations can be performed and we obtain

$$\frac{1}{F(x)} = 1 - \frac{25}{4}\int_0^\infty \frac{dy}{y} F(y) - 9\int_0^\infty \frac{dy}{x} F(y) - \frac{x^2}{8}\int_x^\infty \frac{dy}{y^3} F(y) + \int_0^x \frac{dy}{x}\left[\frac{13}{2}\left(1 - \frac{x}{y}\right) + \frac{7}{4}\left(\frac{x^2}{y^2} - \frac{y}{x}\right) - \frac{1}{4}\frac{y^2}{x^2} + \frac{1}{8}\frac{x^3}{y^3}\right] F(y) + \int_0^{x/4} \frac{dy}{x}\left(6 + \frac{17}{2}\frac{x}{y} - 2\frac{x^2}{y^2} - \frac{1}{8}\frac{x^3}{y^3}\right) \times \left(1 - 4\frac{y}{x}\right)^{1/2} F(y) . \quad (2.13)$$

Here the coupling constant and other multiplicative factors have been scaled away, as in Ref. 1. This form is similar in its general shape to the Mandelstam equation, except for the last term, involving the square root. This comes from principal-value integrals that are engendered by the quotient (2.7).

The infrared analysis of (2.13) is a little more involved than that in Refs. 1 and 2. We replace the unknown function F by ϕ according to

$$F(x) = A/x + Bx + x^3\phi(x) . \quad (2.14)$$

The coefficient of the incipient pole on the rhs of (2.13) is zero if

$$A = -0.3456\int_0^\infty dy\, F(y) , \quad (2.15)$$

while the constant term vanishes if

$$1 = 6.25\int_0^\infty \frac{dy}{y} F(y) . \quad (2.16)$$

Finally, in order to match the linear term on the lhs of (2.14), x/A, we require

$$B = \gamma/A , \quad (2.17)$$

where

$$\gamma = 0.616\,97 . \quad (2.18)$$

These requirements are formal, in the sense that the integrals (2.15) and (2.16) are actually divergent. However, it is worth stressing that the absence of logarithms when the first two terms of (2.14) are inserted into (2.13) is a result of delicate cancellations.

After these manipulations, (2.13) can be cast into the form

$$\frac{\gamma + x^2\phi(x)}{1 + \gamma x^2 + x^4\phi(x)} = -\frac{1}{4}\int_0^x \frac{dy}{x}\left(1 - \frac{y}{x}\right)^3 \phi(y) P\left(\frac{y}{x}\right) + \frac{1}{8}\int_0^{x/4} \frac{dy}{x}\left(1 - \frac{4y}{x}\right)^{3/2} \phi(y) Q\left(\frac{y}{x}\right), \quad (2.19)$$

where

$$P(z) = 1 + 10z + z^2 , \quad (2.20)$$

$$Q(z) = 1 + 20z + 12z^2 . \quad (2.21)$$

No divergences are left, and we propose to study this nonlinear equation for the unknown function $\phi(x)$.

III. ANALYSIS OF EQUATION

In the following, we shall for simplicity replace the polynomials P and Q of (2.20) and (2.21) by constants, in such a way that the threshold value ($x \to 0$) of each integral is unchanged. The averaged equation reads

$$\frac{\gamma + x^2\phi(x)}{1 + \gamma x^2 + x^4\phi(x)} = -\frac{23}{30}\int_0^x \frac{dy}{x}\left(1 - \frac{y}{x}\right)^3 \phi(y) + \frac{53}{168}\int_0^{x/4} \frac{dy}{x}\left(1 - \frac{4y}{x}\right)^{3/2} \phi(y). \tag{3.1}$$

Our previous experience[1,2] leads us to expect that this averaging procedure will only have minor quantitative, but not qualitative effects on the solution.

To examine the nature of the infrared singularity of $\phi(x)$, we linearize the lhs of (3.1), retaining the terms $\gamma - \gamma^2x^2 + x^2\phi(x)$ only. Even with this linearization, we have been unable to give a complete analysis; but there are reasons for expecting the second integral on the rhs of (3.1) to be nondominant in the infrared (see the Appendix). Accordingly, we shall study the linear equation

$$\gamma - \gamma^2x^2 + x^2\phi(x) = -\frac{23}{30}\int_0^x \frac{dy}{x}\left(1 - \frac{y}{x}\right)^3 \phi(y). \tag{3.2}$$

The corresponding homogeneous equation

$$x^6\Psi(x) = -\frac{23}{30}\int_0^x dy(x-y)^3\Psi(y) \tag{3.3}$$

has solutions expressible in terms of Bessel and Neumann functions. The four independent solutions of the corresponding differential equation have the small-x asymptotic behavior

$$\Psi_i(x) \sim x^{-15/4}\exp(73.6x^{-1/2}z_i), \tag{3.4}$$

where z_i are the four fourth roots of -1. These functions may be used to solve the inhomogeneous Eq. (3.2), using variation of parameters on the corresponding differential equation. Each of the homogeneous solutions Ψ_i becomes unbounded in certain sectors of the plane, cut along $-\infty < x < 0$, as x tends to zero. However, we find that there is one (and only one) function $\phi(x)$ which satisfies the integral Eq. (3.2) in the cut x-plane. That function has the following asymptotic behavior as x approaches zero within the cut plane:

$$\phi(x) \sim \gamma x^{-15/4}\sum_{i=1}^{4} z_i \exp(\kappa x^{-1/2} z_i) \times \Gamma(-\tfrac{13}{2}, \kappa x^{-1/2} z_i), \tag{3.5}$$

where κ is a positive number and the incomplete gamma function is given by

$$\Gamma(\alpha,\omega) = \int_\omega^\infty dy\, y^{\alpha-1}e^{-y}. \tag{3.6}$$

It follows from a standard asymptotic expression for $\Gamma(\alpha,\omega)$ that $\phi(x)$ approaches a constant value as x tends to zero throughout the cut plane.

We wish to emphasize that uniform boundedness of ϕ at small x is a significant improvement over the corresponding infrared behavior obtained for solutions of Mandelstam's equation. In the latter case, we obtained asymptotic behavior $x^{-7/2}$ for the corresponding function as x tends to zero on the left-hand cut. In the present case, the linear approximation (3.2) remains under control at small x, even on the left-hand cut. As a consequence, we expect that the solutions of the nonlinear equation will be analytic in the cut plane, at least in the infrared. In the case of Mandelstam's equation, the linear approximation was out of control near the left-hand cut, and manifestation of this was the accumulation of first-sheet branch points of the full nonlinear equation.

We have done an extensive numerical study of Eq. (3.1), with the second integral omitted. One may obtain an asymptotic power series for $\phi(x)$ at small x directly from the nonlinear integral equation, which is used for computation of ϕ at small real x. The solution is then obtained at larger values of x by Runge–Kutta integration of the differential equation

$$[x^4G(x)]^{iv} = -\tfrac{23}{5}\phi(x), \tag{3.7}$$

where

$$G(x) = [\gamma + x^2\phi(x)]/[1 + \gamma x^2 + x^4\phi(x)]. \tag{3.8}$$

Using the techniques of Refs. 1 and 2, we have continued $\phi(x)$ into the complex plane, and have not found any branch points on the first Riemann sheet, except at the infrared point, $x = 0$. On penetrating the cut along the timelike axis, $-\infty < x < 0$, however, we have picked up two branch points at $-0.3782 - 0.2239i$ and $-0.0241 - 0.0783i$. It is a reasonable guess that more exist, probably accumulating at the origin on the second or higher Riemann sheets. The fact that the complex branch points, which were on the first sheet in the Mandelstam approximation, are now on secondary sheets, where they cause no trouble, is a definite improvement. The function $F(x)$ [cf. Eq. (2.14)] does indeed have the infrared asymptote

$$F(x) \sim A/x + Bx + Cx^3, \tag{3.9}$$

as $x \to 0$ in any direction on the first sheet.

When we make an analytic continuation of ϕ to larger x, we find a pole on the real axis at $x = 2.1853$. The pole in $\phi(x)$ may be located directly as a zero of the function $1 - x^2G(x)$. Fortunately, the pole does not interfere with the Wick rotation to the Euclidean region, in contrast to the complex singularities found in Refs. 1 and 2. The pole occurs some distance from the infrared, where our approximation scheme is no longer necessarily good. Further details of the numerical work may be found in Ref. 8.

ACKNOWLEDGMENTS

Two of us (D.A. and P.W. J.) would like to thank A. White for the hospitality accorded to them in the High Energy Physics division of Argonne National Laboratory, where this paper was completed.

W. J. Schoenmaker and K. Stam have carried out this work as scientific members of the Stichting F.O.M. (Foundation for Fundamental Research of Matter), which is financially supported by Z.W.O. (Netherlands Organization for Pure Scientific Research). This work was supported by National Science Foundation Grant No. PHY-83-05283; and was performed under the auspices of the United States Department of Energy under Contract No. W-31-109-ENG-38.

APPENDIX: TECHNICAL DETAIL

We shall justify neglecting the second term on the right side of Eq. (3.1). The first remark is that, in the asymptotic power series expansion of that equation, the contribution from the first term to the coefficient of x^{2n} dominates that of the second, except at the first few values of n. As a consequence, the asymptotic series $\phi(x)$ is, in effect, controlled by the first term.

Let us consider the linearized, homogeneous version of that equation,

$$x^6\Psi(x) = -\frac{23}{30}\int_0^x dy(x-y)^3\Psi(y) + \frac{53}{168}x^3\int_0^{x/4} dy\left(1-\frac{4y}{x}\right)^{3/2}\Psi(y). \tag{A1}$$

We substitute the solutions $\Psi_i(x)$ [Eq. (3.4)] into (A1), and note that, for x small within a sector of boundedness, the second term in (A1) is asymptotically small compared with the first. Consequently, these functions $\Psi_i(x)$ are asymptotic solutions of Eq. (A1) at small x.

Let us consider an extreme situation in which the first term in (A1) is dropped, and let us replace the "fractional $\frac{5}{2}$-power derivative" on the right by $\Psi(z)/z^{5/2}$, to obtain

$$x^2\Psi(x) = \lambda\Psi(x/4). \tag{A2}$$

The solution to the difference Eq. (A2),

$$\Psi(x) = \exp\left[-\frac{1}{2\ln 2}\ln^2 x + \left(\frac{\ln\lambda}{2\ln 2} - 1\right)\ln x\right], \tag{A3}$$

decreases more rapidly than any power of x at small x, although less rapidly than the functions $\Psi_i(x)$ [Eq. (3.4)]. The asymptotic solution of (A1) with the first term dropped is essentially identical to (A3). In particular, it is analytic and uniformly small throughout the cut plane at small x, so that its infrared behavior is less quixotic than that of $\Psi_i(x)$.

In terms of the Mellin transform,

$$\widetilde{\Psi}(p) = \int_0^\infty dx\,\Psi(x)x^{p-1}, \tag{A4}$$

Eq. (A.1) becomes an algebraic difference equation

$$\widetilde{\Psi}(p+2) = \left[-\frac{23}{30}B(-p+1,4) + \frac{53}{42}4^p B\left(-p+1,\frac{5}{2}\right)\right]\widetilde{\Psi}(p), \tag{A5}$$

where $B(\,,\,)$ is the beta function. We have not been able to solve (A5), subject to the requirement that $\widetilde{\Psi}(p)$ be analytic in a vertical strip of width at least 2. However, we can prove that solutions exist, and that they behave as $\Psi_i(x)$ at small x.

The above arguments make it clear that the first integral in (A1) dominates the second in the infrared. Away from $x=0$, however, the latter integral is not necessarily small.

[1]D. Atkinson, J. K. Drohm, P. W. Johnson, and K. Stam, J. Math. Phys. **22**, 2704 (1981).
[2]D. Atkinson, P. W. Johnson, and K. Stam, J. Math. Phys. **23**, 1917 (1982).
[3]G. B. West, Phys. Lett. B **115**, 468 (1982).
[4]E. J. Gardner, J. Phys. G: Nucl. Phys. **9**, 139 (1983); B. K. Jennings and R. M. Woloshyn, J. Phys. G: Nucl. Phys. **9**, 997 (1983); D. Atkinson, P. W. Johnson, W. J. Schoenmaker, and H. A. Slim, in Nuovo Cimento A **77**, 197 (1983).
[5]S. Mandelstam, Phys. Rev. D **20**, 3223 (1979).
[6]S. K. Kim and M. Baker, Nucl. Phys. B **164**, 152 (1980).
[7]W. J. Schoenmaker, thesis, University of Groningen, 1983.
[8]H. Boelens, Groningen internal report No. 193, 1983.

Binding limit in the Hartree approximation

Mary Beth Ruskai[a)] and Frank H. Stillinger
Bell Laboratories, Murray Hill, New Jersey 07974

(Received 8 November 1983; accepted for publication 23 December 1983)

We show that the Hartree approximation cannot predict that H^- has a bound state, i.e., the Hartree energy is greater than -0.5. We also show that the Hartree approximation cannot predict binding for the Coulomb model of a two-electron atom unless the nuclear charge Z is greater than 1.03 and we compute accurate upper and lower bounds to the Hartree energy for H^- and He.

PACS numbers: 31.10. + z

I. INTRODUCTION

It is well known[1-4] that H^- has a single bound state of even spatial parity, i.e., the Hamiltonian (in reduced units)

$$H(Z) = -\frac{1}{2}\Delta_1 - \frac{1}{2}\Delta_2 - \frac{Z}{r_1} - \frac{Z}{r_2} + \frac{1}{r_{12}} \tag{1.1}$$

has exactly one eigenvalue below $-\frac{1}{2}$ when $Z=1$. Although it is easy to produce trial functions which establish that $H(1) < -\frac{1}{2}$, we are not aware of any such functions which have the form of a single-determinant Hartree–Fock function. In this paper we show that, in fact, one cannot predict binding in H^- by using a spin-restricted form of the Hartree–Fock approximation which, for two electrons, is equivalent to the Hartree approximation.

Because confusion regarding terminology exists in the literature, we find it prudent to review some elementary definitions, so that we can state our results unambiguously. In the Hartree–Fock (HF) approximation, the exact ground state function Ψ for an N-particle Hamiltonian H_N is replaced by a trial function $\Psi_{\rm HF}$ consisting of a single Slater determinant, i.e., an antisymmetrized product of N single-particle functions, $\{\phi_j\}_{j=1}^N$. The Hartree–Fock energy is then given by

$$E_{\rm HF} = \inf_{\Psi_{\rm HF}} \frac{\langle \Psi_{\rm HF}, H_N \Psi_{\rm HF}\rangle}{\langle \Psi_{\rm HF}, \Psi_{\rm HF}\rangle}, \tag{1.2}$$

which leads to a set of equations for $\{\phi_j\}$. In practice, the minimization in (1.2) is carried out using only a finite and restricted set of $\{\phi_j\}$. One common restriction, particularly when $N=2n$ is even, which we shall call the spin-restricted HF (SRHF) approximation, is to require that

$$\phi_{2k-1} = u_k(\mathbf{r})\alpha; \quad \phi_{2k} = u_k(\mathbf{r})\beta \tag{1.3}$$

(where α,β are the usual spin eigenfunctions) and to choose $\{u_j\}_{j=1}^n$ so as to minimize (1.2). If $N=2$, $\Psi_{\rm SRHF}$ has the exceedingly simple form $(\alpha\beta-\beta\alpha)\times u(\mathbf{r}_1)u(\mathbf{r}_2)$. The equation for u obtained from minimizing (1.2) is equivalent to that obtained from the so-called Hartree approximation, which we discuss below. Therefore, our results on the Hartree approximation for two-electron Hamiltonians of the form (1.1), also apply to the SRHF approximation.

In principle, no restrictions on the $\{\phi_j\}$ are necessary. Indeed ϕ_j need not even have the form (space function) $\times$ (spin function), but could conceivably be something like $\phi_j = f_j(r)\alpha + g_j(r)\beta$. When the restrictions (1.3) are dropped, one sometimes emphasizes this by applying the term unrestricted HF (UHF) approximation. For two electrons one can make $E_{\rm UHF}(Z)$ arbitrarily close to $-\frac{1}{2}Z^2$ by choosing $\phi_1 = f_1(\mathbf{r})\alpha$; $\phi_2 = f_n(\mathbf{r})\beta$, where f_k is any hydrogenic function satisfying $(-\frac{1}{2}\Delta - Z/r)f_k = -(Z^2/2k^2)f_k$, and making n large. Therefore, either the UHF approximation predicts binding for H^- or $E_{\rm UHF} = -\frac{1}{2}$ exactly. No approximate lower bound procedure can prove absence of binding for H^- in the UHF approximation. However, we know of no trial function which actually demonstrates that $E_{\rm UHF}(1) < -\frac{1}{2}$.

Because single determinants, particularly those which minimize (1.2), need not be eigenfunctions of spin or orbital angular momentum operators, the HF approximation is sometimes generalized[5] to allow Ψ to be a linear combination of the minimal number of Slater determinants needed to make Ψ an eigenfunction of some specified set of angular momentum operators. An example for two electrons is the function

$$\Psi = [f(\mathbf{r}_1)g(\mathbf{r}_2) + g(\mathbf{r}_1)f(\mathbf{r}_2)]\,[\alpha(1)\beta(2) - \beta(1)\alpha(2)].$$

It has been shown that functions of this form do predict binding in H^-.[4,6] If one takes $f = e^{-ar}$, $g = e^{-br}$ with $a = 1.03925$, $b = 0.2831$ one finds that $H(1) \leqslant -0.5133 < -0.5$. Although this "spatial permanent" is reminiscent of a HF function, the f and g do not satisfy the HF equations and Ψ is not a single Slater determinant; it is actually a linear combination of two Slater determinants, one from $\phi_1 = f(r)\alpha$, $\phi_2 = g(r)\beta$ and the other from $\phi_3 = g(r)\alpha$, $\phi_4 = f(r)\beta$.

In this paper we study the Hartree, or equivalently the SRHF, approximation for two-electron Hamiltonians of the form (1.1). The Hartree energy is given by

$$E_{\rm H}(Z) = 2\inf_{\|u\|=1} \Phi(u), \tag{1.4}$$

where

$$\Phi(u) = \frac{1}{2}\int |\nabla u|^2\, d^3r - Z\int \frac{|u|^2}{r}\, d^3r + \frac{1}{2}\iint \frac{u^2(\mathbf{r})u^2(\mathbf{s})}{|\mathbf{r}-\mathbf{s}|}\, d^3r\, d^3s. \tag{1.5}$$

It is well-known that the continuous spectrum for $H(Z) = [-Z^2/2, \infty)$ (see Ref. 7). We will say that the Hartree approximation predicts binding if $E_{\rm H}(Z) < -Z^2/2$. We

[a)] On leave from Department of Mathematics, University of Lowell, Lowell, MA 01854. Present address: The Bunting Institute, Radcliffe College, Cambridge, MA 02138.

will show that the Hartree approximation does not predict binding for H^- or for any $H(Z)$ with $Z \leqslant 1.0268$. We define Z_H by $E_H(Z_H) = -Z_H^2/2$ and show that

$$1.0268 < Z_H < 1.0312.$$

The analogous quantity for the exact ground state energy, $E_0(Z)$ of $H(Z)$ is defined by $-Z_C^2/2 = -E_0(Z_C)$. Stillinger[6,8] calculated $Z_C \approx 0.9112$. Both the exact and the Hartree problems have the unusual feature that they have solutions at the crossing point, i.e., $H(Z)$ does have a square-integrable ground state for $Z = Z_C$,[9] and (1.5) has a minimizing u for $Z = Z_H$.[10]

The question of whether or not the infimum in (1.4) is actually a minimum has been studied extensively.[10–17] By making the transformation

$$\rho(x) = [u(x/Z)/Z^2]^2,$$

one can consider the Hartree functional $\Phi(u)$ to be a special case of the Thomas–Fermi–von Weizsäcker (TFW) functional for which the constant γ is 0 in the term $\gamma \int \rho^{5/3} = \gamma \int u^{10/3}$. Then one can use results of Benguria and Lieb[18,19] to show that there is a Z_M satisfying (1) for $Z < Z_M$, (1.4) has no minimizing u and $E_H(Z) = -Z^3 A$ for some $A > 0$; (2) for $Z \geqslant Z_M$, (1.4) has a minimizing u and $E_H(Z) \geqslant -Z^3 A$ with equality only for $Z = Z_M$; and (3) $\frac{1}{2} < Z_M < 1$. Recently Baumgartner[20] has shown that $Z_M \approx 0.828$. It is interesting to note that, since $Z_M < 1 < 1.0268 < Z_H$, there is a nontrivial range of Z, i.e., (Z_M, Z_H), within which the Hartree functional $\Phi(u)$ has a minimizing u yet the Hartree energy lies in the continuum for $H(Z)$. The region (Z_M, Z_H) includes the physically interesting case of H^- for which $Z = 1$.

Figure 1 schematically indicates the behavior of the energy curves $E_0(Z)$ and $E_H(Z)$, with approximate placement of the distinguished points Z_M, Z_C, and Z_H.

We compute lower bounds to both $E_H(Z)$ and Z_H by using a modification of the Bazley and Seydel method[21] as described in Sec. II. Excellent upper bounds to both $E_H(Z)$ and Z_H were obtained by approximating the Hartree trial function u by a linear combination of simple exponentials; this is discussed in Sec. III. In Sec. IV, we summarize our results for H^- $(Z = 1)$ and He $(Z = 2)$ which are generally better than results reported previously in the literature.

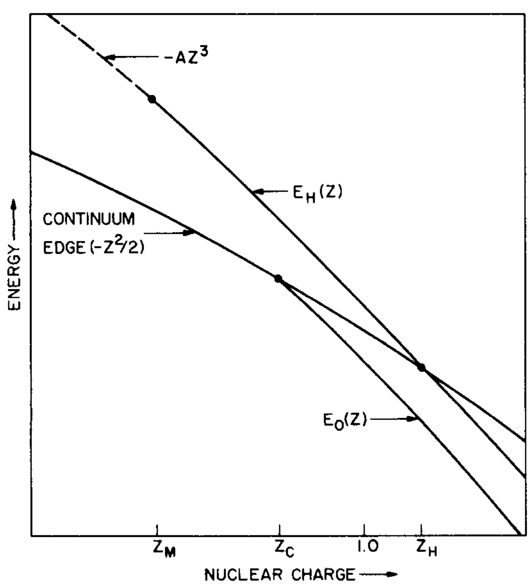


FIG. 1. Schematic diagram of Z-dependent energies for the two-electron problem.

II. LOWER BOUNDS

To obtain a lower bound to $E_H(Z)$, we use the method of Bazley and Seydel[21] which is based upon the inequality

$$\iint \frac{u^2(\mathbf{r})u^2(\mathbf{s})}{|\mathbf{r}-\mathbf{s}|}\, d^3r\, d^3s \geqslant 2\iint \frac{u^2(\mathbf{r})w^2(\mathbf{s})}{|\mathbf{r}-\mathbf{s}|}\, d^3r\, d^3s - \iint \frac{w^2(\mathbf{r})w^2(\mathbf{s})}{|\mathbf{r}-\mathbf{s}|}\, d^3r\, d^3s. \tag{2.1}$$

Substituting (2.1) into (1.5) one finds

$$\Phi(u) \geqslant k_w u - \frac{1}{2}\iint \frac{w^2(\mathbf{r})w^2(\mathbf{s})}{|\mathbf{r}-\mathbf{s}|}\, d^3r\, d^3s, \tag{2.2}$$

where

$$k_w = -\tfrac{1}{2}\Delta - Z/r + q_w(r), \tag{2.3}$$

and

$$q_w(r) = \int \frac{w^2(\mathbf{s})}{|\mathbf{r}-\mathbf{s}|}\, d^3\mathbf{s}. \tag{2.4}$$

Then

$$E_H(Z) \geqslant 2\lambda - \iint \frac{w^2(\mathbf{r})w^2(\mathbf{s})}{|\mathbf{r}-\mathbf{s}|}\, d^3r\, d^3s, \tag{2.5}$$

where λ is a lower bound to k_w. In previous applications of the Bazley–Seydel method,[22] attention was restricted to w for which q_w could be calculated from w by doing the integration in (2.4) exactly, after which one was still faced with the problem of finding a good lower bound to k_w.

Therefore, instead of selecting a trial w, we will choose q_w so that the lowest eigenvalue to k_w can be determined exactly; this eigenvalue will then serve as our lower bound λ in (2.5). We then use the electrostatic equality

$$\iint \frac{w^2(\mathbf{r})w^2(\mathbf{s})}{|\mathbf{r}-\mathbf{s}|}\, d^3r\, d^3s = \frac{1}{4\pi}\int |\nabla q_w|^2\, d^3r$$

to conclude that

$$E_H(Z) \geqslant 2\lambda - \frac{1}{4\pi}\int |\nabla q_w|^2\, d^3r. \tag{2.6}$$

We now choose q_w as follows. Let

$$u = \sum_{k=1}^{N} A_k e^{-a_k r}, \tag{2.7}$$

where $A_k > 0$, $a_{k+1} > a_k > 0$, and

$$\sum_k A_k (Z - a_k) = 0. \tag{2.8}$$

Let

$$q_w(r) = \frac{Z}{r} + \frac{1}{u}\left(\sum_{k=1}^{N-1} \frac{a_k^2 - a_N^2}{2} A_k e^{-a_k r} - \frac{1}{r}\sum_{k=1}^{N} a_k A_k e^{-a_k r} \right). \tag{2.9}$$

Then $k_w u = \lambda u$ with $\lambda = -a_N^2/2$. The constraint $A_k > 0$ insures[23] that $u(r)$ is strictly positive so that λ is the lowest eigenvalue of k_w, and q_w has no singularities, except possibly at $r = 0$. Condition (2.8) guarantees that u satisfies the nuclear cusp condition[24] and consequently that q_w has no pole at $r = 0$. Although we need not compute w explicitly, it can be verified that $w = \sqrt{-\Delta q_w}$ is square-integrable, so that q_w can be obtained from an acceptable Hartree trial function. Since q_w is spherically symmetric,

$$E_{\rm H}(Z) \geqslant -a_N^2 - \int_0^\infty r^2 |q_w'(r)|^2\, dr. \tag{2.10}$$

If one now makes the substitutions

$$r = y/\epsilon Z,$$
$$a_k = Z(1-\gamma_k \epsilon) \quad (k = 1 \ldots N),$$

and lets $\gamma_1 = -1$ so that $a_1 = Z(1+\epsilon)$ defines ϵ, one finds

$$\begin{aligned} q_w(r) &= \epsilon Z^2 \hat{q}_w(y) \\ &= \frac{\epsilon Z^2}{\hat{u}(y)} \sum_k \left(\frac{\epsilon(\gamma_k^2 - \gamma_N^2)}{2} - (\gamma_k - \gamma_N) + \frac{\epsilon}{y}\gamma_k \right) A_k e^{\gamma_k y} \\ &= \frac{\epsilon Z^2}{\hat{u}} \left(\frac{\epsilon \hat{u}''}{2} - \hat{u}' + \frac{\epsilon}{y} \hat{u}' \right) + \text{const}, \end{aligned}$$

where $\hat{u}(y) = \Sigma_k A_k e^{\gamma_k y}$. Then

$$\int_0^\infty r^2 |q'(r)|^2\, dr = \epsilon Z^3 \int_0^\infty \frac{1}{\hat{u}^4(y)} [\epsilon S(y) + y T(y)]^2\, dy, \tag{2.11}$$

where

$$\begin{aligned} T(y) &= (\hat{u}')^2 - \hat{u}\hat{u}'' \\ &= \sum_k \sum_l A_k A_l (\gamma_k \gamma_l - \gamma_l^2) e^{(\gamma_k + \gamma_l)y}, \end{aligned}$$

and

$$\begin{aligned} S(y) + T(y) &= \frac{y}{2}(\hat{u}\hat{u}''' - \hat{u}'\hat{u}'') - \frac{\hat{u}\hat{u}'}{y} \\ &= \sum_k \sum_l A_k A_l \left(\frac{\gamma_l^2 y}{2}(\gamma_l - \gamma_k) - \frac{\gamma_l}{y} \right) e^{(\gamma_k + \gamma_l)y}. \end{aligned}$$

Thus $E_{\rm H}(Z) \geqslant F(Z)$, where

$$\begin{aligned} F(Z) &= -(1-\gamma_N \epsilon)^2 Z^2 - \epsilon Z^3 (A\epsilon^2 + B\epsilon + C), \\ A &= \int_0^\infty \left(\frac{S(y)}{\hat{u}^2(y)} \right)^2 dy, \\ B &= \int_0^\infty \frac{y S(y) T(y)}{\hat{u}^4(y)}\, dy, \\ C &= \int_0^\infty \left(\frac{y T(y)}{\hat{u}^2(y)} \right)^2 dy. \end{aligned} \tag{2.12}$$

The quantity $F(Z)$ gives a lower bound to $E_{\rm H}(Z)$ for all choices of $A_1 \cdots A_{N-1}$, ϵ, and $\gamma_2 \cdots \gamma_N$. [Here A_N is determined by (2.8) and $\gamma_1 = 1$.] Since $F(Z)$ is independent of normalization, we can set $A_1 = 1$ without loss of generality. We now search for values of the remaining parameters which optimize $F(Z)$. For each fixed choice of $A_2 \cdots A_{N-1}$, and $\gamma_2 \cdots \gamma_{N-1}$, the optimal ϵ must satisfy

$$-2\gamma_N + 2\gamma_N^2 \epsilon + Z(3A\epsilon^2 + 2B\epsilon + C) = 0. \tag{2.13}$$

It can be shown that only one of the roots of (2.13) is positive and this positive root maximizes $F(Z)$. A simple linear variation of $A_2 \cdots A_{N-1}$ and $\gamma_2 \cdots \gamma_{N-1}$ can then be used to maximize $F(Z)$ to obtain the best possible lower bound of the form (2.7) for fixed N and Z.

Although one could bound $Z_{\rm H}$ by comparing $F(Z)$ with $-Z^2/2$, we used the following direct approach. Suppose $F(\tilde{Z}) = -\tilde{Z}^2/2$ then $E_{\rm H}(\tilde{Z}) \geqslant -\tilde{Z}^2/2$, which implies $\tilde{Z} < Z_{\rm H}$. To find $\tilde{Z}$ we note that

$$+\tfrac{1}{2} - (1-\gamma_N \epsilon)^2 = \tilde{Z}\epsilon(A\epsilon^2 + B\epsilon + C). \tag{2.14}$$

One can eliminate $\tilde{Z}$ from (2.13) and (2.14) to see that the optimal ϵ for $\tilde{Z}$ must satisfy the quartic equation

$$\gamma_N^2 A\epsilon^4 - 4\gamma_N A\epsilon^3 + (3A/2 - 2\gamma_N B - \gamma_N^2 C)\epsilon^2 + B\epsilon + C/2 = 0. \tag{2.15}$$

Although (2.15) can have four roots, we found that, in practice, it has only one real root in the acceptable range $1 > \gamma_N \epsilon > 1 - 1/\sqrt{2} > 0.29$ [where the second inequality follows from the condition $-Z^2/2 < F(Z) < -(1-\gamma_N\epsilon)Z^2$]. Once ϵ is fixed, $\tilde{Z}$ can be found from either (2.13) or (2.14). A simple variation of $A_2 \cdots A_{N-1}$ and $\gamma_2 \cdots \gamma_{N-1}$ can then be used to maximize $\tilde{Z}$.

Our results are summarized in Table I. Three exponentials suffice to show that $Z_{\rm H} > 1.0$ which implies that the Hartree approximation cannot predict binding in H^-. Four exponentials give $Z_{\rm H} > 1.02683$ which is remarkably close to the upper bound (next section) of 1.031 18. In calculating $\tilde{Z}$, no attempt was made to estimate numerical errors during the variation process itself. Instead, a double precision lower bound was computed for $E_{\rm H}$ (1.026 831) with careful error estimates. These results, namely $E_{\rm H} > -0.527\,190\,945 > -0.527\,190\,951 = -(1.026\,831)^2/2$, verify that $Z_{\rm H} > 1.026\,831$ is indeed a valid bound.

III. UPPER BOUNDS

To obtain upper bounds to $E_{\rm H}$, we again approximate u by a finite linear combination of exponentials. Thus $E_{\rm H} \leqslant 2\Phi(u)$ with u given by (2.7) except that the A_k are required to satisfy the normalization condition, $\Sigma_k \Sigma_l A_k A_l/(a_k + a_l)^3 = 1$, rather than the cusp condition (2.8). If we make the substitutions $t = a_1 r$, $a_j = \gamma_j a_1$ $(j = 1 \cdots N)$ and let a denote a_1, then $2\Phi(u)$ can be written in the form

$$2\Phi(u) = G(a,Z) = a^2 P - 2aZQ + aR. \tag{3.1}$$

Here P, Q, and R are elementary integrals of exponentials which can be evaluated exactly as algebraic combinations of the A_k's and γ_k's. For fixed Z, $G(a,Z)$ has its minimum at $a = (ZQ - R/2)/P$, so that

$$\begin{aligned} E_{\rm H} &\leqslant G(a_{\min}, Z) \\ &= -(ZQ - R/2)^2/P. \end{aligned} \tag{3.2}$$

A linear variation of $A_2 \cdots A_N$, and $\gamma_2 \cdots \gamma_N$ was used to optimize (3.2). Results are summarized in Table II. The quantity $C = \Sigma_k a_k A_k/(Z\, \Sigma_k A_k)$ was calculated as a measure of how closely the cusp condition (2.8) was satisfied.

An upper bound, $Z_{\rm H}^u$, to $Z_{\rm H}$ can be obtained from the condition

$$-(Z_{\rm H}^u)^2/2 = -(Z_{\rm H}^u Q - R/2)^2/P. \tag{3.3}$$

This has solutions

TABLE I. Bounds on Z_H.

N	Type	Z_H	$E_H(Z_H)$	Cusp
1	up	1.066 942	− 0.569 182	0.707
1[a]	lo	0.0	0.0	1.0
2	up	1.031 403	− 0.531 896	0.970
2	lo	0.951 550	− 0.452 724	1.0
3	up	1.031 178	− 0.531 664	1.006
3	lo	1.016 472	− 0.516 608	1.0
4	up	1.031 178	− 0.531 664	1.008
4	lo	1.026 831	− 0.527 191	1.0

[a] Lower bounds with $N=1$ are included here and in Tables II and IV for comparison purposes only. The methods of Sec. II are not strictly valid, since formal application of (2.7) to (2.10) implies $w=0$ when $N=1$.

$$Z_H^{u} = R/(2Q \mp \sqrt{2P}).$$

However, the smaller solution gives a negative expression for $a_{\min}$ which corresponds to a non-square-integrable u. Therefore, we conclude that

$$Z_H \leqslant R/(2Q + \sqrt{2P}).$$

Our results are summarized in Table I. A single exponential has no parameters except a and gives the remarkably good result $Z_H < \frac{3}{8}(1 + 1/\sqrt{2}) < 1.067$. This approximation to Z_H was then used as the first value of Z in the cycle of two-exponential calculations, and the predicted value for Z_H was used for each subsequent variation. This procedure rapidly converged to the bound $Z_H < 1.031\,178$. When combined with our previous lower bound results, we can state with confidence that $Z_H = 1.03$ to three significant figures and $1.0268 \leqslant Z_H < 1.0312$.

IV. H^- AND He

Upper and lower bounds on the Hartree energy for $H^-(Z=1)$, and He$(Z=2)$, were carried out using the procedures described in Secs. II and III. The results are summarized in Table II. We find $-0.489\,651 \leqslant E_H(1) \leqslant -0.487\,929$ for H^-, and $-2.864\,674 \leqslant E_H(2) \leqslant -2.861\,679$ for He.

When compared with previously published bounds, these results are quite gratifying. The difference between the upper and lower bounds for both H^- and He is less than 0.003. We attribute this agreement to the unusually good lower bounds we obtain by circumventing the need to find first a lower bound to k_w, which was constructed instead to be exactly solvable. Our lower bound for He represents an improvement of over 0.018 from the value of $-2.882\,356$ reported by Behling *et al.*[22] We are surprised at the apparent accuracy of upper bounds obtained from a simple linear combination of exponentials. Using only three exponentials we were able to obtain a better upper bound on the Hartree energy for H^- than Froese-Fischer[25] obtained using a linear combination of 11 hydrogenic orbitals. Our upper bound for He is also in remarkably close (10^{-6}) agreement to that reported by Froese-Fischer.[25] A comparison of our bounds to those reported previously is summarized in Table III.

A few comments about the optimizing functions are in order. A list of parameters for the various optimizing functions is given in Table IV. Although the optimal coefficients for the upper and lower bound functions for fixed Z and given N seem rather different, plots of the actual functions indicate that they are pointwise quite close, at least for $N=4$. A comparison of the norm of the difference between the upper- and lower-bound functions shows that they are very close for $N \geqslant 3$. When comparing the upper- and lower-bound functions, it should be kept in mind that one is doing an apples versus oranges type of comparison. The upper bound function is an approximation to the function which minimizes (1.5) and therefore solves the Hartree eigenvalue problem; the lower bound function solves a different, but related, eigenvalue problem, namely $k_w u = (a_N^2/2)u$. However, since both approximations can be expected to converge to the exact Hartree minimizing function, the observed agreement is to be expected.

Our variational procedure found the energy surface to be quite flat; substantial changes in the variation parameters produced insignificant changes in the energy. We attribute this primarily to the fact that first-order changes in the parameters will, in general, produce second- or higher-order

TABLE II. Bounds on Hartree energy for H^- and He.

Z	N	Type	Energy	Cusp	$\|u_{up} - u_{lo}\|$	Overlap
1.0	1	up	− 0.472 656	0.687	0.3194	0.949 003
1.0	1	lo	− 1.0	1.0		
1.0	2	up	− 0.487 824	0.966	0.0759	0.997 116
1.0	2	lo	− 0.517 203	1.0		
1.0	3	up	− 0.487 929 3	1.005	0.0213	0.999 773
1.0	3	lo	− 0.493 691	1.0		
1.0	4	up	− 0.487 929 6	1.007	0.0067	0.999 978
1.0	4	lo	− 0.489 651	1.0		
2.0	1	up	− 2.847 656	0.375	0.1467	0.989 246
2.0	1	lo	− 4.0	1.0		
2.0	2	up	− 2.861 672	1.002	0.0130	0.999 915
2.0	2	lo	− 2.904 016	1.0		
2.0	3	up	− 2.861 679	1.006	0.0027	0.999 996
2.0	3	lo	− 2.870 400	1.0		
2.0	4	up	− 2.861 679		0.0007	0.999 999 7
2.0	4	lo	− 2.864 674	1.0		

TABLE III. Comparison with other energy bounds.

	$Z=1$	$Z=2$
Hartree upper bounds		
3-exponential	− 0.487 929 3	− 2.861 679
4-exponential	− 0.487 929 6	− 2.861 679
Froese-Fischer[a]	− 0.487 927	− 2.861 680
Behling *et al.*[b]	...	− 2.861 58
Hartree lower bounds		
3-exponential	− 0.493 691	− 2.870 400
4-exponential	− 0.489 651	− 2.864 674
Behling *et al.*[b]	...	− 2.882 356
Precise correlated (not HF)		
Pekeris[c]	− 0.527 751	− 2.903 724

[a] Reference 25.
[b] Reference 22.
[c] Reference 3.

TABLE IV. Summary of optimizing parameters.

Z	N	Type	Unnormalized coefficients A_1	A_2	A_3	A_4	Exponents a_1	a_2	a_3	a_4
1.069 9	1	up	1.0				0.687 5[a]			
0.0	1	lo	1.0				1.0[b]			
1.031 403	2	up	1.0	0.529 438			1.268 812	0.493 360		
0.951 550	2	lo	1.0	0.368 683			1.165 260	0.371 892		
1.031 178	3	up	1.0	1.052 094	0.421 745		1.555 878	0.793 467	0.413 478	
1.016 472	3	lo	1.0	0.652 611	0.160 813		1.384 725	0.617 704	0.344 805	
1.031 178	4	up	1.0	3.384 017	3.415 354	1.392 050	1.879 396	1.358 213	0.736 340	0.404 855
1.026 831	4	lo	1.0	0.779 303	0.242 388	0.068 309	1.460 684	0.717 123	0.418 579	0.342 300
1.0	2	up	1.0	0.492 644			1.214 425	0.462 200		
1.0	2	lo	1.0	0.401 462			1.235 018	0.414 595		
1.0	3	up	1.0	1.034 584	0.374 117		1.499 283	0.753 114	0.380 748	
1.0	3	lo	1.0	0.650 161	0.149 785		1.360 014	0.601 520	0.326 114	
1.0	4	up	1.0	1.117 301	0.508 448	0.164 404	1.556 349	0.853 731	0.481 322	0.333 077
1.0	4	lo	1.0	0.780 810	0.228 671	0.055 143	1.419 772	0.689 600	0.390 293	0.311 147
2.0	2	up	1.0	1.640 902			2.906 239	1.452 963		
2.0	2	lo	1.0	1.224 015			2.747 452	1.389 344		
2.0	3	up	1.0	0.730 100	1.419 830		3.074 542	1.307 938	1.627 090	
2.0	3	lo	1.0	0.774 050	0.988 864		2.917 102	1.630 428	1.361 859	
2.0	4	up	1.0	0.299 382	0.875 678	1.071 215	3.095 529	1.705 568	1.696 055	1.347 738
2.0	4	lo	1.0	0.761 538	0.544 252	0.740 025	2.988 910	1.744 292	1.415 324	1.356 822

[a] $= Z - \frac{5}{16}$.
[b] $= Z$.

changes in the energy. The parameters we report should probably be regarded as sufficient to prove our results rather than optimal. It is curious to note that the cusp factor worsens rather than improves as one increases the number of exponentials. Presumably this illustrates the principle that one expects to obtain a lower approximating energy if one does not impose unnecessary constraints on approximating functions, even when the exact solution is known to satisfy these constraints.

ACKNOWLEDGMENTS

One of us (MBR) is grateful to Professor N. Bazley for arranging a visit to the University of Cologne at which time Ref. 21 was brought to her attention.

All calculations were carried out on a Cray-1 computer, using the numerical integration routines in the PORT FORTRAN library of Bell Laboratories.

[1]H. Massey, *Negative Ions* (Cambridge U. P., Cambridge, 1976), pp. 1–15.
[2]R. N. Hill, Phys. Rev. Lett. **38**, 643 (1977); J. Math. Phys. **18**, 2316 (1977).
[3]C. J. Pekeris, Phys. Rev. **112**, 1649 (1958).
[4]S. Chandrasekhar, Astrophys. J. **100**, 176 (1944); see also, H. Shull and P.-O. Löwdin, J. Chem. Phys. **25**, 1035 (1956).
[5]C. Froese-Fischer, *The Hartree-Fock Method for Atoms* (Wiley, New York, 1977).
[6]F. H. Stillinger and D. K. Stillinger, Phys. Rev. A **10**, 1109 (1974).
[7]See, for example, W. Thirring, *Quantum Mechanics of Atoms and Molecules* (Springer-Verlag, Berlin, 1979), Sec. 4.3.5.
[8]F. Stillinger, J. Chem. Phys. **45**, 3623 (1966).
[9]M. Hoffmann-Ostenhof, T. Hoffmann-Ostenhof, and B. Simon, J. Phys. A **16**, 1125 (1983).
[10]E. H. Lieb, Rev. Mod. Phys. **53**, 603 (1981). See Theorem 7.7 and Remark (iii) on p. 630.
[11]N. Bazley and B. Zwahlen, Manuscripta Math. **2**, 365 (1970).
[12]M. Reeken, J. Math. Phys. **11**, 2505 (1970).
[13]J. Wolkowsky, Ind. Univ. Math. J. **22**, 551 (1972); C. Stuart, Arch. Rat. Mech. Anal. **51**, 60 (1973).
[14]K. Gustafson and D. Sather, Rend. Mat. **4**, 723 (1971).
[15]E. H. Lieb and B. Simon, J. Chem. Phys. **61**, 735 (1974); Commun. Math. Phys. **53**, 185 (1974).
[16]P. Bader, Proc. R. Phys. Soc. Edinburgh Ser. A **82**, 27 (1978).
[17]P. L. Lions, Nonlinear Anal. **5**, 1245 (1981).
[18]R. Benguria and E. Lieb, Phys. Rev. Lett. **50**, 1771 (1983).
[19]See Ref. 10, Theorems 7.15 and 7.23 and Ref. 20.
[20]B. Baumgartner (private communication).
[21]N. Bazley and R. Seydel, Chem. Phys. Lett. **24**, 128 (1974).
[22]R. Behling, A. Bongers, and T. Küpper, Int. J. Quantum Chem. **10**, 985 (1976).
[23]That all $A_k > 0$ is stronger than needed to insure $u = 0$. However, an unrestricted numerical variation did not yield trial u with negative coefficients.
[24]See, for example, Ref. 7, Sec. 4.3.45.
[25]See Ref. 5, p. 165.

A Lagrangian theory of the classical spinning electron

Patrick L. Nash
NASA Langley Research Center, Hampton, Virginia 23665

(Received 1 February 1980; accepted for publication 20 January 1984)

A Lagrangian is defined that governs the dynamics of a classical electron with spin, moving under the influence of electromagnetic forces. The Euler–Lagrange equations associated with this Lagrangian for space-time position x^α provide a generalization of the Lorentz force law. The remaining Euler–Lagrange equations lead directly to the (generalized) Frenkel–Thomas–BMT equations.

PACS numbers: 41.70. + t, 03.20. + i

I. INTRODUCTION

There are many formalisms for defining the dynamics of a classical electron, modeled as a point particle with spin.[1-9] One of the most intuitive approaches is to assume that, in addition to position x^α in Minkowski space-time, an orthogonal tetrad of vectors transported along the world line $x^\alpha = x^\alpha(s)$ of the electron comprise an (incomplete) set of classical dynamical variables (observables) for the electron.[6,7,10,11] In order to provide a satisfactory description of the electron's motion, one must set down the law of transport for the tetrad in the presence of electromagnetic fields.

This paper contains an epigenesis of an idea put forward by Proca.[4] Proca suggested that the four-velocity of the electron be constructed from ("classical") Dirac spinors. He defined equations of motion for these spinors from which one could deduce the law of transport for the four-velocity.

The ansatz adopted in this paper is to build an orthogonal tetrad from a real eight component $\overline{\mathrm{O}(3,3)}$ spinor (denoted by ψ). A Poincaré-invariant parameter-invariant Lagrangian is constructed from ψ, $\dot{\psi}$, $\dot{x}^\alpha$, and external fields. The Euler–Lagrange equations associated with this Lagrangian provide equations of motion for ψ and x^α, which, in turn, determine in a straightforward manner evolution equations for the tetrad. We find that the law of transport for the three spacelike members of the tetrad is given by a generalization of the Frenkel–Thomas–Bargmann–Michel–Telegdi equations.[1,2,12] The timelike member of the tetrad is parallel to the four-velocity, whose dynamical evolution is governed by a generalized Lorentz force law.

In this model of the classical electron, the Frenkel condition[1] is satisfied, so that, as is observed in nature, the electron's electric dipole moment is zero in a rest frame. However, as is shown below, the incorporation of this feature into this model necessarily forces the equations of motion for x^α to be nondeterministic for all but one value of $\hbar$, which is a Lagrange multiplier in this formalism. This particular value for $\hbar$ is $\hbar = 0$, but this value is not required by anything (other than our own desire for determinism) in the formalism. This fact appears to be a major flaw within the model.

II. LAGRANGIAN FORMALISM

The fundamental dynamical variables that characterize a classical electron with spin are the particle's position coordinates x^α relative to external field sources, the four-velocity $\dot{x}^\alpha = dx^\alpha/dw$ (where w is an arbitrary parameter up to the constraint $dw/ds > 0$, with $ds^2 = -g_{\alpha\beta}\, dx^\alpha\, dx^\beta > 0$ for dx^α timelike), and a tetrad of vectors $E^{(\mu)}_{\ \alpha}$ transported along its world line. The intrinsic spin tensor $\Sigma^{\alpha\beta}$ of the electron might, at first glance, be added to this list. However, when the Frenkel condition $\Sigma^{\alpha}_{\ \beta}\dot{x}^\beta = 0$, is satisfied, the spin tensor $\Sigma^{\alpha\beta}$ becomes a simple bilinear function of the $E^{(\mu)}_{\ \alpha}$ [see Eq. (28)]. We desire to determine the transport law for this tetrad, in addition to equations of motion for x^α. In accordance with the idea put forward by Proca, instead of attempting to directly generate dynamical equations for the observables $E^{(\mu)}_{\ \alpha}$, we shall introduce the "potentials" ψ (which are not directly observable) and determine equations of motion for them. In order to generate the law of transport for the $E^{(\mu)}_{\ \alpha}$, one differentiates the definitions of the $E^{(\mu)}_{\ \alpha}$ in terms of ψ with respect to proper time s, and then substitutes for $\dot{\psi}$.

Since the construction of the tetrad from ψ is known, and not of primary interest, the reader may desire only a statement of the construction. Details can be found in Refs. 13 and 14. We use the notation and conventions of Ref. 15.

Let ψ denote a real eight component $\overline{\mathrm{O}(3,3)}$ spinor, where ψ is assumed to be dimensionless. Put

$$\psi = \begin{pmatrix} \lambda \\ \tilde{\xi} \end{pmatrix},$$

where λ and ξ are real four component spinors, and the tilde denotes transpose. The ξ transforms inversely to λ under $\overline{\mathrm{SO}(3,3)}$, and the restriction of $\overline{\mathrm{SO}(3,3)}$ to a $\overline{\mathrm{SO}(3,1)}$ subgroup. Let γ^α $\alpha, \beta, \ldots = 1,2,3,4$ be a real 4×4 irreducible representation of the Dirac matrices, where

$$\gamma^\alpha\gamma^\beta + \gamma^\beta\gamma^\alpha = 2\gamma_0\, g^{\alpha\beta}, \tag{1}$$

$$g^{\alpha\beta} = g_{\alpha\beta} = \mathrm{diag}(1,1,1,-1), \tag{2}$$

and

$$\gamma_0 = \mathrm{diag}(1,1,1,1). \tag{3}$$

Define

$$\gamma^5 = -\gamma^1\gamma^2\gamma^3\gamma^4, \tag{4}$$

$$\epsilon = -\gamma^4, \tag{5}$$

and

$$S^{\alpha\beta} = -\tfrac{1}{4}[\gamma^\alpha, \gamma^\beta]. \tag{6}$$

One finds that

 0022-2488/84/062104-05$02.50

$$\tilde{\gamma}^{\alpha}\epsilon = -\epsilon\gamma^{\alpha}, \tag{7}$$

where the tilde denotes transpose;

$$\tilde{\gamma}^{5}\epsilon = \epsilon\gamma^{5}, \tag{8}$$

$$\tilde{S}^{\alpha\beta}\epsilon = -\epsilon S^{\alpha\beta}, \tag{9}$$

and

$$[S^{\alpha\beta}, \gamma_{\mu}] = \delta^{\alpha}_{\mu}\,\gamma^{\beta} - \delta^{\beta}_{\mu}\,\gamma^{\alpha}. \tag{10}$$

A tetrad may be defined as follows:

$$E^{\alpha}_{(1)} = \tfrac{1}{2}\xi\gamma^{\alpha}\gamma^{5}\lambda; \tag{11}$$

$$E^{\alpha}_{(2)} = \tfrac{1}{2}\xi\gamma^{\alpha}\lambda; \tag{12}$$

$$n^{\alpha} = \tilde{\lambda}\epsilon\gamma^{\alpha}\lambda \tag{13}$$

(future-pointing lightlike vector);

$$m^{\alpha} = -\xi\gamma^{\alpha}\epsilon^{-1}\tilde{\xi} \tag{14}$$

(future-pointing lightlike vector);

$$M = \tfrac{1}{2}\xi\lambda; \tag{15}$$

$$N = -\tfrac{1}{2}\xi\gamma^{5}\lambda; \tag{16}$$

n^{α} and m^{α} are linearly independent whenever $M^2 + N^2 \neq 0$. Under this assumption, one may define

$$E^{\alpha}_{(3)} = \tfrac{1}{4}(m^{\alpha} - n^{\alpha}), \tag{17}$$

and

$$E^{\alpha}_{(4)} = \tfrac{1}{4}(m^{\alpha} + n^{\alpha}). \tag{18}$$

The $E^{(\mu)}_{\alpha}$ satisfy[13,14]

$$E^{\alpha}_{(\mu)}E_{(\rho)\alpha} = \eta_{(\mu)(\rho)}(M^2 + N^2), \tag{19}$$

where

$$\eta_{(\mu)(\rho)} = \mathrm{diag}(1,1,1,-1). \tag{20}$$

One also finds that[13,14]

$$m^{\alpha}n_{\alpha} = -8(M^2 + N^2). \tag{21}$$

Equations (19) and (21) may be derived in a straightforward manner using the identities listed in the Appendix. Lastly, let us define the spin tensor of the electron to be

$$\Sigma^{\alpha\beta} = \xi S^{\alpha\beta}\lambda. \tag{22}$$

With the use of Eqs. (A1)–(A3) of the Appendix, one may show that[13,14]

$$\Sigma^{\alpha}_{\ \beta}\, n^{\beta} = Mn^{\alpha}, \tag{23}$$

$$\Sigma^{\alpha}_{\ \beta}\, m^{\beta} = -Mm^{\alpha}, \tag{24}$$

$$^{*}\Sigma^{\alpha}_{\ \beta}\, n^{\beta} = -Nn^{\alpha}, \tag{25}$$

and

$$^{*}\Sigma^{\alpha}_{\ \beta}\, m^{\beta} = Nm^{\alpha}, \tag{26}$$

where

$$^{*}\Sigma^{\alpha\beta} = \tfrac{1}{2}\, g^{\alpha\rho}\, g^{\beta\mu}\, \epsilon_{\rho\mu\sigma\kappa}\, \Sigma^{\sigma\kappa} = \xi\gamma^{5}S^{\alpha\beta}\lambda. \tag{27}$$

On physical grounds it is clear that $E^{\alpha}_{(4)}$ must be parallel to $\dot{x}^{\alpha}$ in order for the tetrad to be a legitimate set of dynamical variables. For $\dot{x}^{\alpha}\dot{x}_{\alpha} = -1$, consider the kinematical constraint $\epsilon^{-1}\tilde{\xi} = -\gamma^{5}\gamma_{\alpha}\,\lambda\dot{x}^{\alpha}$. With ξ so determined, one finds that $M = 0$ (because $\tilde{\lambda}\epsilon\gamma^{5}\gamma^{\alpha}\lambda = 0$, due to the skew symmetry of $\epsilon\gamma^{5}\gamma^{\alpha}$), $N = -\tfrac{1}{2}\, n_{\alpha}\, \dot{x}^{\alpha} > 0$ for $\psi \neq 0$, and $m^{\alpha} = -(-\dot{x}^{\mu}\tilde{\lambda}\epsilon\gamma_{\mu}\gamma^{5})\,\gamma^{\alpha}(-\gamma^{5}\gamma_{\beta}\,\lambda\dot{x}^{\beta}) = 4N\dot{x}^{\alpha} - n^{\alpha}$. Substituting this result into Eq. (18) yields $E^{\alpha}_{(4)} = N\dot{x}^{\alpha}$, as desired. Moreover, one sees from Eqs. (23) and (24) that $M = 0$ implies that $\Sigma^{\alpha}_{\ \beta}\,\dot{x}^{\beta} = 0$ (Frenkel condition) when one identifies $N^{-1}E^{\alpha}_{(4)}$ with $\dot{x}^{\alpha}$. Having $M = 0$ also guarantees that

$$\Sigma^{\alpha\beta} = N^{-1}(E^{\alpha}_{(1)}\,E^{\beta}_{(2)} - E^{\beta}_{(1)}\,E^{\alpha}_{(2)}), \tag{28}$$

a result that follows from expanding $\Sigma^{\alpha\beta}$ in terms of $E^{\alpha}_{(\mu)}\,E^{\beta}_{(\rho)}$. Equation (28) in turn implies that $E^{\alpha}_{(3)}$, which is proportional to $\epsilon^{\alpha\beta\mu\rho}\,E^{(1)}_{\beta}E^{(2)}_{\mu}E^{(4)}_{\rho}$, is therefore proportional to $\tfrac{1}{2}\,\epsilon^{\alpha\beta\mu\rho}\,\Sigma_{\beta\mu}\dot{x}_{\rho}$, i.e., $N^{-1}E^{\alpha}_{(3)}$ is the normalized Pauli–Lubanski spin vector.

In order to put the kinematical constraint into a manifestly covariant form, it is convenient to define 8×8 matrix analogs of Dirac's γ matrices. Let

$$\Gamma^{\alpha} = \begin{pmatrix} 0 & \gamma^{5}\gamma^{\alpha}\epsilon^{-1} \\ -\epsilon\gamma^{\alpha}\gamma^{5} & 0 \end{pmatrix}, \tag{29}$$

$$\Gamma^{5} = \begin{pmatrix} 0 & \epsilon^{-1} \\ -\epsilon & 0 \end{pmatrix}, \tag{30}$$

and

$$\Gamma^{6} = \begin{pmatrix} 0 & \gamma^{5}\epsilon^{-1} \\ \epsilon\gamma^{5} & 0 \end{pmatrix}. \tag{31}$$

One finds that $(A,B,\ldots = 1,\ldots,6)$

$$\Gamma^{A}\Gamma^{B} + \Gamma^{B}\Gamma^{A} = 2g^{AB}I_{8\times 8}, \tag{32}$$

where

$$g^{AB} = \mathrm{diag}(1,1,1,-1,-1,-1), \tag{33}$$

and $I_{8\times 8} = \mathrm{diag}(1,1,1,1,1,1,1,1)$. Expressed covariantly, the relationships $(-\dot{x}^{\alpha}\dot{x}_{\alpha})^{1/2}\,\epsilon^{-1}\tilde{\xi} = -\gamma^{5}\gamma_{\alpha}\lambda\dot{x}^{\alpha}$ and $(-\dot{x}^{\alpha}\,\dot{x}_{\alpha})^{1/2}\lambda = -\dot{x}^{\alpha}\gamma_{\alpha}\gamma^{5}\,\epsilon^{-1}\tilde{\xi}$ (an elementary consequence of the first relationship) are given by

$$\left(\Gamma_{\alpha}\,\dot{x}^{\alpha} + \sqrt{-\dot{x}^{\alpha}\dot{x}_{\alpha}}\;\Gamma^{7}\right)\psi = 0, \tag{34}$$

where

$$\Gamma^{7} = \Gamma^{1}\Gamma^{2}\Gamma^{3}\Gamma^{4}\Gamma^{5}\Gamma^{6}. \tag{35}$$

Here Γ^{7} satisfies

$$\Gamma^{7}\Gamma^{A} + \Gamma^{A}\Gamma^{7} = 0, \tag{36}$$

$$(\Gamma^{7})^{2} = I_{8\times 8}, \tag{37}$$

and, in the particular representation of the Γ^{A} given above, is

$$\Gamma^{7} = \begin{pmatrix} \gamma_{0} & 0 \\ 0 & -\gamma_{0} \end{pmatrix}. \tag{38}$$

We shall ensure that ψ satisfies Eq. (34) by incorporating this constraint into the Lagrangian via Lagrange multipliers.

Before writing down the Lagrangian one comment is in order. As shown in Ref. 14, $E^{\alpha}_{(j)}$ $(j = 1,2,3)$, $\Sigma^{\alpha\beta}$, and N transform as pseudotensors under spatial reflection (and as absolute tensors otherwise), whereas $E^{\alpha}_{(4)}$ transforms as a pseudovector under time reversal (and as an absolute vector otherwise). This means that $N^{-1}E^{\alpha}_{(j)}$, $[-\det(g_{\mu\rho})]^{1/2}\,N^{-1}\,E^{\alpha}_{(4)}$, and $N^{-1}\,\Sigma^{\alpha\beta}$ transform as absolute tensors under all O(3,1) transformations. However, it substantially complicates the equations of motion for ψ if one constructs the Lagrangian out of these latter quantities. In order to construct a Lagrangian that is invariant under the reflections contained in the full Poincaré group, we shall introduce a "quasiscalar" factor that multiplies every expression quadratic in ψ that appears in the Lagrangian. This quasiscalar is to transform as a scalar under proper Lorentz and time reversal transformations, but a scalar density under spatial reflections. Denoting this quasiscalar by $\hbar$, this transformation law may be stated covariantly as

$$\hbar \rightarrow \hbar' = (\operatorname{sgn} L^4_{\ 4})^{-1} (\det L^\alpha_{\ \beta})\, \hbar, \tag{39}$$

where $L^\alpha_{\ \beta}$ denotes the Lorentz matrix. Thus, $\hbar\Sigma^{\alpha\beta}$ is an absolute type (2,0) tensor; $\hbar E^\alpha_{(j)}$ and $\hbar[-\det(g_{\mu\nu})]^{1/2} E^\alpha_{(4)}$ are absolute vectors. Let us agree to give $\hbar$ the dimensions of angular momentum in order that $\hbar\Sigma^{\alpha\beta}$ may have dimensions of angular momentum (ψ is dimensionless).

We shall bring $\hbar$ into the formalism as a Lagrange multiplier. Its transformation properties are not to be given *a priori*, but are to be determined from the requirements of Poincaré invariance of the Lagrangian. That is to say, one demands that $\hbar$ transform according to Eq. (39) so that the Lagrangian is invariant under the full Poincaré group.

The Lagrangian postulated to govern the classical dynamics of a massive point electron with spin, possessing charge $e = -|e|$, is

$$\mathscr{L} = -M\sqrt{-\dot{x}^\alpha \dot{x}_\alpha} - (\hbar/2)\,\tilde{\psi}\Omega\dot{\psi} + eA_\alpha \dot{x}^\alpha + \hbar\tilde{\Lambda}\Omega\left(\Gamma_\alpha \dot{x}^\alpha + \sqrt{-\dot{x}^\alpha \dot{x}_\alpha}\,\Gamma^7\right)\psi. \tag{40}$$

Here $\dot{} = d/dw$, where w is an arbitrary parameter up to the requirement that $dw/ds > 0$. The 8×8 matrix Ω, appearing for the first time in Eq. (40), is a $\overline{\mathrm{SO}(3,3)}$ invariant symplectic form defined by

$$\tilde{\Gamma}^A\Omega = \Omega\Gamma^A, \tag{41}$$

which implies that

$$\tilde{\Gamma}^7\Omega = -\Omega\Gamma^7. \tag{42}$$

A particular representation of Ω is given by

$$\Omega = \Gamma^1\Gamma^2\Gamma^3 = \begin{pmatrix} 0 & \gamma_0 \\ -\gamma_0 & 0 \end{pmatrix}. \tag{43}$$

Here Λ is a column matrix of eight Lagrange multipliers; Λ transforms as ψ under $\overline{\mathrm{SO}(3,1)}$.

The quantity A_α is the vector potential of the (total) electromagnetic field tensor $F_{\alpha\beta} = A_{\beta,\alpha} - A_{\alpha,\beta}$, which describes all external fields plus that due to the electron.

Here, M denotes the effective mass of the electron, and it is postulated to be given by

$$M = m\left[1 - (ge\hbar/2m^2)\,\Sigma^{\alpha\beta}F_{\alpha\beta}\right]^{1/2}, \tag{44}$$

where m is a constant parameter with dimensions of mass. In the weak field limit, $F_{\alpha\beta} \rightarrow 0$, this definition of effective mass M reduces to the definition of mass $= m - (ge\hbar/4m)\,\Sigma^{\alpha\beta}F_{\alpha\beta}$ used by, for example, Corben[9] and Barut.[6] However, one must use the definition of mass given by Eq. (44) in order to ensure that the effective mass enters into the Euler–Lagrange equations for both x^α and ψ in a consistent manner.

In Eq. (44), $g = 2 + \delta g$ is a dimensionless parameter, where δg is arbitrary. The fact that g is assumed to be a scalar is a reflection of the implicit assumption that the intrinsic electromagnetic dipole moment tesnor $\mu^{\alpha\beta}$ of the point electron is given by $\mu^{\alpha\beta} = (\mathrm{const})\,\Sigma^{\alpha\beta}$.

It is straightforward to generalize the electromagnetic interaction appearing in Eq. (40) to a general Yang–Mills interaction. However, the classical approach, which does not allow for the experimentally observed phenomena of particle creation and annihilation, presumably has little validity, even as a limiting form of dynamics, at energies high enough to produce $W^\pm$'s and Z's. A classical approach to dynamics only makes sense in the presence of weak fields such as the electromagnetic and gravitational fields.

The Euler–Lagrange equations for Λ and $\hbar$ imply that (putting $w = s$ after performing the differentiations, so that $-\dot{x}_\alpha\dot{x}^\alpha = 1$)

$$(\Gamma_\alpha \dot{x}^\alpha + \Gamma^7)\,\psi = 0 \tag{45}$$

and

$$\tilde{\psi}\Omega\dot{\psi} = (ge/2M)\,\Sigma^{\alpha\beta}F_{\alpha\beta}, \tag{46}$$

along a dynamical trajectory. The canonical momenta are given by

$$\zeta = \frac{\partial\mathscr{L}}{\partial\dot{\psi}} = -\frac{\hbar}{2}\,\tilde{\psi}\Omega, \tag{47}$$

and

$$p_\alpha = \frac{\partial\mathscr{L}}{\partial\dot{x}^\alpha} = M\dot{x}_\alpha + eA_\alpha + \hbar\tilde{\Lambda}\Omega\,(\Gamma_\alpha - \dot{x}_\alpha\Gamma^7)\,\psi \tag{48}$$

$$= M\dot{x}_\alpha + eA_\alpha + (\delta^\beta_\alpha + \dot{x}_\alpha\dot{x}^\beta)\,\hbar\tilde{\Lambda}\Omega\Gamma_\beta\psi, \tag{49}$$

where we have again put $w = s$ after differentiation, and used Eq. (45) to replace $\Gamma^7\psi$ in Eq. (48) by $-\Gamma_\beta\psi\dot{x}^\beta$, which yields Eq. (49).

Before evaluating the Euler–Lagrange equations for ψ, we must make one last definition. [A summary of $\overline{\mathrm{SO}(3,3)}$ definitions and relationships is given in the Appendix.] Let

$$M^{AB} = -\tfrac{1}{4}[\Gamma^A,\Gamma^B]. \tag{50}$$

One finds that

$$[M^{AB},\Gamma_R] = \delta^A_R\,\Gamma^B - \delta^B_R\,\Gamma^A, \tag{51}$$

$$[M^{AB},\Gamma^7] = 0, \tag{52}$$

and

$$\tilde{M}^{AB}\Omega = -\Omega M^{AB}. \tag{53}$$

The Euler–Lagrange equations for ψ are ($w = s$)

$$\hbar\dot{\psi} + \tfrac{1}{2}\dot{\hbar}\psi = -(ge\hbar/4M)\,M^{\alpha\beta}\psi F_{\alpha\beta} - \hbar(\Gamma_\alpha\dot{x}^\alpha - \Gamma^7)\,\Lambda, \tag{54}$$

where we have used $\Sigma_{\alpha\beta} = -\tfrac{1}{2}\tilde{\psi}\Omega M^{\alpha\beta}\psi$, which follows from Eqs. (29), (43), and (50). Upon multiplying this expression from the left with $\Gamma_\mu\dot{x}^\mu + \Gamma^7$, and using Eqs. (45) and (51) along with $(\Gamma_\alpha\dot{x}^\alpha + \Gamma^7)\,\dot{\psi} = -\Gamma_\alpha\psi\ddot{x}^\alpha$, and then solving for $(\Gamma_\alpha\dot{x}^\alpha - \Gamma^7)\,\Lambda$, one finds that

$$(\Gamma_\alpha\dot{x}^\alpha - \Gamma^7)\,\Lambda = \tfrac{1}{2}\Gamma^7\Gamma_\alpha\psi b^\alpha, \tag{55}$$

$$= M_{\alpha\beta}\psi\dot{x}^\alpha b^\beta, \tag{56}$$

where

$$b^\alpha = \ddot{x}^\alpha - (ge/2M)\,F^\alpha_{\ \beta}\,\dot{x}^\beta. \tag{57}$$

To get Eq. (56) from Eq. (55), we have used Eq. (45) and the fact that $b_\alpha \dot{x}^\alpha = 0$ when $w = s$. Using Eq. (56) we may rewrite Eq. (54) as

$$\hbar\dot{\psi} + \tfrac{1}{2}\dot{\hbar}\psi = -\hbar M^{\alpha\beta}\psi[(ge/4M)F_{\alpha\beta} + \dot{x}_\alpha b_\beta]. \quad (58)$$

Upon differentiating the definitions of Eqs. (11)–(22) with respect to s, and then substituting according to Eq. (58), one finds that

$$\frac{d}{ds}\hbar E^{\alpha}_{(j)} = \frac{ge\hbar}{2M}F^{\alpha}_{\beta}E^{\beta}_{(j)} + \dot{x}^\alpha \hbar E^{\beta}_{(j)} b_\beta, \quad (59)$$

for $j = 1, 2, 3$; also

$$\frac{d}{ds}\hbar N = 0, \quad (60)$$

and

$$\frac{d}{ds}\hbar\Sigma^{\alpha\beta} = \frac{ge\hbar}{2M}(F^{\alpha\mu}\Sigma_\mu{}^\beta - F^{\beta\mu}\Sigma_\mu{}^\alpha) + \hbar b_\mu(\Sigma^{\alpha\mu}\dot{x}^\beta - \Sigma^{\beta\mu}\dot{x}^\alpha). \quad (61)$$

Substituting for b_α from Eq. (57) into Eq. (59) yields a generalization of the Frenkel,[1] Thomas,[2] and BMT[12] equations [it is the BMT equation when $\ddot{x}^\alpha = (e/M)F^{\alpha}_{\beta}\dot{x}^\beta$]. These equations were introduced to describe the evolution of the Pauli–Lubanski pseudovector in a classical context. Here it is seen to govern the evolution of the entire spacelike triad. Frenkel–Thomas–BMT transport preserves both the orthogonality and normalization of the tetrad, as does Eq. (59).

In order to display the Euler–Lagrange equations for x^α, the last term in Eq. (49) must be evaluated, and dM/ds calculated. The $(\delta^\beta_\alpha + \dot{x}_\alpha \dot{x}^\beta)\tilde{\Lambda}\Omega\Gamma_\beta\psi$ may be found as follows. Equation (55) implies that $(\Gamma^7\Gamma_\beta\dot{x}^\beta - 1)\Lambda = \frac{1}{2}\Gamma_\beta\psi b^\beta$, so that

$$\begin{aligned}\tilde{\psi}\Omega\Gamma_\alpha(\Gamma^7\Gamma_\beta\dot{x}^\beta - 1)\Lambda &= \tfrac{1}{2}\tilde{\psi}\Omega\Gamma_\alpha\Gamma_\beta\psi b^\beta \\ &= \tfrac{1}{2}\tilde{\psi}\Omega(g_{\alpha\beta} - 2M_{\alpha\beta})\psi b^\beta \\ &= -\tilde{\psi}\Omega M_{\alpha\beta}\psi b^\beta = 2\Sigma_{\alpha\beta}b^\beta,\end{aligned}$$

since $\Sigma_{\alpha\beta} = -\frac{1}{2}\tilde{\psi}\Omega M_{\alpha\beta}\psi$. This result combined with the alternative evaluation

$$\begin{aligned}&\tilde{\psi}\Omega\Gamma_\alpha(\Gamma^7\Gamma_\beta\dot{x}^\beta - 1)\Lambda \\ &\quad= -\tilde{\psi}\Omega\Gamma_\alpha\Lambda + \tilde{\psi}\Omega\Gamma_\alpha\Gamma^7\Gamma_\beta\Lambda\dot{x}^\beta \\ &\quad= \tilde{\Lambda}\Omega\Gamma_\alpha\psi - \tilde{\Lambda}\Omega\Gamma^7(-\Gamma_\alpha\Gamma_\beta + 2g_{\alpha\beta})\psi\dot{x}^\beta \\ &\quad= \tilde{\Lambda}\Omega\Gamma_\alpha\psi - \tilde{\Lambda}\Omega\Gamma^7\Gamma_\alpha\Gamma^7\psi + 2\dot{x}_\alpha\tilde{\Lambda}\Omega\Gamma_\beta\psi\dot{x}^\beta \\ &\quad= 2(\delta^\beta_\alpha + \dot{x}_\alpha\dot{x}^\beta)\tilde{\Lambda}\Omega\Gamma_\beta\psi\end{aligned}$$

yields

$$(\delta^\beta_\alpha + \dot{x}_\alpha\dot{x}^\beta)\tilde{\Lambda}\Omega\Gamma_\beta\psi = \Sigma_{\alpha\beta}b^\beta. \quad (62)$$

Upon substituting this result into Eq. (49), one finds that p_α is given by

$$p_\alpha = M\dot{x}_\alpha + eA_\alpha + \hbar\Sigma_{\alpha\beta}b^\beta. \quad (63)$$

Next, let us consider

$$\frac{dM}{ds} = -\frac{ge}{4M}\frac{d}{ds}\hbar\Sigma^{\alpha\beta}F_{\alpha\beta} = -\frac{ge\hbar}{4M}\Sigma^{\alpha\beta}F_{\alpha\beta,\mu}\dot{x}^\mu - \frac{ge}{4M}F_{\alpha\beta}\frac{d}{ds}\hbar\Sigma^{\alpha\beta}.$$

Multiplying Eq. (58) from the left with $\dot{\tilde{\psi}}\Omega$ gives

$$\begin{aligned}&\tfrac{1}{2}\dot{\hbar}\dot{\tilde{\psi}}\Omega\psi \\ &\quad= -\tfrac{1}{2}\dot{\hbar}\tilde{\psi}\Omega\dot{\psi} \\ &\quad= -\hbar\dot{\tilde{\psi}}\Omega M^{\alpha\beta}\psi[(ge/4M)F_{\alpha\beta} + \dot{x}_\alpha b_\beta] \\ &\quad= \hbar\dot{\Sigma}^{\alpha\beta}[(ge/4M)F_{\alpha\beta} + \dot{x}_\alpha b_\beta].\end{aligned}$$

From Eq. (46) [or Eq. (58)], $\tilde{\psi}\Omega\dot{\psi} = (ge/2M)\Sigma^{\alpha\beta}F_{\alpha\beta}$; substituting this into the previous expression yields

$$\frac{ge}{4M}F_{\alpha\beta}\frac{d}{ds}\hbar\Sigma^{\alpha\beta} = -\hbar\dot{\Sigma}_{\alpha\beta}\dot{x}^\alpha b^\beta = -\dot{x}^\mu\frac{d}{ds}\hbar\Sigma_{\mu\beta}b^\beta.$$

Thus

$$\frac{dM}{ds} = -\frac{ge\hbar}{4M}\Sigma^{\alpha\beta}F_{\alpha\beta,\mu}\dot{x}^\mu + \dot{x}^\mu\frac{d}{ds}(\hbar\Sigma_{\mu\beta}b^\beta). \quad (64)$$

The Euler–Lagrange equations for x^α are ($w = s$)

$$\frac{d}{ds}p_\alpha = eA_{\beta,\alpha}\dot{x}^\beta + \frac{ge\hbar}{4M}\Sigma^{\mu\nu}F_{\mu\nu,\alpha}. \quad (65)$$

Substituting for p_α from Eq. (63) and using Eq. (60) gives

$$M\ddot{x}^\alpha = eF^\alpha{}_\beta\dot{x}^\beta + (g^{\alpha\mu} + \dot{x}^\alpha\dot{x}^\mu)\times\left(\frac{ge\hbar}{4M}\Sigma^{\beta\lambda}F_{\beta\lambda,\mu} - \frac{d}{ds}(\hbar\Sigma_{\mu\beta}b^\beta)\right). \quad (66)$$

III. DISCUSSION

Equations (61) and (66) are coupled nonlinear equations. They are further complicated by the fact that $\dddot{x}^\alpha$ terms arise in the coupled equations. For example, the expression $\Sigma^\alpha_\beta\dddot{x}^\beta$ appears explicitly in Eq. (66) when one performs the indicated differentiation. Moreover, this cannot be solved for $\dddot{x}^\alpha$ since, along a dynamical trajectory, $\det(\Sigma^{\alpha\beta}) = 0$ [this is a consequence of Eq. (28)]. However, these equations are almost identical in form to those given by Corben,[9] whose method of derivation is completely different from the variational approach adopted here. Corben's derivation is essentially algebraic, and his results possess a general applicability commensurate with the validity of his definitions.

Corben discusses some applications and peculiarities of dynamical equations of the form of Eqs. (61) and (66) in his book.[9] However, he, as well as Barut[6] and other authors, uses the definition of mass $= m - (ge\hbar/4m)\Sigma^{\alpha\beta}F_{\alpha\beta}$. For small spin-field interaction energies the definition of (effective) mass M used in this paper reduces to Corben's, and presumably the approximation of the square root in this case does not greatly alter the nature of the solutions to the equations of motion [compare with the approximation $E = m(1 - v^2)^{-1/2} \simeq m + \frac{1}{2}mv^2$]. The appearance of the square root in the Lagrangian of Eq. (40) does influence the nature of the electromagnetic field due to the electron, and will introduce a nonlinearity in the classical field equations for the electromagnetic field in the presence of classical sources with spin.

The term $-(ge\hbar/2m)\Sigma^{\alpha\beta}F_{\alpha\beta}$, representing the spin-field interaction energy, modifies the $e = 0$ mass of the electron. Here $F_{\alpha\beta}$ should be understood as representing both the external electromagnetic field and the self-field of the electron. Unphysical solutions to the equations of motion will occur if the self-field effects are entirely neglected.

In ending, it should be emphasized that the introduction of "classical" spinors to the canonical formalism generates sufficient additional degrees of freedom so that the properties of a classical point particle possessing intrinsic spin can be modeled utilizing a variational formalism. Heretofore, various approaches incorporating a tetrad as a set of fundamental variables have not been completely satisfactory. For example, nowhere in Ref. 7 is the complete Lagrangian actually defined; all results are purely formal.

ACKNOWLEDGMENT

This work was completed, some time ago, at the University of North Carolina at Chapel Hill.

APPENDIX: SOME MATRIX ALGEBRA

A basic identity satisfied by Dirac's γ matrices is[15]

$$X - \gamma^5 X \gamma^5 + \epsilon^{-1}\tilde{X}\epsilon - \gamma^5\epsilon^{-1}\tilde{X}\epsilon\gamma^5 = \gamma_0 \operatorname{tr} X - \gamma^5 \operatorname{tr} \gamma^5 X, \tag{A1}$$

where X denotes an arbitrary 4×4 matrix that transforms in the same way as $\lambda\xi$ under $\overline{\mathrm{SO}(3,1)}$. Another useful identity is[15]

$$\gamma_\alpha \lambda\xi\gamma^\alpha = \gamma_0\xi\lambda + \gamma^5\xi\gamma^5\lambda + \epsilon^{-1}\tilde{\xi}\tilde{\lambda}\epsilon + \gamma^5\epsilon^{-1}\tilde{\xi}\tilde{\lambda}\epsilon\gamma^5. \tag{A2}$$

From this identity one may derive Eq. (23), for example, by considering $2\Sigma^{\alpha}_{\ \beta}\, n^\beta = \xi\, 2S^{\alpha\beta}\lambda\tilde{\lambda}\epsilon\gamma_\beta\lambda$ $= \xi\,(g^{\alpha\beta} - \gamma^\alpha\gamma^\beta)\,\lambda\tilde{\lambda}\epsilon\gamma_\beta\lambda = 2Mn^\alpha$, since $\gamma^\beta\lambda\tilde{\lambda}\epsilon\gamma_\beta\,\lambda = 0$, owing to the identity

$$-\gamma_\alpha\lambda\tilde{\lambda}\epsilon\gamma^\alpha = \lambda\tilde{\lambda}\epsilon + \gamma^5\lambda\tilde{\lambda}\epsilon\gamma^5. \tag{A3}$$

Equation (A3) follows trivially from Eq. (A2) upon setting $\xi = \tilde{\lambda}\epsilon$ in Eq. (A2), and remembering that ϵ and $\epsilon\gamma^5$ are both skew-symmetric matrices, so that $\tilde{\lambda}\epsilon\lambda = 0 = \tilde{\lambda}\epsilon\gamma^5\lambda$.

A summary of definitions relevant to the pseudo-Clifford algebra C_6 and the Lie algebra SO(3,3) follows.

Let Γ^A $A,B,... = 1,...,6$ denote six real 8×8 matrices that verify

$$\Gamma^A\Gamma^B + \Gamma^B\Gamma^A = 2g^{AB}I_{8\times 8}, \tag{A4}$$

where

$$g^{AB} = g_{AB} = \operatorname{diag}(1,1,1,-1,-1,-1). \tag{A5}$$

The generators of $\overline{\mathrm{SO}(3,3)}$ are

$$M^{AB} = -\tfrac{1}{4}[\Gamma^A, \Gamma^B]. \tag{A6}$$

The M^{AB} verify

$$[M^{AB}, M^{RS}] = g^{AR}M^{BS} - g^{AS}M^{BR} - g^{BR}M^{AS} + g^{BS}M^{AR}, \tag{A7}$$

and

$$[M^{AB}, \Gamma_R] = \delta^A_R\,\Gamma^B - \delta^B_R\Gamma^A. \tag{A8}$$

A $\overline{\mathrm{SO}(3,3)}$ invariant symplectic form Ω may be defined on the vector space "$\lambda + \xi$" by requiring that

$$\tilde{\Gamma}^A\Omega = \Omega\Gamma^A. \tag{A9}$$

Since $\tilde{\Omega} = -\Omega$, each $\Omega\Gamma^A$ is skew symmetric. Equations (A6) and (A9) imply that

$$\tilde{M}^{AB}\Omega = -\Omega M^{AB}, \tag{A10}$$

hence each ΩM^{AB} is symmetric.

Define

$$\Gamma^7 = \Gamma^1\Gamma^2\Gamma^3\Gamma^4\Gamma^5\Gamma^6; \tag{A11}$$

whence,

$$\Gamma^7\Gamma^A + \Gamma^A\Gamma^7 = 0, \tag{A12}$$

$$(\Gamma^7)^2 = I_{8\times 8}, \tag{A13}$$

and

$$\tilde{\Gamma}^7\Omega = -\Omega\Gamma^7, \tag{A14}$$

so that $\Omega\Gamma^7$ is symmetric. In fact, $\Omega\Gamma^7$ is an $\overline{\mathrm{O}(3,3)}$ invariant, covariant rank two, $\overline{\mathrm{O}(3,3)}$ spinor on the eight-dimensional vector space of which ψ is an element.

[1] W. J. Frenkel, Z. Phys. **37**, 243 (1926).
[2] L. H. Thomas, Philos. Mag. **3**, 1 (1927).
[3] H. J. Bhaba and H. C. Corben, Proc. R. Soc. London, Ser. A **178**, 243 (1941).
[4] A. Proca, J. Phys. Radium **15**, 65 (1954).
[5] R. Schiller, Phys. Rev. **125**, 1116; **128**, 1402 (1962).
[6] A. O. Barut, *Electrodynamics and Classical Theory of Fields and Particles* (Dover, New York, 1980), 2nd ed.
[7] A. J. Hanson and T. Regge, Ann. Phys. **87**, 498 (1974).
[8] J. Stachel and J. Plebanski, J. Math. Phys. **18**, 2368 (1977).
[9] H. C. Corben, *Classical and Quantum Theories of Spinning Particles* (Holden–Day, San Francisco, 1968).
[10] J. B. Hughes, Nuovo Cimento Suppl. **20**, 89, 149 (1961).
[11] F. Halbwachs, P. Hillion, and J. P. Vigier, Nuovo Cimento **10**, 817 (1958).
[12] V. Bargmann, L. Michel, and V. L. Telegdi, Phys. Rev. Lett. **2**, 435 (1959).
[13] P. L. Nash, J. Math. Phys. **21**, 1024 (1980).
[14] P. L. Nash, dissertation, University of North Carolina at Chapel Hill, 1981 (unpublished).
[15] P. L. Nash, J. Math. Phys. **25**, 204 (1984).

Inverse scattering problems: A study in terms of the zeros of entire functions

M. Nieto-Vesperinas[a)]
Instituto de Optica del C.S.I.C., Serrano 121, Madrid-6, Spain

(Received 30 March 1983; accepted for publication 14 October 1983)

The study of the zeros of functions of several complex variables recently reported is applied to scalar scattered wavefields. First, an analysis of the role of the zeros of the scattering amplitude leads to an interpretation of the nonuniqueness of inverse source and scattering problems. Secondly, the analytic character of the scattered wavefunction outside the scattering volume and its zero description, as well as its role for information transmission, is discussed.

PACS numbers: 42.10.Hc, 03.65.Sq, 02.30.Dk, 02.30.Jr

1. INTRODUCTION

The analyticity and unique continuation of solutions to partial differential equations has been discussed in several contexts in physics (see, e.g., Ref. 1). One of the applications has been, among others, the description of scalar scattered wavefields in the far zone, i.e., when the scattered field $u^{(s)}(\mathbf{r})$, $(\mathbf{r} = (x,y,z))$ and the spectral amplitude $A(k_0\mathbf{s}_\perp)$, $[k_0\mathbf{s} = k_0(s_1,s_2,s_3)$, $k_0 = 2\pi/\lambda$, and $\mathbf{s}$ being a unit vector in the direction of observation. $k_0\mathbf{s}_\perp = k_0(s_1,s_2)$ is the projection of the 3-D wavevector $k_0\mathbf{s}$ onto the XY-plane so that $|\mathbf{s}_\perp|^2 + s_3^2 = 1, s_3 = +\sqrt{1 - s_\perp{}^2}$ and $|k_0\mathbf{s}_\perp| \leqslant |k_0\mathbf{s}| = k_0]$are connected by the asymptotic expression[2]

$$u^{(s)}(\mathbf{r}) = -2\pi i k_0 s_3 A(k_0\mathbf{s}_\perp)\frac{e^{ik_0 r}}{r}. \tag{1}$$

So far these discussions have been first restricted to functions $A(s)$ of one scalar variable s (Refs. 3–5) and, within this limitation, they have addressed the ambiguity in object reconstruction from far-zone intensity data (phase problem) and put forward methods of solution based on the localization of the zeros of $A(s)$ in the complex plane $\hat{s} = s + is'$.[5–8] Analytic Fourier Optics, as this analysis of $A(s)$ was coined,[9] also provides a means of encoding information, which is alternative to the Shannon description of sampling and interpolation of bandlimited functions. Recently, some generalizations of the study of the above ambiguity have been done by means of the theory of entire functions of two and more complex variables.[10–12]

The purpose of this paper is twofold. We shall see that the angular spectrum $A(K_1,K_2)$ is the restriction of the 3-D Fourier transform $F(K_1,K_2,K_3)$ of the source function to values of $K_3 = q$ satisfying $|\mathbf{K}|^2 + q^2 = k_0^2$, $\mathbf{K} = (K_1,K_2)$, for any real value of $|\mathbf{K}|$. Thus, based on the theory of Fourier transforms in several complex variables we shall discuss first in Sec. 2 how the study of the analytic continuation of F into three complex variables allows an interpretation of the nonuniqueness of the inverse source and scattering problems,[13,14] within the framework of the theory of entire functions. Specifically, we shall base our discussion on the recent paper by Manolitsakis[12] on two-dimensional scattered fields, which are given by a Fourier transform, considered as functions of two-complex variables. As such, we shall use its description in terms of "zero lines" of a function of several complex variables [where a "zero line" in the n-dimensional complex space means an $(n-1)$-dimensional surface in which the function vanishes]. We shall not present in detail these mathematical essentials as they are presented in depth in Ref. 12. We will instead concentrate on some consequences extracted from that theory. In particular, a description of the source Fourier transforms $F(\hat{K}_1,\hat{K}_2,\hat{K}_3)$ in terms of the zero lines of its analytic continuation into the 6-D complex space $(\hat{K}_1,\hat{K}_2,\hat{K}_3)$ justifies, by virtue of a theorem due to Osgood, the existence of identically zero scattering amplitudes given as data of a single or a discrete number of scattering experiments.[14,15]

Secondly, in Sec. 3 we shall address the study of the analyticity of the scattered field without the approximations introduced for studying its far-zone behavior. We shall see that between this region and the scattering object, and more specifically, outside the scattering volume and its boundary, the field may be described in its free propagation by an analytic function of the spatial complex variables $(\hat{x},\hat{y},\hat{z})$. The kind of analyticity may be even more specified in those regions in which the scattered field admits a representation in terms of the angular spectrum of plane waves. As a matter of fact, given a certain level of noise this field may be approximated at optical frequencies by an entire function of exponential type at those points where evanescent waves from a

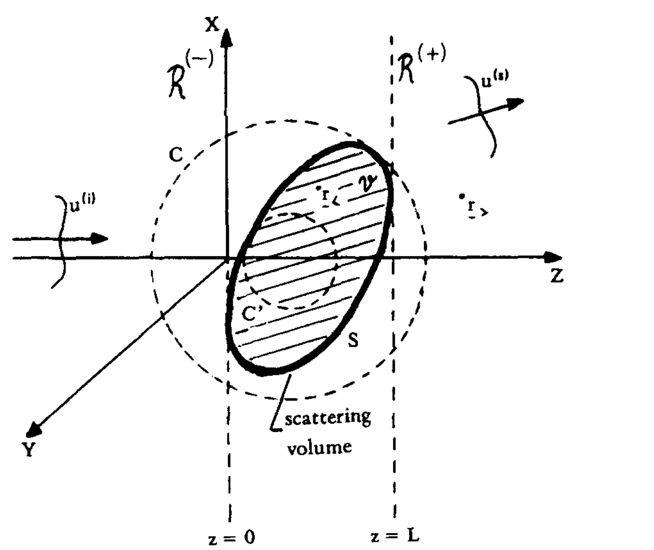


FIG. 1. Scattering geometry.

a) Temporary address: The Institute of Optics, The University of Rochester, Rochester, New York 14627.

 0022-2488/84/062109-07$02.50

certain spatial frequency $K=\sqrt{K_1^2+K_2^2}>k_0$ onwards are not detected and thus may be neglected. The characterization of this scattered field by "zero lines" is a difficult one in several complex variables.[12] However, in those problems exhibiting cylindrical symmetry, the scattered field may be described at each distance z from the object by an entire function of exponential type $u^{(s)}(\hat{x})$ of a single variable; as such, its propagation may be described by means of its zero trajectories giving a zero distribution with a certain density at each complex plane associated with a given distance z, which decreases, as this distance increases, up to the limit marked by the evanescent waves. The reconstruction of the field in the process reversal to that of propagation from the object poses, therefore, the problem of reproducing higher zero densities as one gets closer to the scattering volume.

2. ANALYTIC PROPERTIES OF THE 3-D SOURCE FOURIER TRANSFORM

Let $u(\mathbf{r})$ represent a scalar monochromatic wavefield in the presence of a scattering medium that occupies a finite volume v in the strip $0<z<L$ (Fig. 1). $u(\mathbf{r})$ satisfies at any point $\mathbf{r}=(x,y,z)$ of the space the inhomogeneous Helmholtz equation

$$(\nabla^2+k_0^2)u(\mathbf{t})=V(\mathbf{r})u(\mathbf{r}), \tag{2}$$

where the scattering potential is given in terms of the refractive index $n(\mathbf{r})$ of the scattering medium by

$$\begin{aligned} V(\mathbf{r}) &= -k_0^2[n^2(\mathbf{r})-1], && \mathbf{r}\in v,\\ V(\mathbf{r}) &= 0, && \mathbf{r}\notin v, \end{aligned} \tag{3}$$

$k_0=2\pi/\lambda$ being the wavenumber in free space.

In order to write the integral equations describing the propagation of the field $u(\mathbf{r})$ we shall follow a procedure already employed by Wolf.[15]

Let $\mathbf{r}_>$ and $\mathbf{r}_<$ denote respectively the position vectors of a point *exterior* and *interior* to the scattering volume v, then the application of Green's theorem yields for the exterior field[16]:

(i) In terms of the field $u(\mathbf{r}')$ inside the volume v:

$$u(\mathbf{r}_>)=u^{(i)}(\mathbf{r}_>)-\frac{1}{4\pi}\int_v d^3\mathbf{r}'\; G(\mathbf{r}_>,\mathbf{r}')V(\mathbf{r}')u(\mathbf{r}'). \tag{4}$$

(ii) As a function of the values of the field and its normal derivative on the surface S which constitutes the boundary of v:

$$u(\mathbf{r}_>)=u^{(i)}(\mathbf{r}_>)+\frac{1}{4\pi}\int_S ds\left[u(\mathbf{r}')\frac{\partial G(\mathbf{r}_>,\mathbf{r}')}{\partial n} - G(\mathbf{r}_>,\mathbf{r}')\frac{\partial u(\mathbf{r}')}{\partial n}\right]. \tag{5}$$

In Eqs. (4) and (5) $u^{(i)}(\mathbf{r})$ denotes the incident field, and the integral of the second term represents the scattered field $u^{(s)}(\mathbf{r})$. $G(\mathbf{r},\mathbf{r}')$ is the outgoing free space Green's function:

$$G(\mathbf{r},\mathbf{r}')=\frac{e^{ik_0|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}, \tag{6}$$

which is the solution of the equation

$$(\nabla^2+k_0^2)G(\mathbf{r},\mathbf{r}')=-4\pi\delta(\mathbf{r}-\mathbf{r}'). \tag{7}$$

$G(\mathbf{r},\mathbf{r}')$ admits the decomposition (Weyl)[15]

$$G(\mathbf{r},\mathbf{r}')=\frac{i}{2\pi}\int\int_{-\infty}^{\infty}\frac{d^2\mathbf{K}}{q}e^{i[\mathbf{K}\cdot(\mathbf{R}-\mathbf{R}')+q|z-z'|]}, \tag{8}$$

where $\mathbf{r}=(\mathbf{R},z)$ and $\mathbf{r}'=(\mathbf{R}',z')$ ($\mathbf{R}$ and $\mathbf{R}'$ being vectors in a plane perpendicular to 0Z). Also

$$q=+\sqrt{k_0^2-K^2}\quad\text{when } |\mathbf{K}|\leqslant k_0, \tag{9a}$$

$$q=+i\sqrt{K^2-k_0^2}\quad\text{when } |\mathbf{K}|>k_0. \tag{9b}$$

Denoting by $R^{(+)}$ and $R^{(-)}$, respectively, the half-spaces $z<0$ and $z>L$ (Fig. 1), the absolute value $|z-z'|$ may be written as

$$|z-z'|=+(z-z'),\ \text{when } \mathbf{r}\in R^{(+)} \text{ and } \mathbf{r}'\in v+S, \tag{10a}$$

$$|z-z'|=-(z-z'),\ \text{when } \mathbf{r}\in R^{(-)} \text{ and } \mathbf{r}'\in v. \tag{10b}$$

By substituting Eq. (8) into Eqs. (4) and (5), making use of Eqs. (10) and changing the order of integration, we obtain for the scattered field[17]:

$$u^{(s)}(\mathbf{r}_>)=\int\int_{-\infty}^{\infty}d^2\mathbf{K}\;A^{\pm}(\mathbf{K})e^{i(\mathbf{K}\cdot\mathbf{R}_>\pm qz_>)}, \tag{11}$$

where the angular spectrum, or scattering amplitude, $A^{(\pm)}(\mathbf{K})$ may be written either as a volume integral [cf. Eq. (4)]

$$A^{(\pm)}(\mathbf{K})=\frac{-i}{8\pi^2}\frac{1}{q}\int_v d^3\mathbf{r}\;V(\mathbf{r})u(\mathbf{r}')e^{-i(\mathbf{K}\cdot\mathbf{R}'\pm qz')} \tag{12}$$

or as a surface integral [cf. Eq. (5)]

$$A^{(\pm)}(\mathbf{K})=\frac{i}{8\pi^2}\frac{1}{q}\int_S ds\left[u(\mathbf{r}')\frac{\partial}{\partial n}(e^{-i(\mathbf{K}\cdot\mathbf{R}'\pm qz')}) - e^{-i(\mathbf{K}\cdot\mathbf{R}'\pm qz')}\frac{\partial u}{\partial n}\right]. \tag{13}$$

But one has

$$\frac{\partial}{\partial n}[e^{-i(\mathbf{K}\cdot\mathbf{R}'\pm qz')}]=-i\mathbf{n}\cdot(\mathbf{K},\pm q)e^{-i(\mathbf{K}\cdot\mathbf{R}'\pm qz')} \tag{14}$$

That is

$$A^{(\pm)}(\mathbf{K})=\frac{-i}{8\pi^2 q}\int_S ds\left[i\mathbf{n}\cdot(\mathbf{K},\pm q)u(\mathbf{r}') + \frac{\partial u(\mathbf{r}')}{\partial n}\right]e^{-i(\mathbf{K}\cdot\mathbf{R}'\pm qz')} \tag{15}$$

which reduces to the usual expression for the angular spectrum in terms of $u(\mathbf{R}',L)$ when the surface S is taken as the plane $z=L$ [cf. Eq. (2.4) of Ref. 18] and one chooses Sommerfeld's Green's function.[19]

The upper $(+)$ and lower $(-)$ signs in Eqs. (12)–(15) are taken according to whether the point $\mathbf{r}_>$ belongs either to $R^{(+)}$ or $R^{(-)}$.

As is well known, those components of the scattered field, Eq. (1), for which Eq. (9a) holds, represent *homogeneous waves* and those components satisfying Eq. (9b) are *evanescent waves*, whose contribution may be significant at points $\mathbf{r}_>$ near the boundary S; and certainly, if the dimensions of the scattering volume are of the order of, or smaller than, the wavelength λ of the incident field (or if the scale of inhomogeneity of the scattering medium is of the order of, or smaller than λ), then, the contribution of the evanescent waves in the expression (11) may be important.[20–22]

Considering the representation (12) for $A^{(\pm)}(\mathbf{K})$, we shall introduce

$$f(\mathbf{r}') = V(\mathbf{r}')u(\mathbf{r}') \tag{16}$$

which is sometimes called the object wave or source function[4,13] and it is such that, due to Eqs. (3), becomes zero outside the volume v; i.e., $f(\mathbf{r}')$ is a function of finite support v. As a consequence, its three-dimensional Fourier transform $F(\mathbf{k})$:

$$F(\mathbf{k}) = \frac{1}{(2\pi)^3}\int_{-\infty}^{\infty} d^3\,\mathbf{r}' f(\mathbf{r}')e^{-\mathbf{k}\cdot\mathbf{r}'}, \qquad \mathbf{k} = (K_1,K_2,K_3), \tag{17}$$

is, by virtue of the Plancherel–Polya theorem,[23] the boundary value for real values of $\mathbf{k}$ of an *entire function of exponential type* in the $\hat{\mathbf{k}}$ complex space of the three complex variables $\hat{\mathbf{k}} = (\hat{K}_1,\hat{K}_2,\hat{K}_3)$ $(\hat{K}_1 = K_1 + iK'_1, \hat{K}_z = K_2 + iK'_2, \hat{K}_3 = K_3 + iK'_3)$, i.e., such that $F(\hat{\mathbf{k}})$ always satisfies the inequality

$$|F(\hat{K}_1,\hat{K}_2,\hat{K}_3)| \leqslant Ce^{a(|\hat{K}_1| + |\hat{K}_2| + |\hat{K}_3|)}, \tag{18}$$

where C and a are positive constants.

Comparing Eqs. (11) and (16) we obtain

$$A^{(\pm)}(\mathbf{K}) = \frac{\pi}{iq}F(\mathbf{k}^{(\pm)}), \qquad \mathbf{k}^{(\pm)} = (K_1,K_2,\pm q), \tag{19}$$

i.e., *Eq. (19) allows us to express the angular spectrum* $A^{(\pm)}(\mathbf{k})$ *in terms of the 3-D Fourier transform* $F(\mathbf{k})$ *of the source function restricted to the conditions* (9). It has been already noted[24] that for values $|\mathbf{k}| > K_0$, q is imaginary and $A^{(\pm)}(\mathbf{k})$ represents the value of $F(\hat{\mathbf{k}})$ along the "line" given by Eq. (9b).

So far only real values of $\mathbf{K}$ have been considered in the path of integration corresponding to Eq. (8). If we make a transformation to cylindrical coordinates

$$K_1 = \zeta\cos\alpha, \quad K_2 = \zeta\sin\alpha, \tag{20a}$$

then Eq. (7) may be written

$$G(\mathbf{r},\mathbf{r}') = i\int_0^{\infty}\frac{d\zeta\,\zeta}{\sqrt{k_0^2-\zeta^2}}J_0(\zeta\,|\mathbf{R}-\mathbf{R}'|)e^{i\sqrt{k_0^2-\zeta^2}|z-z'|}.$$

The above integral may be extended to negative values of ζ by expressing the zero order Bessel function J_0 in terms of the Hankel function of the first kind and order zero H_0^1

$$J_0(z) = \tfrac{1}{2}[H_0^1(z) - H_0^1(-z)]$$

so that

$$G(\mathbf{r},\mathbf{r}') = \frac{i}{2}\int_{-\infty}^{\infty}\frac{d\zeta\,\zeta}{\sqrt{k_0^2-\zeta^2}}H_0^1(\zeta\,|\mathbf{r}-\mathbf{r}'|)e^{i\sqrt{k_0^2-\zeta^2}|z-z'|}. \tag{20b}$$

The path of integration in (20b) may be modified in the complex ζ-plane by choosing any other path with ends $(-\infty,\infty)$ providing one avoids to cross the cuts $\zeta \equiv \mathrm{Re}\{\hat{\zeta}\} = \mathrm{Re}\{\lambda k_0\}$, $\zeta' \equiv \mathrm{Im}\{\hat{\zeta}\} \gtrless 0$ arising from the two branch points at $\zeta = \pm k_0$. The choice of such a path with complex values $\hat{\zeta}$ would imply a decomposition of the exterior scattered field $u^{(s)}(\mathbf{r}_>)$, Eq. (4), into inhomogeneous waves, alternative to the representation (11) in terms of homogeneous and evanescent components.

Following now the analysis of Ref. 12 and returning to Eq. (17), a central result of the theory of functions of several complex variables is that a function $F(\hat{K}_1,\hat{K}_2,\hat{K}_3)$ that is an entire function of exponential type in the $\hat{\mathbf{k}}$-space has no isolated zero points[23]; i.e., their zeros are always continuous lines [here "line" means of course a 4-dimensional variety of the 6-D complex $\hat{\mathbf{k}}$-space given by a functional dependence between the three complex variables $(\hat{K}_1,\hat{K}_2,\hat{K}_3)$]. An entire function of exponential type has an infinite number of zero lines.[12]

The set of zeros of $F(\hat{K}_1,\hat{K}_2,\hat{K}_3)$ is called a reducible analytic set if it can be decomposed into a union of subsets. The entire function $F(\hat{K}_1,\hat{K}_2,\hat{K}_3)$ is called globally reducible if it may be written as the product of two entire functions. Otherwise it is called globally irreducible. (Global irreducibility does not imply that F is irreducible everywhere.)

The main fundamental theorem that is required in order to characterize an entire function of several complex variables is the following due to Osgood (see a proof in Ref. 25):

Theorem: An entire function $F(\hat{K}_1,\hat{K}_2,\hat{K}_3)$ that has zeros and does not vanish identically can be uniquely decomposed into a finite or infinite product of globally irreducible factors in the form:

$$F(\hat{K}_1,\hat{K}_2,\hat{K}_3) = \prod_{m=1}^{N}[F_m(\hat{\mathbf{k}})e^{\gamma_m(\hat{\mathbf{k}})}]^{l_m}, \tag{21}$$

where $F_m(\hat{\mathbf{k}})$ are globally irreducible entire functions, exp $[\gamma_m(\hat{\mathbf{k}})]$ are convergence factors with $\gamma_m(\hat{\mathbf{k}})$ being polynomials in $\hat{\mathbf{k}}$ and the exponents l_m are integers.

Let us compare the decomposition given by Eq.(21) with that corresponding to an entire function of one complex variable $\hat{K}$[5]:

$$F(\hat{K}) = C\hat{K}^{\mu}e^{\alpha\hat{K}}\prod_{n=-\infty}^{\infty}\left(1-\frac{\hat{K}}{\hat{K}_n}\right)e^{\hat{K}/\hat{K}_n}, \tag{22}$$

where C, μ, and α are constants and $\hat{K}_n$ are the zero points of $F(\hat{K})$.

The main difference between (21) and (22) is that each factor $F_m(\hat{\mathbf{k}})e^{\gamma_m(\hat{K})}$ is not known and depends entirely on the zero set of $F(\hat{\mathbf{k}})$. Also, each factor of Eq. (21) may contain not only one, but a certain number (finite or infinite) of zero lines with their possible branches or pseudobranches that cannot be separated.[12] While in the case of functions of a single complex variable all the information about the function is contained in the position of its zeros $\hat{K}_n$. For functions of several complex variables, however, it is necessary to know not only the position but also the relationship between zero lines and its combination in each irreducible factor. The uniqueness of irreducible zero sets ensures that each irreducible factor determines in an unique manner the function $F(\hat{K}_1,\hat{K}_2,\hat{K}_3)$. The uniqueness of the product ensures that each function $F(\hat{K}_1,\hat{K}_2,\hat{K}_3)$ is uniquely determined by the zero lines.

The asymptotic behavior of $F(\hat{K}_1,\hat{K}_2,\hat{K}_3)$ is defined by the shape of the support of $f(\mathbf{r})$, i.e., by the shape of the volume v. For example, a function $f(\mathbf{r})$ contained in a spherical volume will yield a 3-D Fourier transform that behaves asymptotically as $(1/\hat{k}^2)[\sin a\hat{k} - a\hat{k}\cos a\hat{k}]$, $\hat{k} = \sqrt{K_1^2 + K_2^2 + K_3^2}$ which is the 3-D Fourier transform of a sphere of radius a. Its zeros are given by the lines $\hat{k} = \hat{k}_0$, where $\hat{k}_0$ are the values that are solutions of the equation $\tan a\hat{k} = a\hat{k}$.

The existence of an infinity of zero lines of an entire function of exponential type $F(\hat{K}_1,\hat{K}_2,\hat{K}_3)$ allows an interpretation of nonuniqueness of object wave reconstruction from the value of $A^{(\pm)}(\mathbf{K})$ obtained from a single experiment.[14] In fact, *there may exist an object wave $f_0(\mathbf{r})$ such that, for a given incident field, its 3-D Fourier transform $F(\hat{K}_1,\hat{K}_2,\hat{K}_3)$ given by Eq. (16) has one of its zero lines given by Eq. (8), i.e., such that $A_0{}^{\pm}(\mathbf{k})$ is zero. Then, due to Eq. (11) the corresponding scattered field $u_0^{(s)}(\mathbf{r}_>)$ would be zero in $R^{(+)}$ or $R^{(-)}$.* Also since the object wave, Eq. (16), is a function localized within v then the angular spectrum

$$A'^{(\pm)}(\mathbf{K}) = A_0^{(\pm)}(\mathbf{K}) + A^{(\pm)}(\mathbf{K}) \tag{23}$$

would be equal to $A^{(\pm)}(\mathbf{K})$ and thus the corresponding scattered field

$$u'^{(s)}(\mathbf{r}_>) = u_0^{(s)}(\mathbf{r}_>) + u^{(s)}(\mathbf{r}_>) \tag{24}$$

would be equal to $u^{(s)}(\mathbf{r}_>)$ in $R^{(+)}$ or $R^{(-)}$.

Moreover, as was noted in Ref. 19, from Eqs. (12) and (23) it is evident that within v one has

$$V'(\mathbf{r}) = \frac{V(\mathbf{r})u(r) + f_0(\mathbf{r})}{u(\mathbf{r}) + u_0(\mathbf{r})}. \tag{25}$$

For a given incident field $u^{(i)}(\mathbf{r}) = \exp(i\mathbf{k}^{(i)}\cdot\mathbf{r})$, $\mathbf{k}^{(i)} = (K_1^{(i)},K_2^{(i)},K_3^{(i)})$, $k_0^2 = K_1^{(i)^2} + K_2^{(i)^2} + q^{(i)^2}$, and within the domain of validity of the first Born approximation the angular spectrum $A_B^{(\pm)}(\chi)$ is given by

$$A_B^{(\pm)}(\chi) = \frac{\pi}{iq} F(\mathbf{x}^{(\pm)}) \tag{26}$$

with the transfer vector $\mathbf{x}^{(\pm)} = (\chi, x_3^{(\pm)}) = (K_1 - K_1^{(i)}, K_2 - K_2^{(i)}, \pm q - q^{(i)})$ lying on the surface

$$\mathbf{x}^{(\pm)}\cdot\mathbf{x}^{(\pm)} = 2k_0^2 - 2\mathbf{k}^{(\pm)}\cdot\mathbf{k}^{(i)}. \tag{27}$$

For a discrete number of incident wave vectors $\mathbf{k}^{(i)}$, one may have a discrete number of lines (27) being zeros of $F(\hat{\mathbf{x}}^{(\pm)})$ and, thus, allowing to express $F(\hat{\mathbf{x}}^{(\pm)})$ by means of a product of irreducible factors of the form of Eq. (21). Thus, as it was already noted[14] it follows as above that a discrete number of experiments, each with an incident plane wave of wave vector $\mathbf{k}^{(i)}$ does not allow, within the first Born approximation, to uniquely determine the object wave from the angular spectrum $A_B^{(\pm)}(\chi)$.

3. ANALYTICITY OF THE SCATTERED FIELD

In order to assess the analytic characteristics of the scattered field $u^{(s)}(\mathbf{r})$, we shall recall the Cauchy–Kowalewski theorem of the theory of partial differential equations (see Ref. 1 for a proof and detailed discussion):

Theorem: There is exactly one analytic solution of the partial differential equation

$$\nabla^2 u(x,z) = g\left(x,z,u,\frac{\partial u}{\partial x},\frac{\partial u}{\partial z}\right) \tag{28}$$

that satisfies the Cauchy boundary conditions on a smooth boundary C of a bounded simply connected domain D: $\phi = u(x,z)|_C$, $\psi = \partial u(x,z)/\partial n|_C$, g being an analytic function of its five arguments and $u(x,z)$ being twice continuously differentiable in a real neighborhood of $x = 0$, $z = 0$.

Besides (see Refs. 1, 26, and 27), $u(x,z)$ may be analytically continued to complex values $\hat{x} = x + ix'$, $\hat{z} = z + iz'$, so that $u(\hat{x},\hat{z})$ satisfies the Cauchy–Riemann equations with respect to each complex variable, and reduces for $x' = z' = 0$ to the function $u(x,z)$.

These results established for two variables x,z may be also obtained for more independent variables (cf. Ref. 1 and references therein).

Equation (1) does not satisfy the conditions of the above theorem since the source function $f(\mathbf{r})$ is discontinuous across the boundary S. However the Helmholtz equation holds the above conditions in the domain exterior to the scattering volume v and its boundary S where $f(\mathbf{r}_>) \equiv 0$. This implies that $u^{(s)}(\mathbf{r})$ fails to be analytic inside the volume v or on its boundary S. In fact the scattered field $u^{(s)}(\mathbf{r})$ must have singularities associated to the scattering sources somewhere in v or on its boundary S since if it were analytic everywhere it would be identically zero.[26] Wave functions analytic in the whole space are the incident field $u^{(i)}(\mathbf{r})$ and any other source-free field that satisfies the equation $(\nabla^2 + k_0^2)u(\mathbf{r}) = 0$ everywhere.[28] Also, from the representations (4) or (5) for $u(\mathbf{r}_>)$ one can see that since $G(\mathbf{r}_>,\mathbf{r}')$ depends analytically on $\mathbf{r}_>$ and $\mathbf{r}'$ then $u^{(s)}(\mathbf{r}_>)$ is an analytic function $\mathbf{r}_>$ provided that $|\mathbf{r}_> - \mathbf{r}'| \neq 0$ and it may be analytically continued to complex values $(\hat{x}_>,\hat{y}_>,\hat{z}_>)$.[27] According to Refs. 28 and 30 a sufficient condition for this is that $\mathbf{r}_>$ be outside the circumscribed sphere C to S (Fig. 1).[30] This is certainly true when $\mathbf{r}_>$ is either in $R^{(-)}$ or $R^{(+)}$ which is also when the representation (10) by means of the angular spectrum converges.[30] If moreover, the boundary S is analytic, then the values $u^{(s)}(\mathbf{r}_>)$ may be analytically continued up to S (Refs. 29 and 31); this is also a necessary, although not sufficient, condition, for the validity of the angular spectrum representation (1) between S and $z = 0$ and $z = L$; this, or even the validity of (1) between S and C is the Rayleigh hypothesis.[29,30] A necessary and sufficient condition for this hypothesis to hold is that $u^{(s)}(\mathbf{r})$ may be analytically continued up to the inscribed sphere C', i.e., its singularities lie inside that sphere C'.[29,30]

Having seen that the scattered field $u^{(s)}(\mathbf{r})$ is an analytic function at least in the exterior domain and providing that $|\mathbf{r}_> - \mathbf{r}'| \neq 0, (\mathbf{r}' \in v)$, it remains to see with more detail what sort of analyticity may characterize $u^{(s)}(\mathbf{r})$.

From either Eq. (12) or (13) one can see that

$$|A^{(\pm)}(\mathbf{K})| < M\gamma^{(\pm)}(\mathbf{K}) \tag{29}$$

for some constant M, where

$$\gamma^{(+)}(\mathbf{K}) = e^{+|q|L}, \tag{30a}$$

$$\gamma^{(-)}(\mathbf{K}) = 1. \tag{30b}$$

Thus, at points $\mathbf{r}_>$ where the angular spectrum representation Eq. (11), is valid, one has that the spatial frequency spectrum $A^{(\pm)}(\mathbf{K})e^{\pm iqz_>}$ is bounded by

$$|A^{(+)}(\mathbf{K})e^{iqz_>}| < Me^{-|q|(z_> - L)} \quad \text{for } z_> \in R^{(+)}, \tag{31a}$$

$$|A^{(-)}(\mathbf{K})e^{-iqz_>}| < Me^{-|q|z_>} \quad \text{for } z_> \in R^{(-)}. \tag{31b}$$

For optical frequencies, given a certain level of noise, and at a distance $z_>$ large enough it would be impossible to distinguish from zero the bound given by Eqs. (31) for

$$|\mathbf{K}| \geqslant Nk_0, \quad (N \gg 1) \tag{32}$$

so that under those conditions the spectrum $A^{(+)}(\mathbf{K})e^{\pm iqz_>}$ is effectively bandlimited to a finite range given by Eq. (32) (see also Ref. 18). Thus, in practice, $u(\mathbf{r}_>)$ may be approximately characterized by a function to which the Plancherel–Polya theorem[23] may be applied for constant $z_>$, i.e., *the function* $u(\mathbf{r}_>)$ *may be considered as the limit for real values* $x_>, y_>$ *of an entire function of exponential type in the space of complex variables* $\hat{x}_> = x_> + ix'_>, \hat{z}_> = z_> + iz'_>$. [When $u(\mathbf{r}_>)$ is source-free then $A^{(\pm)}(\mathbf{K}) = 0$ for $|\mathbf{K}| > k_0$ (Ref. 28) and the above will also hold]. *This is, of course, an approximation taking into account the presence of noise and at* $z_> \in (R^{(+)}, R^{(-)})$ *large enough to record the frequency spectrum as of effective finite support.* (Of course in an ideal, noise-free situation, the bounds (31) will have a certain value, however small, and the above characterization could not be made.)

This scheme is particularly valuable in situations of cylindrical symmetry along the OY axis. Then the scattered field $u^{(s)}(\mathbf{r}_>) = u^{(s)}(x_>, z_>)$ may be extended as an entire function of exponential type into the complex plane of the single complex variable $\hat{x}_> = x_> + ix'_>$ at each $z_>$ constant. As it has been already seen, the information of these functions of one complex variable (in contrast with those of several complex variables) is encoded by the position of its zeros $\hat{x}_n$ in the $\hat{x}_>$-complex plane by means of the Hadamard product, Eq. (22). These zeros lie asymptotically on a line parallel to the real axis (see Refs. 4 and 5 for a detailed discussion of information encoding by zeros), with a spacing given by the Nyquist rate, or Shannon points[19]: $x_n = n\pi/Nk_0$ and enables one to define a zero density[4,5]:

$$\mathscr{D} = \lim_{r\to\infty} \frac{n(r)}{r} = \frac{Nk_0}{\pi} = \frac{2N}{\lambda}, \quad r = |\hat{x}_>|. \tag{33}$$

As the distance $z_>$ varies, the configuration of the zeros $\hat{x}_n$ at each complex plane $\hat{x}_>$ associated to the corresponding value of $z_>$ changes. Therefore, as the scattered field propagates in the exterior domain, the information conveyed by these zeros propagates along "zero trajectories" in a four-dimensional complex space (Fig. 2). This yields another interpretation to Gabor's treatment of information[32]: each zero encodes a degree of freedom or logon of the scattered wavefunction. The propagation of information along zero trajectories may be associated to the concept of tubes of information.[32]

As $z_>$ increases however, the effective support of $A^{(\pm)}(\mathbf{K})e^{\pm iqz_>}$ marked by Eqs. (31) decreases until there is no contribution of evanescent waves. This implies that as the scattered field propagates, the zero configuration at each

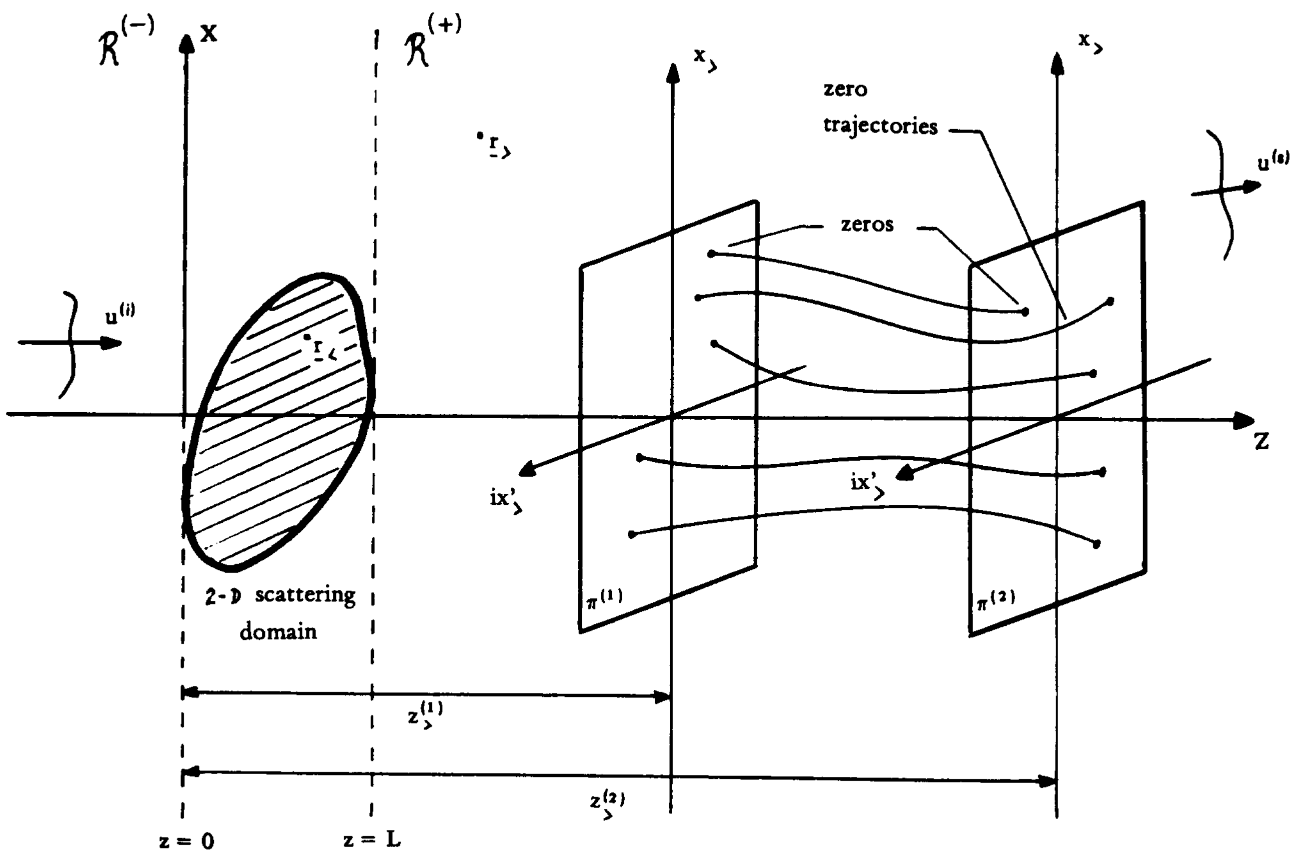


FIG. 2. Propagation of zeros of the scattered field along zero trajectories from the configuration in the $\hat{x}_>$-complex plane $\pi^{(1)}$ associated to the distance $z_>^{(1)}$ to the positions in the $\hat{x}_>$-complex plane $\pi^{(2)}$ associated to the distance $z_>^{(2)} > z_>^{(1)}$.

complex plane associated to the corresponding $z_>$ not only varies, but also the zero density decreases from the value $2N/\lambda$ until a certain $z_>^{(c)}$ at which one reaches the limit marked by the evanescent waves

$$\mathscr{D}_0 = \frac{2}{\lambda}, \tag{34}$$

and from that value of $z_>^{(c)}$ onwards the density remains constantly equal to $\mathscr{D}_0$ as $z_>$ increases.

This impossibility of recovering higher zero density in the reversal process shows the irreversible nature of information loss as the scattered field propagates. From a zero configuration of density $\mathscr{D}_0 = 2/\lambda$ in a complex plane $\hat{x}_>$ associated with a certain distance $z_>$ one can reconstruct the zero configurations at all prior complex planes (i.e., at smaller $z_>$) with the same density $\mathscr{D}_0 = 2/\lambda$; however, from a well known theorem of the theory of entire functions[33] (Carlson's theorem), no entire function can have a density of zeros larger than T/π, if T is its bandwidth, unless it is identically zero. Thus, no entire function can have a zero density larger than $\mathscr{D}_0 = 2/\lambda$ if evanescent waves are excluded. This fact was already noted by Schmidt-Weinmar[21] who quoted the impossibility of obtaining information about the field near the sources from the far field containing only homogeneous waves.

Thus, as one considers the situation in which the conjugate wave reproduces in the reversal process the field propagation; i.e., as $z_>$ tends to a plane $z = z_0$, which is the nearest plane to the center of the scattering volume, at which the angular spectrum representation (11) converges, the computation of $u^{(s)}(\hat{x}_>, z_>)$ at $z_> > z_>^{(c)} > z_0$ will yield only a replica of $u^{(s)}(\hat{x}_0, z_0)$ with lower resolution limited to $\lambda/2$, i.e., with smaller zero density $\mathscr{D}_0 = 2/\lambda$. This is in agreement with the lack of convergence as a simple limit of $u^{(s)}(\hat{x}_>, z_>)$ to $u^{(s)}(\hat{x}_0, z_0)$, but as a limit in the mean, put forward by Sherman.[28] Only when $u^{(s)}(\hat{x}_0, z_0)$ also has a zero density $\mathscr{D}_0 = 2/\lambda$, the convergence will be in the form of an ordinary limit [in particular, when $u^{(s)}(\mathbf{r})$ is source-free[28]], since then all the zeros of $u^{(s)}(\hat{x}_0, z_0)$ may be found by following backwards the zero trajectories starting from the zero configuration of $u^{(s)}(\hat{x}_>, z_>)$.

4. SUMMARY AND CONCLUSIONS

We have generalized and exploited further some aspects of the analyticity properties of scattered scalar fields.

First, performing the analytic continuation of the 3-D Fourier transform F of the source function, we have seen that a single scattering experiment, or a discrete number of them, yields an angular spectrum which is given by the values of F along a line, or lines, on which it may have a zero since it is the boundary value on the real axes of an entire function of exponential type in a 6-D complex space and, as such, it possesses an infinity of zero lines.

Scattering and diffraction are usually regarded as processes creating evanescent waves that are added to the usual homogeneous components of the scattered field. Changing the ζ-path of integration we have seen that scattering and diffraction may be equally envisaged as phenomena in which only inhomogeneous (leaky) wave are produced (a study of these waves at planar interfaces may be found in Ref. 34). These inhomogeneous waves correspond to a certain analytic continuation of the scattering amplitude into the complex space $\hat{\mathbf{K}}$ avoiding the corresponding branch cuts. [This continuation of $A(\mathbf{K})$ has been already used in order to put forward the equivalence between the solution to the phase problem and the location of the zeros of $A(\hat{K})$.[4] Also the use of inhomogeneous wave representations has been employed in order to obtain information of subwavelength scattering sources[35]].

Let us observe that if the source is planar then $A(\hat{\mathbf{K}})$ is an entire function (see also Wolf in Refs. 2 and 21).

Secondly, the scattered field is described in its free propagation outside the scattering volume by an analytic function of the spatial variables. (On the other hand, the generalized analytic character of this field in 2-D problems in the complex variable $x + iz$ leads to integral theorems such as the Helmholtz–Kirchhoff formula and the extinction theorem[36]). *The description of the field propagation based on Huygens principle is ultimately a consequence of the possibility of performing an analytic continuation of the wave function outside the scattering volume.* At optical frequencies, when evanescent waves from a certain spatial frequency onwards may be neglected given a certain level of noise, this analyticity is more specific: the field may be approximated by an entire function of exponential type. In two-dimensional problems, at each distance $z_>$ from the object the scattered wave function is therefore described by its zeros in the $\hat{x}_>$-plane having a certain density $2N/\lambda$ ($N \geqslant 1$). As $z_>$ increases this density decreases until one reaches the region at which the evanescent waves may be neglected. This yields a density of zeros $2/\lambda$ and establishes the impossibility of fully reconstructing the field with higher zero densities at smaller distances from the object in the reversal process; and also poses the resolution limit $\lambda/2$ for the reconstruction of object details (subwavelength source reconstruction requires therefore the localization of zeros with densities above $2/\lambda$). Only in the case of source-free fields, for which the density keeps constantly equal to $2/\lambda$ in propagation, this reconstruction by means of the inverse process is straightforward.

ACKNOWLEDGMENTS

The author wishes to acknowledge many interesting discussions with Drs. G. Ross and M. A. Fiddy of Queen Elizabeth College, London, that motivated this work.

He also wants to thank a referee for many valuable comments.

[1]D. L. Colton, *Analytic Theory of Partial Differential Equations* (Pitman, London, 1980).

[2]K. Miyamoto and E. Wolf, J. Opt. Soc. Am. **52**, 615–625 (1962); E. Wolf, J. Opt. Soc. Am. **68**, 1597–1605 (1978).

[3]G. Ross, M. A. Fiddy, M. Nieto-Vesperinas, and I. Manolitsakis, Opt. Acta **26**, 229–238 (1979).

[4]G. Ross, M. A. Fiddy, and M. Nieto-Vesperinas, "The inverse scattering problem in structural determinations," in *Inverse Scattering Problems in Optics*, edited by H. P. Baltes, *Topics in Current Physics*, Vol. 20 (Springer-Verlag, Berlin, 1980), Chap. 2, pp. 15–17.

[5]M. A. Fiddy, G. Ross, M. Nieto-Vesperinas, and A. M. J. Huiser, Opt. Acta **29**, 23–40 (1982).

[6]G. Ross, M. A. Fiddy, and H. Moezzi, Opt. Acta **27**, 1433–1451 (1980).
[7]J. G. Walker, Opt. Acta **28**, 1017–1019 (1981); Opt. Acta **28**, 735–738 (1981).
[8]J. W. Wood, M. A. Fiddy, and R. E. Burge, Opt. Lett. 514–516 (1982).
[9]M. A. Fiddy and G. Ross, Opt. Acta **26**, 1139–1146 (1979).
[10]Y. M. Bruck and L. G. Sodin, Opt. Commun. **30**, 304–308 (1979).
[11]A. M. J. Huiser and P. Van Toorn, Opt. Lett. **5**, 499–501 (1980).
[12]I. Manolitsakis, J. Math. Phys. **23**, 2291–2298 (1982).
[13]N. Bleistein and J. K. Cohen, J. Math. Phys. **18**, 194–201 (1977).
[14]A. J. Devaney, J. Math. Phys. **19**, 1526–1531 (1978).
[15]E. Wolf, Opt. Comm. **1**, 153–156 (1969); A. Baños, *Dipole Radiation in the Presence of a Conducting Half-space* (Pergamon, New York, 1966), Eq. (2–19).
[16]D. N. Pattanayak and E. Wolf, Phys. Rev. D **13**, 913–923 (1976).
[17]As is well known, Eq. (1) represents the leading term of the asymptotic expression of (11) for $k_0 r \rightarrow \infty$.[2] As such, the argument $k_0\mathbf{s}$ in Eq. (1) is restricted to be $|k_0\mathbf{s}_\perp| \leqslant |k_0\mathbf{s}| = k_0$ at difference with the vector $\mathbf{K}$ defined in Eq. (11) that can take any real value.
[18]J. R. Shewell and E. Wolf, J. Opt. Soc. Am. **58**, 1596–1603 (1968).
[19]J. W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill, New York, 1968).
[20]W. H. Carter, J. Opt. Soc. Am. **65**, 1054–1058 (1975).
[21]H. G. Schmidt-Weinmar, J. Opt. Soc. Am. **65**, 1059–1066 (1975).
[22]G. S. Agarwal and E. Wolf, J. Opt. Soc. Am. **72**, 321–326 (1982).
[23]B. A. Fucks, "Introduction to the Theory of Analytic Functions of Several Complex Variables," Trans. Math. Monographs, Vol. 8 (Amer. Math. Soc., Providence, RI, 1983).
[24]A. T. Friberg and E. Wolf, J. Opt. Soc. Am. **73**, 26–32 (1983).
[25]W. F. Osgood, *Lehrbuch der Funktionentheorie* (Teubner, Leipzig, 1929), Vol. 2.
[26]R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1962), Vol. 2.
[27]P. R. Garabedian, *Partial Differential Equations* (Wiley, New York, 1964).
[28]G. C. Sherman, Phys. Rev. Lett. **21**, 761–764 (1968), J. Opt. Soc. Am. **59**, 697–711 (1969).
[29]R. F. Millar, Proc. Camb. Phil. Soc. **65**, 773–791 (1969).
[30]R. F. Millar, Radio Sci. **8**, 785–796 (1973); A. J. Devaney and G. C. Sherman, SIAM Rev. **15**, 765–786 (1973).
[31]A result that may be obtained from the Mie scattering is that the scattered field is analytic up to the origin of a perfectly conducting sphere. Although the analytic continuation of the field inside the sphere differs from the physical field which is actually zero, [V. H. Weston and J. J. Bowman, Ergun Ar, Arch. Rational Mech. Anal. **31**, 199–213 (1968); B. J. Hoenders, "The Uniqueness of Inverse Problems," in *Inverse Source Problems in Optics*, edited by H. P. Baltes, Chap. 3, pp.41–82, *Topics in Current Physics*, Vol. 8 (Springer-Verlag, Berlin, 1978)]. I recognize a referee for having drawn attention to this point.
[32]D. Gabor, "Light and information," in *Proceedings of the Symposium on Astronomical Optics and Related Subjects*, edited by Z. Kopal (North-Holland, Amsterdam, 1956), pp. 17–30.
[33]R. P. Boas, *Entire Functions* (Academic, New York, 1954).
[34]T. Tamir, Optik **38**, 269–297 (1973).
[35]H. G. Schmidt-Weinmar, "Spatial Resolution of Subwavelength Sources from Optical-Far-Zone Data," in *Inverse Source Problems in Optics*, edited by H. P. Baltes, Chap. 4, *Topics in Current Physics*, Vol. 9 (Springer-Verlag, Berlin, 1978).
[36]M. Nieto-Vesperinas, "On the representation of two-dimensional scattered fields in the complex plane," J. Math. Phys. **25**, 1592 (1984).

Some remarks on the application of the QCA to the determination of the overall elastic response of a matrix/inclusion composite

J. R. Willis
School of Mathematics, University of Bath, Bath, BA2 7AY, England

(Received 10 June 1983; accepted for publication 6 October 1983)

A "multiple scattering" formulation for the elastic response of a matrix containing a random distribution of inclusions is considered. The "ordinary" quasicrystalline approximation (QCA) is reviewed, together with a recently proposed "self-consistent" QCA. It is shown that the basic "self-consistent" postulate can be implemented without making an extra approximation in relation to scattering from the "background" material and that, when this extra approximation is dispensed with, the alternative self-consistent prescription that results is capable in principle of generating the overall properties exactly. In practice, there is still the need to close a hierarchy by making some assumption about interactions between inclusions. When the QCA is adopted, the alternative self-consistent prescription yields precisely the same estimates for overall behavior as the "ordinary" QCA. The formulation is given explicitly for elastostatics but the conclusions depend only upon the algebraic structure of the problem and apply equally well to problems of wave propagation.

PACS numbers: 46.30.Cn, 03.40.Kf, 61.70.Rj

I. INTRODUCTION

Attempts to determine the overall response of a medium containing a random distribution of inclusions usually aim at finding equations which govern the ensemble averages of the fields in question. This is done by selecting some particular form of the equations that describe the behavior of any realization of the random medium and then taking ensemble averages, conditional upon a certain number of inclusions being fixed. This procedure generates a hierarchy of equations which involve a sequence of increasingly more complicated conditional expectations. The hierarchy is reduced to a closed set of equations by performing some kind of truncation, which may be achieved either by truncating a perturbation series[1,2] or by adopting some *ad hoc* closure assumption. The truncation of a perturbation series provides an approximation whose status is clear enough, but it is not useful in practice unless the properties of the medium vary only slightly. Approximations based upon a closure assumption have no absolute validity (except that in some circumstances they can be shown to provide bounds) but they do yield estimates for overall behavior which are not restricted to small variations in local properties.

One of the most widely used closure assumptions is the quasicrystalline approximation (QCA), which states essentially that the expected field in one inclusion is insensitive to the precise location of any other inclusion. In its basic form, it is applied to a formulation in which each inclusion acts as a "scatterer," embedded in matrix material.[3–5] An extension, formulated for elastic waves by Devaney,[6] describes the scattering of the inclusions relative to a "background" or "comparison" medium, which is different from the matrix. A "self-consistent" description of overall behavior then follows from the postulate that the mean "scattering" is zero, when the "comparison" medium is chosen as the "overall" medium; this postulate generates an equation (or set of equations) for the overall properties. In practice, the prescription can be implemented only approximately, because the hierarchy of equations has to be closed; Devaney proposed for this purpose an extension of the QCA, in which (in a sense made precise later) the response of the "background" material is insensitive to the positions of inclusions. It is shown in this work that the basic self-consistent prescription of Devaney can be implemented while invoking only the "ordinary" QCA, because the response of the "background" medium can be determined explicitly in terms of the response of the inclusions. When this is done, it emerges that the self-consistent scheme in fact generates precisely the same estimates for overall response as the "ordinary" QCA, formulated relative to the matrix.

For simplicity, the development is carried out explicitly for the elastostatic response of a matrix containing a single population of inclusions, but it is pointed out that the conclusions remain valid for any matrix/inclusion composite. The reasoning is algebraic and the relationships that are displayed apply to any problem with the same algebraic structure, even though its analytic structure might be different. In particular, the relationships remain valid for the problems of elastic wave propagation considered by Devaney.

II. FORMULATION

An elastic body occupies a domain Ω and is subjected to body force f per unit volume and boundary conditions that either displacement or traction components, or some combination of these, are prescribed on the boundary $\partial\Omega$. Symbolically, stress σ is related to strain e by

$$\sigma = Le; \tag{2.1}$$

in components this would read $\sigma_{ij} = L_{ijkl}\, e_{kl}$. Components of strain are related to components of displacement by $e_{ij} = u_{(i,j)}$, the suffix j representing differentiation with respect to x_j and the bracket representing symmetrization. The L_{ijkl} denote the components of the tensor of elastic moduli. The basic problem is to solve the equilibrium equations

$$\operatorname{div} \sigma + f = 0, \quad x \in \Omega, \tag{2.2}$$

 0022-2488/84/062116-05$02.50

where σ is related to e through (2.1), together with the boundary conditions.

The body to be considered comprises a matrix with tensor of moduli L_2, in which are embedded inclusions, composed of material with tensor of moduli L_1. The inclusions occupy domains Ω_A, $A = 1,2,..., N$, whose locations are defined by specifying the position x_A of a chosen point, called for convenience the center of Ω_A. The body is considered to be a random medium in the sense that the points x_A are not given precisely but are distributed according to some stochastic process. The stress, strain, and displacement fields depend upon the locations x_A of the inclusions and the objective will be to find their ensemble averages, $\langle\sigma\rangle$, $\langle e\rangle$, $\langle u\rangle$. In principle, this could be done by finding $e(x)$, etc., for all possible locations x_A and then taking the ensemble average. In practice, this is not feasible, both because $e(x)$ could not be found and because the stochastic process will not be completely known. Instead, only limited statistical information will be available, such as the probability density $P_A = P(x_A)$ for finding an inclusion centered at x_A, and the conditional density $P_{B|A} = P(x_B|x_A)$ for finding an inclusion centered at x_B, given that an inclusion is centered at x_A. Correspondingly, approximations to $\langle e\rangle$ and other mean fields must be sought, which make use of this limited information.

It is convenient to replace the system (2.1) and (2.2) by an integral equation, by introducing a "comparison medium" with moduli L_0 and setting

$$\sigma = L_0 e + \tau, \tag{2.3}$$

so that

$$\tau = (L - L_0)e, \tag{2.4}$$

the tensor L_0 at the moment being arbitrary. Then, (2.2) gives

$$\operatorname{div}(L_0 e) + \operatorname{div}\tau + f = 0, \tag{2.5}$$

and, if Green's function for the comparison medium is known, there follows the representation

$$e = e_0 - \Gamma_0\tau. \tag{2.6}$$

Here, e_0 is the solution of the given boundary value problem for the comparison medium and Γ_0 is an operator related to its Green's function G_0. Explicitly, Γ_0 is an integral operator with components

$$(\Gamma_0)_{ijkl} = \frac{\partial^2[G_0(x, x')]_{jk}}{\partial x_i\,\partial x_l'}\bigg|_{(ij)(kl)}, \tag{2.7}$$

and its singularity is interpreted in the sense of generalized functions. Equations (2.4) and (2.6) now generate the operator equation

$$(L - L_0)^{-1}\tau + \Gamma_0\tau = e_0. \tag{2.8}$$

Now from (2.6)

$$\langle e\rangle = e_0 - \Gamma_0\langle\tau\rangle, \tag{2.9}$$

so that the problem is solved once $\langle\tau\rangle$ is found. It is convenient to use (2.9) to replace (2.8) by

$$(L - L_0)^{-1}\tau + \Gamma_0(\tau - \langle\tau\rangle) = \langle e\rangle. \tag{2.10}$$

Suppose that this equation has solution

$$\tau = T_0\langle e\rangle, \tag{2.11}$$

so that T_0 is an operator, closely related to Devaney's "transition operator." Then, from (2.3),

$$\langle\sigma\rangle = \hat{L}\,\langle e\rangle, \tag{2.12}$$

where $\hat{L}$ is an "effective modulus operator" defined by

$$\hat{L} = L_0 + \langle T_0\rangle. \tag{2.13}$$

Also, from (2.9) and (2.11),

$$\langle e\rangle = (I + \Gamma_0\langle T_0\rangle)^{-1}e_0. \tag{2.14}$$

If, in particular, e_0 is the strain produced by a delta-function body force, with homogeneous boundary conditions, it becomes the strain E_0 associated with the Green's function G_0 for the comparison body, and $\langle e\rangle$ becomes the strain $\hat{E}$ associated with the Green's function $\hat{G}$ for the "effective" body with elastic modulus operator $\hat{L}$. In this case, (2.14) implies

$$\hat{E} = (I + \Gamma_0\langle T_0\rangle)^{-1}E_0. \tag{2.15}$$

Here, E_0 satisfies

$$\operatorname{div}(L_0 E_0) + I = 0, \tag{2.16}$$

and, in view of (2.12), $\hat{E}$ satisfies

$$\operatorname{div}(\hat{L}\hat{E}) + I = 0. \tag{2.17}$$

This relation also follows from (2.15) and (2.16), since, by its definition, Γ_0 has the property

$$\operatorname{div}(L_0\Gamma_0\tau) = \operatorname{div}\tau, \tag{2.18}$$

for any τ. Furthermore, if approximations to $\hat{L}$ and $\hat{E}$ are generated from (2.13) and (2.15) using an approximate $\langle T_0\rangle$, then these approximations also satisfy (2.17).

This section is concluded by noting properties of the operator Γ_0 that will be useful in the sequel. If (2.4) is substituted into (2.6), there results the equation

$$e = e_0 - \Gamma_0(L - L_0)e. \tag{2.19}$$

This equation applies to any boundary value problem, for a medium with tensor of elastic moduli L. If, in particular, $e_0 = E_0$, as above, but L is taken equal to L_2, so that, correspondingly, $e = E_2$, then (2.19) gives

$$E_2 = E_0 - \Gamma_0(L_2 - L_0)E_2. \tag{2.20}$$

It is also possible to apply a system of force dipoles so that the strain $e_0 = \Gamma_0$ is produced in the comparison material. Then, the strain e generated in a material with tensor of moduli L_2 is $e = \Gamma_2$ and, from (2.19),

$$\Gamma_2 = \Gamma_0 - \Gamma_0(L_2 - L_0)\Gamma_2. \tag{2.21}$$

Finally, the roles of L_0 and L_2 can be interchanged to give also

$$\Gamma_0 = \Gamma_2 - \Gamma_2(L_0 - L_2)\Gamma_0. \tag{2.22}$$

Equation (2.19) is the Lipmann–Schwinger equation and the relations that follow from it are of course well known.[7]

III. THE QUASICRYSTALLINE APPROXIMATION (QCA)

This section addresses the problem of constructing approximations to $\langle T_0\rangle$, by applying various forms of the quasicrystalline approximation (QCA).

A. The ordinary QCA

If L_0 is chosen equal to L_2 (so that Γ_0, e_0, T_0 become Γ_2, e_2, T_2, respectively), then $\tau = 0$ throughout the matrix and τ can be written

$$\tau(x) = \sum_A \tau^A(x), \tag{3.1}$$

where τ^A represents the restriction of τ to Ω_A. Equation (2.10) then gives

$$(L_1 - L_2)^{-1}\tau^A + \Gamma_2\tau^A + \Gamma_2\Big(\sum_{B\neq A} \tau^B - \langle\tau\rangle\Big) = \langle e\rangle,$$
$$x \in \Omega_A, \tag{3.2}$$

and is both indeterminate and irrelevant when x lies in the matrix. Now take the expectation of (3.2), conditional upon finding an inclusion centered at x_A. This gives

$$(L_1 - L_2)^{-1}\tau_A^A + \Gamma_2(\tau_A^A - \langle\tau\rangle^A) + \Gamma_2 g = \langle e\rangle,$$
$$x \in \Omega_A, \tag{3.3}$$

where $\langle\tau\rangle^A$ represents the restriction of $\langle\tau\rangle$ to Ω_A and

$$g(x) = \begin{cases} \displaystyle\int dx_B\,[\tau_{BA}^B(x)P_{B|A} - \tau_B^B(x)P_B], & x \notin \Omega_A, \\ 0, & x \in \Omega_A \end{cases}. \tag{3.4}$$

Here, τ_A^A represents the expected value of τ^A, conditional upon finding an inclusion centered at x_A, and τ_{BA}^B is the expected value of τ^B, conditional upon finding distinct inclusins centered at x_B and x_A. Equation (3.3) cannot be solved for τ_A^A because of the presence of τ_{BA}^B in the definition of $g(x)$. An approximate equation, which can be solved, will be discussed below. However, τ_A^A depends upon $\langle e\rangle$ and it is useful to write

$$\tau_A^A = T_A^A\langle e\rangle, \tag{3.5}$$

where T_A^A is an operator, equal to the indicated conditional expectation of T_2. Then, since τ is zero in the matrix,

$$\langle T_2\rangle = \int dx_A\, T_A^A\, P_A, \tag{3.6}$$

and, from (2.13),

$$\widehat{L} = L_2 + \int dx_A\, T_A^A\, P_A. \tag{3.7}$$

The "ordinary" QCA consists of replacing (3.3) by a closed equation, by making the approximation

$$\tau_{BA}^B = \tau_B^B. \tag{3.8}$$

The solution of the resulting equation yields an approximation to T_A^A and so by (3.7), to an approximation $\widehat{L}^{\text{ord}}$ for $\widehat{L}$, the superscript "ord" being used as a reminder that it is obtained from the ordinary QCA. This result can also be obtained by substituting simple trial fields into variational principles of the Hashin–Shtrikman type, so that $\widehat{L}^{\text{ord}}$ also has some status as a variational approximation independent of the validity or otherwise of the QCA.[8]

B. The self-consistent QCA

The foregoing reasoning can be repeated with an arbitrary L_0, which can even be allowed to be a nonlocal operator of the same type as $\widehat{L}$. Then, in place of (3.1),

$$\tau(x) = \sum_A \tau^A(x) + \tau^m(x), \tag{3.9}$$

where τ^m represents the restriction of τ to the matrix material. Also, (3.2) is replaced by

$$(L_1 - L_0)^{-1}\tau^A + \Gamma_0\tau^A + \Gamma_0\Big(\sum_{B\neq A}\tau^B + \tau^m - \langle\tau\rangle\Big) = \langle e\rangle, \quad x \in \Omega_A, \tag{3.10}$$

and, by taking expectations conditional upon x_A being prescribed,

$$(L_1 - L_0)^{-1}\tau_A^A + \Gamma_0(\tau_A^A - \langle\tau\rangle^A) + \Gamma_0(g + f) = \langle e\rangle,$$
$$x \in \Omega_A, \tag{3.11}$$

where $g(x)$ is defined as in (3.4) and

$$f(x) = \begin{cases} \tau_{mA}^m(x)p_{m|A}(x) - \tau_m^m\, p_m(x), & x \notin \Omega_A, \\ 0, & x \in \Omega_A. \end{cases} \tag{3.12}$$

Here, $p_m(x)$ represents the probability of finding matrix at x and $p_{m|A}$ is the corresponding probability, conditional upon finding an inclusion centered at x_A; τ_m^m and τ_{mA}^m are defined similarly. In (3.10) and (3.11),

$$\langle\tau\rangle(x) = \int dx_A\, \tau_A^A(x)P_A + \tau_m^m(x)p_m(x), \tag{3.13}$$

and, from (2.4), there is the additional relation

$$(L_1 - L_0)^{-1}\int dx_A\, \tau_A^A(x)P_A + (L_2 - L_0)^{-1}\tau_m^m(x)p_m(x) = \langle e\rangle(x). \tag{3.14}$$

Equation (3.11) is not soluble for τ_A^A because of the presence of τ_{BA}^B and τ_{mA}^m; in different notation, Devaney[6] proposes the approximations

$$\tau_{BA}^B = \tau_B^B, \quad \tau_{mA}^m = \tau_m^m \tag{3.15}$$

as an extension of the QCA. With these approximations, (3.11) and (3.14) can be solved for τ_A^A and τ_m^m for any choice of L_0. The "self-consistent" prescription is that the estimate obtained for $\widehat{L}$ would be exact, if L_0 were identified with $\widehat{L}$. In view of (2.13), this is equivalent to asserting that $L_0 = \widehat{L}^{\text{sc}}$ when it is chosen so that

$$\langle\tau\rangle = 0. \tag{3.16}$$

The prescription generates the approximation $\widehat{L}^{\text{sc}}$ rather than $\widehat{L}$, because it depends upon the approximations (3.15). However, like the ordinary QCA, the estimate for $\widehat{L}$ outlined here can be obtained from variational principles of the Hashin–Shtrikman type, even for general L_0 and, in this framework too, it is possible to postulate that (3.16) holds when $L_0 = \widehat{L}$. The self-consistent prescription (3.16) has been discussed from this standpoint by Willis[9] for statics and Talbot and Willis[10] for dynamics.

C. An "alternative" self-consistent QCA

It is, in fact, possible to proceed with a modification of the foregoing "self-consistent" scheme, which does not require the second of the assumptions (3.15). The information that fixes τ_{mA}^m comes from averaging (2.10), conditionally upon Ω_A being fixed, when x lies outside Ω_A. This gives

$$(L_1 - L_0)^{-1} \int dx_B\, \tau_{BA}^B(x) P_{B|A} + (L_2 - L_0)^{-1} \tau_{mA}^m(x) p_{m|A}(x) + \Gamma_0(g + f) + \Gamma_0(\tau_A^A - \langle \tau \rangle^A) = \langle e \rangle, \quad x \notin \Omega_A, \tag{3.17}$$

where g and f are defined as in (3.4) and (3.12). It may be noted that, when x is far from Ω_A, (3.17) implies (3.14) precisely. Subtracting (3.14) from (3.17) gives

$$(L_1 - L_0)^{-1} g(x) + (L_2 - L_0)^{-1} f(x) + \Gamma_0(g + f) + \Gamma_0(\tau_A^A - \langle \tau \rangle^A) = 0, \quad x \notin \Omega_A. \tag{3.18}$$

This equation can be solved for f, given g, τ_A^A, and $\langle \tau \rangle^A$. In fact, f can be represented in the form

$$f = h^A - (L_2 - L_0)\Gamma_2[h^A + \tau_A^A - \langle \tau \rangle^A + (L_1 - L_2)(L_1 - L_0)^{-1} g] - (L_2 - L_0)(L_1 - L_0)^{-1} g, \tag{3.19}$$

where h^A is defined over Ω_A and satisfies the equation

$$(L_0 - L_2)^{-1} h^A + \Gamma_2[h^A + \tau_A^A - \langle \tau \rangle^A + (L_1 - L_2)(L_1 - L_0)^{-1} g] = 0, \quad x \in \Omega_A. \tag{3.20}$$

This result is justified in Appendix A, which also provides some motivation for seeking this type of representation.

Substituting (3.19) for f into (3.11) and using (2.21) gives the remarkably simple relation

$$(L_1 - L_0)^{-1} \tau_A^A + (L_2 - L_0)^{-1} h^A = \langle e \rangle, \quad x \in \Omega_A. \tag{3.21}$$

Eliminating h^A between (3.20) and (3.21) now gives

$$(L_1 - L_0)^{-1} \tau_A^A + \Gamma_2[(L_1 - L_2)(L_1 - L_0)^{-1}(\tau_A^A + g) (L_2 - L_0)\langle e \rangle^A - \langle \tau \rangle^A] = \langle e \rangle, \quad x \in \Omega_A, \tag{3.22}$$

and this can be rearranged, using (3.13) and (3.14), to give

$$(L_1 - L_2)^{-1}[(L_1 - L_2)(L_1 - L_0)^{-1} \tau_A^A] + \Gamma_2\left[(L_1 - L_2)(L_1 - L_0)^{-1}\left(\tau_A^A - \left(\int dx_B\, \tau_B^B P_B\right)^A + g\right)\right] = \langle e \rangle, \quad x \in \Omega_A. \tag{3.23}$$

This equation is identical to (3.3), except that τ_A^A and τ_{BA}^B are now multiplied by $(L_1 - L_2)(L_1 - L_0)^{-1}$. Its solution, therefore, is

$$\tau_A^A = (L_1 - L_0)(L_1 - L_2)^{-1} T_A^A \langle e \rangle. \tag{3.24}$$

Now from (3.13), (3.14), and (3.24),

$$\langle \tau \rangle = (L_2 - L_0)\langle e \rangle + \int dx_A\, T_A^A P_A \langle e \rangle. \tag{3.25}$$

The self-consistent condition (3.16) implies, this time, that $L_0 = \hat{L}$, as in (3.7). The solution would be exact if T_A^A were known exactly. If T_A^A is estimated by making the QCA (3.8), but no other approximation, then (3.16) generates $L_0 = \hat{L}^{\text{ord}}$ as an "alternative" self-consistent estimate.

IV. DISCUSSION

The principal interest of the "alternative" self-consistent scheme is that it replaces the *approximation* $(3.15)_2$ by the *exact relation* (3.17); this motivated the writer to describe the "alternative" scheme as an "improved" scheme until this terminology was resisted by a referee. It is perhaps worth admitting that some brand new estimate for overall moduli was initially expected to result from the use of (3.17), and that the realization that it had to lead to $\hat{L}^{\text{ord}}$ came only after laborious explicit solution of (3.17) in particular cases. It is not suggested that the "alternative" scheme is always likely to produce better agreement with observation than the "ordinary" self-consistent scheme, since it is clearly possible that the errors in the two approximations (3.15) might tend to cancel in particular cases; indeed, Talbot and Willis[11] have expressed a preference for the "ordinary" scheme when applied to a suspension of glass spheres in an epoxy matrix. There are, however, examples,[12] even for static behavior, where the self-consistent scheme clearly overestimates the effects of interactions: rigid spheres generate infinite overall moduli, and spherical cavities generate zero overall moduli, at concentration 0.5. The "alternative" scheme, coupled with the QCA, has been shown to give the same estimate as the "ordinary" QCA, and this tends to underestimate interactions (producing lower or upper bounds, respectively, for the two extreme cases just quoted). It would appear, therefore, that *systematic* improvement could be guaranteed *for any composite* only by making more detailed allowance for interactions between inclusions.

A similar discussion can be given for a matrix containing any number of different types of inclusion; the only complication is in the notation. Section III C of course offers no alternative to the usual self-consistent estimate[9,13] for a composite such as a polycrystal, which has no clearly defined matrix phase.

Finally, although the discussion has been given for static problems, it is remarked that the relations that have been noted depend only upon the algebraic property (2.21). For dynamic problems a four-dimensional formalism can be introduced, in which strain e is augmented to strain and velocity $(e, \dot{u})$, stress is augmented to stress and momentum density (σ, p), and the constitutive relation (2.1) becomes augmented by

$$p = \rho \dot{u}, \tag{4.1}$$

where ρ represents mass density. Finally, the stress polarization τ is augmented by momentum polarization π, defined by[5]

$$\pi = (\rho - \rho_0)\dot{u}. \tag{4.2}$$

Further details are given in Appendix B. Although this formalism provides no quick route to finding explicit solutions, it demonstrates that the dynamic problem has the same algebraic structure as the static problem, even though its analytic structure is different. The relations between the various approximations discussed in Sec. III, being purely algebraic, thus apply equally well to the dynamic problem.

APPENDIX A: A REPRESENTATION FOR THE FUNCTION f

The solution of (3.18) can be expressed in the form

$$f(x) = f_1(x) - (L_2 - L_0)(L_1 - L_0)^{-1} g(x), \tag{A1}$$

where f_1 satisfies the equation

$$(L_2 - L_0)^{-1} f_1(x) + (\Gamma_0 f_1)(x) + (\Gamma_0 g_1)(x) = 0, \quad x \notin \Omega_A, \tag{A2}$$

with

$$g_1(x) = \begin{cases} \tau_A^A - \langle\tau\rangle^A, & x \in \Omega_A, \\ (L_1 - L_2)(L_1 - L_0)^{-1} g(x), & x \notin \Omega_A. \end{cases} \tag{A3}$$

To motivate the solution of (A2), let

$$e_1 = -\Gamma_0 f_1. \tag{A4}$$

Then, by multiplying (A2) through by $(L_2 - L_0)$, taking the divergence and using the property (2.18), it follows that

$$\mathrm{div}(L_2 e_1) - \mathrm{div}[(L_2 - L_0)\Gamma_0 g_1] = 0, \quad x \notin \Omega_A. \tag{A5}$$

Thus, the strain field e_1 would be produced in an infinite medium with moduli L_2, by the body-force distribution indicated in (A5) for $x \notin \Omega_A$, and some body-force, as yet unknown, for $x \in \Omega_A$. This motivates writing

$$e_1 = -\Gamma_2[(L_2 - L_0)\Gamma_0 g_1 + h^A] = -\Gamma_2 h^A + (\Gamma_0 - \Gamma_2) g_1, \tag{A6}$$

where h^A is zero except over Ω_A. The second equality in (A6) follows from (2.22). Now also, directly from (A2),

$$f_1 = -(L_2 - L_0)\Gamma_2(h^A + g_1), \quad x \notin \Omega_A. \tag{A7}$$

To restrict h^A so that f_1 really does satisfy (A2), it is convenient to define f_1^* by (A7), for all x. Then, by use of (2.21) and (A6), it follows that

$$\Gamma_0 f_1^* = -e_1 - \Gamma_0 h^A, \tag{A8}$$

and so, from (A4),

$$\Gamma_0(f_1^* + h^A - f_1) = 0. \tag{A9}$$

This suggests that

$$f_1 = h^A - (L_2 - L_0)\Gamma_2(h^A + g_1), \tag{A10}$$

and, since $f_1 = 0$ when $x \in \Omega_A$,

$$(L_0 - L_2)^{-1} h^A + \Gamma_2(h^A + g_1) = 0, \quad x \in \Omega_A. \tag{A11}$$

Equation (A11) is interesting in that it defines a problem for an "inclusion" with moduli L_0 occupying Ω_A, in a matrix with moduli L_2. That (A10) actually provides the solution to (A2) can be checked by direct substitution, using (2.21).

APPENDIX B: GENERALIZATION TO DYNAMICS

This appendix gives in slightly more detail the correspondence between the static case discussed explicitly in the text and its generalization to dynamics. It is relegated to this position because it is a rather unattractive hybrid, being neither fully three- or four-dimensional. First, if σ^* is taken to denote the set of quantities (σ_{ij}, p_i) and e^* denotes the set $(e_{ij}, \dot{u}_i)$, the relation

$$\sigma^* = L^* e^* \tag{B1}$$

corresponding to (2.1) represents, in components,

$$\sigma_{ij} = L_{ijkl}\, e_{kl}, \quad p_i = \rho \dot{u}_i, \tag{B2}$$

and the equations of motion [which correspond to the equilibrium equation (2.2)] can be written

$$\mathrm{div}^* \sigma^* + f = 0, \tag{B3}$$

with $\mathrm{div}^* \sigma^*$ representing a three-vector with components $\sigma_{ij,j} - \dot{p}_i$, $i = 1,2,3$. If τ^* represents the set (τ_{ij}, π_i), Eqs. (2.3)–(2.5) remain valid with asterisks placed as indicated for (B1), (B3), and, as shown by Willis,[5]

$$\begin{aligned} e &= e_0 - S_x\, \tau - M_x\, \pi, \\ \dot{u} &= \dot{u}_0 - S_t\, \tau - M_t\, \pi, \end{aligned} \tag{B4}$$

where the integral operators S_x, M_x, S_t, M_t have kernels whose components are

$$(S_x)_{ijkl} = \left.\frac{\partial^2[G_0(x, x', t - t')]_{jk}}{\partial x_i\, \partial x_l'}\right|_{(ij)(kl)}, \tag{B5}$$

$$(M_x)_{ijk} = \left.\frac{[\partial^2 G_0(x, x', t - t')]_{ik}}{\partial x_j\, \partial t}\right|_{(ij)} = (S_t)_{kij}, \tag{B6}$$

$$(M_t)_{ij} = \frac{\partial^2[G_0(x, x', t - t')]_{ij}}{\partial t^2}. \tag{B7}$$

Equation (B4) can be expressed in the form

$$\begin{bmatrix} e \\ \dot{u} \end{bmatrix} = \begin{bmatrix} e_0 \\ \dot{u}_0 \end{bmatrix} - \begin{bmatrix} S_x & M_x \\ S_t & M_t \end{bmatrix} \begin{bmatrix} \tau \\ \pi \end{bmatrix}, \tag{B8}$$

or, still more concisely, as

$$e^* = e_0^* - \Gamma_0^*\, \tau^*, \tag{B9}$$

with the obvious definition of Γ_0^*. This equation corresponds to (2.6). All of the equations of the main text now generalize to the dynamic case by the addition of asterisks. The analytic structure is different in the dynamic case, with τ^*, etc., having more components than τ and div^* representing a different differential operator from div. The algebraic structure is all that is used, however, and this is unchanged.

[1]F. C. Karal, Jr. and J. B. Keller, J. Math. Phys. **5**, 537 (1964).
[2]M. J. Beran, *Statistical Continuum Theories* (Wiley–Interscience, New York, 1968).
[3]V. Twersky, J. Math. Phys. **18**, 2468 (1977).
[4]V. K. Varadan, V. V. Varadan, and Y. H. Pao, J. Acoust. Soc. Am. **63**, 1310 (1978).
[5]J. R. Willis, J. Mech. Phys. Solids **28**, 307 (1980).
[6]A. J. Devaney, J. Math. Phys. **21**, 2603 (1980).
[7]E. Kröner and H. Koch, Solid Mech. Arch. **1**, 183 (1976).
[8]J. R. Willis, in *Variational Methods in Mechanics of Solids*, edited by S. Nemat-Nasser (Pergamon, Oxford, 1980), p. 59.
[9]J. R. Willis, J. Mech. Phys. Solids **25**, 185 (1977).
[10]D. R. S. Talbot and J. R. Willis, Int. J. Solids Struct. **18**, 673 (1982).
[11]D. R. S. Talbot and J. R. Willis, Int. J. Solids Struct. **18**, 685 (1982).
[12]B. Budiansky, J. Mech. Phys. Solids **13**, 223 (1965).
[13]A. V. Hershey, J. Appl. Mech. **21**, 236 (1954).