New commutator identities on the Riemann tensor

C. Martin Pereira

Center for Theoretical Physics and Department of Physics & Astronomy, University of Maryland, College Park, Maryland 20742

(Received 8 August 1973)

A number of new generally covariant identities which involve second derivatives of the Riemann tensor are presented. Each of these new identities can be expressed by equating to zero either (a) a particular sum of terms each of which contains an operator of the form $(\nabla_\mu\nabla_\nu - \nabla_\nu\nabla_\mu)$ acting on the Riemann tensor; or (b) a particular sum of terms each of which contains an operator of the form ∇_α acting either on the expression $(\nabla_\mu R_{\alpha\beta\tau}{}^{\omega} + \nabla_\beta R_{\mu\alpha\tau}{}^{\omega} + \nabla_\alpha R_{\beta\mu\tau}{}^{\omega})$ or on the expression $(\nabla_\mu R_{\alpha\beta\tau\omega} + \nabla_\beta R_{\mu\alpha\tau\omega} + \nabla_\alpha R_{\beta\mu\tau\omega})$; or (c) a particular sum of algebraic terms each of which contains no derivatives of the Riemann tensor, but rather is quadratic in the Riemann tensor. Each of the new identities can be expressed in all three of the above-described forms. Furthermore, each of these new identitites can be thought of as an integrability condition derived from the equations that define the Riemann tensor in terms of the $\Gamma^{\omega}_{\alpha\beta}$ or the $g_{\mu\nu}$. The requirements of Riquier's existence theorem are used to guide the derivation of the identities. The operator ∇_μ denotes covariant differentiation. All the new identities assume the existence of a symmetric connection $\Gamma^{\mu}_{\alpha\beta}$ and one of the new identities assumes the existence of a metric. Schouten's identity and Walker's identity are also discussed.

1. INTRODUCTION

Einstein's nonvacuum field equations may be written in the form (1.1a) below

$$(-g)^{1/2}(R^{\mu\nu} - \tfrac{1}{2}g^{\mu\nu}R) = -8\pi(-g)^{1/2}T^{\mu\nu}, \qquad \mu = 1,\ldots,4. \quad (1.1a)$$

These equations imply the law of conservation of energy-momentum

$$[(-g)^{1/2}T^{\mu\nu}]_{,\nu} = 0, \quad \mu = 1,\ldots,4, \quad (1.1b)$$

and the algebraic equations

$$T^{[\mu\nu]} = 0, \quad \mu = 1,\ldots,4. \quad (1.1c)$$

Now Eq. (1.1b) is usually thought of as a restriction on the energy–momentum tensor $T^{\mu\nu}$. However, one can also think of Eq. (1.1b) as restricting only the metric, and it is possible to show[1] that one can give the quantities $(-g)^{1/2}T^{\mu\nu}$ arbitrarily as functions of the x^μ and still be certain of the existence of a solution to Eqs. (1.1a) and (1.1b) provided only that the given quantities satisfy the algebraic restriction (1.1c).

It can also be shown[1] that Eqs. (1.1b) and (1.1c) are integrability conditions for the system (1.1a) in the case where the quantities $(-g)^{1/2}T^{\mu\nu}$ are taken as the given functions.

Consider the system of equations which define the Riemann tensor

$$\partial_{[\nu}\Gamma_{\mu]\lambda}{}^{\alpha} + \Gamma_{\rho[\nu}{}^{\alpha}\Gamma_{\mu]\lambda}{}^{\rho} = \tfrac{1}{2}R_{\nu\mu\lambda}{}^{\alpha}, \mu = 1,\ldots,n, \quad n \geq 4. \quad (1.2a)$$

(Please refer to the Appendix for an explanation of the bracket notation used here.) In the present note, the attempt will be made to examine Eqs. (1.2) from the same standpoint as that just used to describe Eqs. (1.1a) above.

One tries then to think of the components of the Riemann tensor in Eq. (1.2a) as being given while the $\Gamma_{\lambda\alpha}{}^{\rho}$ are the unknown functions. It is well known[2] that this viewpoint leads to the Bianchi identities

$$\partial_{[\alpha}R_{\nu\mu]\lambda}{}^{\omega} - \Gamma_{\lambda[\alpha}{}^{\rho}R_{\nu\mu]\rho}{}^{\omega} + \Gamma_{\rho[\alpha}{}^{\omega}R_{\nu\mu]\lambda}{}^{\rho} = 0 \quad (1.2b)$$

as the integrability conditions of the system (1.2a). In Eq. (1.2b), the $R_{\nu\mu\rho}{}^{\omega}$ are given functions of x, and the symmetric $\Gamma_{\lambda\alpha}{}^{\rho}$ are unknowns. Equations (1.2b) are algebraic equations for the $\Gamma_{\lambda\alpha}{}^{\rho}$. Or in the terminology of Ref. 3, they contain only the zeroth-order derivatives of the $\Gamma_{\lambda\alpha}{}^{\rho}$.

Following the method described on p. 1544 of Ref. 1, combine Eq. (1.2b) with Eq. (1.2a) and consider the integrability conditions of the combined system, Eq. (1.2). Ones hope would be to prove that Eq. (1.2a) has (in the terminology of Ref. 1) integrability of the second kind. A familiarity with Riquier's procedure for deriving integrability conditions leads to the following conjectures about the nature of the integrability conditions of the system (1.2). First, the integrability conditions should be generally covariant. Second, they should involve first derivatives of Eq. (1.2b). [They should not involve first derivatives of Eq. (1.2a).] In addition, many of these integrability conditions should be quadratic identities on the Riemann tensor.

In order to clearly understand the procedure being u used here, note that when considering the integrability of the system (1.1), there are two different integrability problems that can be solved. One is the problem which takes Eq. (1.1b) as a restriction on the derivatives $T^{\mu 4}{}_{,4}$ as is usually done. The other is the problem which takes Eq. (1.1b) as a restriction on the derivatives $g_{\mu 4,4}$.

Similarly, when one is considering the integrability of the system (1.2), there are two different integrability problems that can be solved. One is the problem which takes Eq. (1.2b) as a restriction on certain of the derivatives $\partial_\alpha R_{\nu\mu\lambda}{}^{\omega}$. The other is the problem which takes Eq. (1.2b) as a restriction on certain of the $\Gamma_{\rho\alpha}{}^{\omega}$ and leaves the $\partial_\alpha R_{\nu\mu\lambda}{}^{\omega}$ to be chosen as freely as possible. The first integrability problem for the system (1.2) has been discussed in Reference 2, p. 146, and in Ref. 4. The second integrability problem for the system (1.2) is discussed (but not solved) in the present paper.

Thus, the proof of integrability of the system (1.2a) in the inhomogeneous case (case of nonzero curvature) has two steps. The second step has two branchs. The first branch is well-known and has been thoroughly investigated. The second branch, which has been largely overlooked, is discussed in the present paper.

Please note also that in the present discussion, when an equation is an integrability condition in the sense of Riquier, it will be called an integrability condition *whether it is identically satisfied or not.* Many well-known authors are in the habit of saying that an equation like Eq. (3.1) of Sec. 3 is not an integrability condition merely because it is identically satisfied. (See Ref. 5) This practice (which is contrary to Riquier's practice) leads to minor confusion in the treatment of linear systems of equations. However, in the case of nonlinear inhomogeneous systems, this practice leads to major confusion. Such confusion in regard to the system (1.2a) is, in the opinion of the present author, largely responsible for the fact that the discovery of the new integrability conditions presented in Sec. 3 was so long delayed.

2. QUADRATIC IDENTITIES

One quadratic identity which satisfies the preceding conjectures was found by Schouten.[6] This identity may be written

$$\nabla_{[\nu}\nabla_{\mu}R_{\alpha\beta]\tau}{}^{\omega} = 0. \tag{2.1}$$

Equation (2.1) can also be written[7]

$$\nabla_{\nu}\nabla_{[\mu}R_{\alpha\beta]\tau}{}^{\omega} - \nabla_{\mu}\nabla_{[\nu}R_{\alpha\beta]\tau}{}^{\omega} + \nabla_{\alpha}\nabla_{[\nu}R_{\mu\beta]\tau}{}^{\omega} - \nabla_{\beta}\nabla_{[\nu}R_{\mu\alpha]\tau}{}^{\omega} = 0. \tag{2.2}$$

It is clear by inspection that Eq. (2.2) is a consequence of Bianchi's identity (1.2b) as the conjectures of Sec. 1 require. Equation (2.1) can be converted to quadratic form in the following manner. First rewrite Eq. (2.1) in the commutator form

$$\nabla_{[\mu}\nabla_{\nu]}R_{\alpha\beta\tau}{}^{\sigma} + \nabla_{[\beta}\nabla_{\nu]}R_{\mu\alpha\tau}{}^{\sigma} + \nabla_{[\beta}\nabla_{\mu]}R_{\alpha\nu\tau}{}^{\sigma} + \nabla_{[\alpha}\nabla_{\nu]}R_{\beta\mu\tau}{}^{\sigma} + \nabla_{[\alpha}\nabla_{\mu]}R_{\nu\beta\tau}{}^{\sigma} + \nabla_{[\alpha}\nabla_{\beta]}R_{\mu\nu\tau}{}^{\sigma} = 0. \tag{2.3}$$

Then substitute Eq. (A17) of the Appendix into Eq. (2.3) and obtain after considerable simplification using the identity (A7).

$$R_{[\nu\mu\alpha]}{}^{\sigma}R_{\sigma\beta\tau}{}^{\omega} + R_{[\nu\mu\beta]}{}^{\sigma}R_{\alpha\sigma\tau}{}^{\omega} - R_{[\nu\beta\alpha]}{}^{\sigma}R_{\sigma\mu\tau}{}^{\omega} - R_{[\beta\mu\alpha]}{}^{\sigma}R_{\sigma\nu\tau}{}^{\omega} = 0. \tag{2.4}$$

Furthermore the connection $\Gamma_{\mu\nu}{}^{\alpha}$ is symmetric and therefore

$$R_{[\nu\mu\alpha]}{}^{\sigma} = 0. \tag{2.5}$$

(See Appendix.) Substitute Eq. (2.5) into Eq. (2.4) and obtain

$$0 = 0. \tag{2.6}$$

Thus Eq. (2.1) is identically satisfied by virtue of the linear identities (A7) and (A9). Thus the quadratic identity (2.4) gives no new algebraic restrictions on the Riemann tensor. Despite this fact, Eq. (2.1) is by no means trivial since it constitutes an important part of the integrability conditions for the system (1.2).

Note the general method used to obtain Eq. (2.4). First one derives an identity by taking an appropriate algebraic combination of the first covariant derivatives of Eq. (A20). Second, one tries to write the identity in commutator form. If the identity can be expressed in commutator form, then use Eq. (A17) to rewrite the identity in quadratic form. Finally, one checks the new quadratic identity to see whether or not it is identically satisfied by virtue of the previously known linear identities.

Examine Eq. (2.1). From its quadratic form, Eq. (2.4), it is clear that Eq. (2.1) is independent of the $\Gamma_{\alpha\beta}{}^{\mu}$. It involves only the given functions $R_{\mu\nu\alpha}{}^{\tau}(x)$. Thus, Eq. (2.1) has the following properties:

(a) It is obtained from Eqs. (1.2b) by differentiation.

(b) The highest derivative of the $\Gamma_{\alpha\beta}{}^{\mu}$ which it contains is not of higher order than the highest derivatives of the $\Gamma_{\alpha\beta}{}^{\mu}$ which occur in Eqs. (1.2b); i.e., the highest derivative which it contains is not higher than the zeroth order.

These properties are by themselves enough to imply that Eq. (2.1) is an integrability condition for the system (1.2). They do not prove however that Eq. (2.1) is the only integrability condition for the system (1.2). Such a proof would require a complete Riquier analysis of (1.2). In Sec. 3 other identities with these same general properties will be derived. The new identities of Sec. 3 will therefore also be integrability conditions.

Another quadratic identity has been found by Walker[8] in the case that the Riemann tensor is derivable from a metric. The commutator form of Walker's identity is

$$\nabla_{[\mu}\nabla_{\nu]}R_{\alpha\beta\sigma\tau} + \nabla_{[\sigma}\nabla_{\tau]}R_{\mu\nu\alpha\beta} + \nabla_{[\alpha}\nabla_{\beta]}R_{\sigma\tau\mu\nu} = 0. \tag{2.7}$$

When Eq. (A17) is used to put Eq. (2.7) in quadratic form, one finds that Eq. (2.7) is identically satisfied by virtue of the linear identities (A7), (A9), (A10), and (A11). It is not known at present whether Eq. (2.7) can be though of as an integrability condition or not.

3. NEW QUADRATIC INTEGRABILITY CONDITIONS

One quadratic integrability condition for the system (1.2), namely Eq. (2.1), has already been described. In the present section, a search for other integrability conditions will be made.

Motivated by the conjectures of Sec. 1, one attempts to find all quadratic algebraic identities on the Riemann tensor that involve exclusively the first covariant derivatives of the Bianchi identity. One also keeps in mind the fact that these integrability conditions obtained by differentiating Eq. (1.2b) must contain as their highest derivatives, derivatives of the form $\partial_{[\mu}\Gamma_{\nu]\alpha}{}^{\beta}$. This is necessary since this combination of derivatives is the only combination of the derivatives of the $\Gamma_{\nu\alpha}{}^{\beta}$ which (when evaluated at the pole of a geodesic coordinate system) is equivalent to the Riemann tensor. (It is assumed also that the integrability condition is generally covariant so that the usual theorems on the differential concomitants of a symmetric connection become relevant. See Ref. 2, p. 164.)

One proceeds as follows. Examine Eqs. (A20) and (A22) and note that the linear differential operators $\nabla_{[\mu..]}$ and $\nabla_{[\nu|..|..]}$ both give zero when applied to the Riemann tensor in the metric case. Furthermore, the commutator of these two operators can be shown to correspond to a sum of commutators of the form $\nabla_{[\mu}\nabla_{\nu]}$ and therefore leads to a quadratic identity on the Riemann tensor. The detailed procedure begins with a consideration of the equation

$$\epsilon^{\eta\nu\tau\omega}(\nabla_{\nu}\nabla_{[\mu}R_{\alpha\beta]\tau\omega}) = 0. \tag{3.1}$$

Equation (3.1) is satisfied as a result of Eq. (A20). Next, consider the equation

$$\epsilon^{\eta\nu\tau\omega}\nabla_{[\mu}\nabla_{|\nu|}R_{\alpha\beta]\tau\omega} = 0. \tag{3.2}$$

Equation (3.2) is satisfied as a result of Eq. (A22). Subtract Eq. (3.2) from Eq. (3.1) and obtain

$$\epsilon^{\eta\nu\tau\omega}(\nabla_\nu\nabla_{[\mu}R_{\alpha\beta]\tau\omega}) - \epsilon^{\eta\nu\tau\omega}\nabla_{[\mu}\nabla_{|\nu|}R_{\alpha\beta]\tau\omega} = 0. \quad (3.3)$$

Equation (3.3) may also be rewritten as a sum of commutators acting on the Riemann tensor. In this form, Eq. (3.3) becomes

$$\begin{aligned}&\nabla_{[\mu}\nabla_{\nu]}R_{\tau\omega\alpha\beta} + \nabla_{[\mu}\nabla_{\omega]}R_{\nu\tau\alpha\beta} + \nabla_{[\mu}\nabla_{\tau]}R_{\omega\nu\alpha\beta} + \nabla_{[\beta}\nabla_{\nu]}R_{\tau\omega\mu\alpha}\\ &\quad + \nabla_{[\beta}\nabla_{\omega]}R_{\nu\tau\mu\alpha} + \nabla_{[\beta}\nabla_{\tau]}R_{\omega\nu\mu\alpha}\\ &\quad + \nabla_{[\alpha}\nabla_{\nu]}R_{\tau\omega\beta\mu} + \nabla_{[\alpha}\nabla_{\omega]}R_{\nu\tau\beta\mu}\\ &\quad + \nabla_{[\alpha}\nabla_{\tau]}R_{\omega\nu\beta\mu} = 0. \quad (3.4)\end{aligned}$$

Equation (3.3) may also be written

$$\epsilon^{\eta\nu\tau\omega}\epsilon^{\rho\mu\alpha\beta}(\nabla_\nu\nabla_\mu R_{\alpha\beta\tau\omega} - \nabla_\mu\nabla_\nu R_{\alpha\beta\tau\omega}) = 0. \quad (3.5)$$

Relabel the appropriate dummy indices of Eq. (3.5) and obtain

$$\epsilon^{\eta\nu\tau\omega}\epsilon^{\rho\mu\alpha\beta}\nabla_\nu\nabla_\mu R_{\alpha\beta\tau\omega} - \epsilon^{\eta\mu\alpha\beta}\epsilon^{\rho\nu\tau\omega}\nabla_\nu\nabla_\mu R_{\tau\omega\alpha\beta} = 0. \quad (3.6)$$

Equation (3.6) may be written

$$\epsilon^{[\eta|\nu\tau\omega|}\epsilon^{\rho]\mu\alpha\beta}\nabla_\nu\nabla_\mu R_{\alpha\beta\tau\omega} = 0. \quad (3.7)$$

It is clear by inspection that in the four-dimensional case, Eq. (3.7) involves six independent equations. Substitute Eq. (A17) into Eq. (3.4) to obtain the quadratic identity

$$\epsilon^{\rho\mu\alpha\beta}(R_{\mu[\nu\tau}{}^{\sigma}R_{|\sigma|\omega]\beta\alpha} - R_{\mu[\nu|\beta|}{}^{\sigma}R_{\tau\omega]\alpha\sigma}) = 0. \quad (3.8)$$

Equation (3.8) is a new quadratic identity. Other new commutator identities will now be derived. The calculation begins with Bianchi's identity in the form

$$\nabla_\mu R_{\alpha\beta\tau}{}^{\sigma} + \nabla_\beta R_{\mu\alpha\tau}{}^{\sigma} + \nabla_\alpha R_{\beta\mu\tau}{}^{\sigma} = 0. \quad (3.9)$$

Operate on Eq. (3.9) to obtain

$$\nabla_{[\nu}(\nabla_{|\mu|}R_{\alpha\beta]\tau}{}^{\sigma} + \nabla_\beta R_{|\mu|\alpha]\tau}{}^{\sigma} + \nabla_\alpha R_{\beta]\mu\tau}{}^{\sigma}) = 0. \quad (3.10)$$

This can be written

$$\nabla_{[\nu}\nabla_{|\mu|}R_{\alpha\beta]\tau}{}^{\sigma} - 2\nabla_{[\nu}\nabla_\beta R_{\alpha]\mu\tau}{}^{\sigma} = 0. \quad (3.11)$$

The second term of Eq. (3.11) is already expressible as a sum of commutators acting on the Riemann tensor. To put the first term in commutator form, operate on Eq. (A20) and obtain

$$\nabla_\mu\nabla_{[\nu}R_{\alpha\beta]\tau}{}^{\sigma} = 0 \quad (3.12)$$

and then subtract Eq. (3.12) from Eq. (3.11) to obtain

$$\nabla_{[\nu}(\nabla_{|\mu|}R_{\alpha\beta]\tau}{}^{\sigma}) - \nabla_\mu\nabla_{[\nu}R_{\alpha\beta]\tau}{}^{\sigma} - 2\nabla_{[\nu}\nabla_\beta R_{\alpha]\mu\tau}{}^{\sigma} = 0. \quad (3.13)$$

Equation (3.13) is expressible as a sum of commutators. It can be written in the form.

$$\begin{aligned}&\nabla_{[\nu}\nabla_{\mu]}R_{\alpha\beta\tau}{}^{\sigma} + \nabla_{[\beta}\nabla_{\mu]}R_{\nu\alpha\tau}{}^{\sigma} + \nabla_{[\alpha}\nabla_{\mu]}R_{\beta\nu\tau}{}^{\sigma}\\ &\quad - \nabla_{[\nu}\nabla_{\beta]}R_{\alpha\mu\tau}{}^{\sigma} - \nabla_{[\alpha}\nabla_{\nu]}R_{\beta\mu\tau}{}^{\sigma}\\ &\quad - \nabla_{[\beta}\nabla_{\alpha]}R_{\nu\mu\tau}{}^{\sigma} = 0. \quad (3.14)\end{aligned}$$

Equation (3.14) can be shown to be equivalent to Schouten's identity. [Compare Eq. (2.3) with Eq. (3.14) to demonstrate this.] To derive a second new identity, operate on Eq. (3.9) and obtain

$$\nabla_{[\nu}(\nabla_{|\mu}R_{\alpha|\beta\tau]}{}^{\sigma} + \nabla_\beta R_{|\mu\alpha|\tau}{}^{\sigma} - \nabla_{|\alpha}R_{\mu|\beta\tau]}{}^{\sigma}) = 0. \quad (3.15)$$

Next consider the covariant derivative of Eq. (A21)

$$\nabla_\mu\nabla_{[\nu}R_{|\alpha|\beta\tau]}{}^{\sigma} = 0. \quad (3.16)$$

This may also be written after relabeling of indices:

$$-\nabla_\alpha\nabla_{[\nu}R_{|\mu|\beta\tau]}{}^{\sigma} = 0. \quad (3.17)$$

Subtract Eqs. (3.16) and (3.17) from Eq. (3.15) and obtain

$$\begin{aligned}&\nabla_{[\nu}\nabla_{|\mu}R_{\alpha|\beta\tau]}{}^{\sigma} - \nabla_\mu\nabla_{[\nu}R_{|\alpha|\beta\tau]}{}^{\sigma} + \nabla_{[\nu}\nabla_\beta R_{|\mu\alpha|\tau]}{}^{\sigma}\\ &\quad - \nabla_{[\nu}\nabla_{|\alpha}R_{\mu|\beta\tau]}{}^{\sigma} + \nabla_\alpha\nabla_{[\nu}R_{|\mu|\beta\tau]}{}^{\sigma} = 0. \quad (3.18)\end{aligned}$$

The commutator form of Eq. (3.18) is

$$\begin{aligned}&2\nabla_{[\nu}\nabla_{\mu]}R_{\alpha[\beta\tau]}{}^{\sigma} + 2\nabla_{[\mu}\nabla_{\beta]}R_{\alpha[\nu\tau]}{}^{\sigma}\\ &\quad + 2\nabla_{[\tau}\nabla_{\mu]}R_{\alpha[\nu\beta]}{}^{\sigma} + \nabla_{[\nu}\nabla_{\beta]}R_{\mu\alpha\tau}{}^{\sigma}\\ &\quad - \nabla_{[\nu}\nabla_{\tau]}R_{\mu\alpha\beta}{}^{\sigma} + \nabla_{[\beta}\nabla_{\tau]}R_{\mu\alpha\nu}{}^{\sigma}\\ &\quad - 2\nabla_{[\nu}\nabla_{\alpha]}R_{\mu[\beta\tau]}{}^{\sigma} + 2\nabla_{[\beta}\nabla_{\alpha]}R_{\mu[\nu\tau]}{}^{\sigma}\\ &\quad - 2\nabla_{[\tau}\nabla_{\alpha]}R_{\mu[\nu\beta]}{}^{\sigma} = 0. \quad (3.19)\end{aligned}$$

In each of the above new commutator identities one can substitute Eq. (A17) and obtain an identity in quadratic form. These quadratic forms of the identities will be presented elsewhere.

Finally it should be emphasized that the system (1.2) may conceivably have other integrability conditions not described here. A demonstration that the above identities are the only nonlinear identities will require a complete Riquier or Cartan existence proof for the system (1.2). It is also not claimed here that all the above identities are independent of one another either in their commutator form or in their quadratic form. In the four-dimensional metric case, it has been shown previously[1] that ten independent linear combinations of the Riemann tensor can be chosen arbitrarily. (These arbitrary components are closely related to the ten components of the Ricci tensor.) This fact places a limit on the number of independent new algebraic identities of the type discussed here.

It is also of interest to note that all the new identities discussed here become trivial in the three-dimensional metric case. See Ref. 9.

One of the most intriguing aspects of the identities is their nonlinear nature. Indeed, the entire calculation disappears if the linearized form of the system (1.2) is used. This suggests that these new identities carry information mainly about the nonlinear aspects of the structure of Eqs. (1.2). An investigation of the properties of these identities is currently in progress.

ACKNOWLEDGMENTS

I wish to express my grateful appreciation to the following investigators for the stimulating remarks or stimulating publications which led me over a ten-year period to the present result: G. E. Tauber and J. W. Weinberg, G. Kuerti, C. W. Misner and J. A. Wheeler, R. Penrose, P. G. Bergmann, J. N. Goldberg, J. R. Stachel, J. W. York.

APPENDIX

The bracket notation is used to denote antisymmetrization. For any quantity $A_{\mu\nu\alpha\sigma}$

$$A_{[\mu\nu]\alpha\sigma} = (1/2!)(A_{\mu\nu\alpha\sigma} - A_{\nu\mu\alpha\sigma}), \quad (A1)$$

$$A_{[\mu\nu\alpha]\sigma} = (1/3!)(A_{\mu\nu\alpha\sigma} + A_{\alpha\mu\nu\sigma} + A_{\nu\alpha\mu\sigma} - A_{\nu\mu\alpha\sigma} - A_{\mu\alpha\nu\sigma} - A_{\alpha\nu\mu\sigma}). \quad (A2)$$

Vertical bars may be used to set apart indices that are not antisymmetrized.

$$A_{[\mu|\nu|\alpha]\sigma} = (1/2!)(A_{\mu\nu\alpha\sigma} - A_{\alpha\nu\mu\sigma}). \quad (A3)$$

The quantity $A_{[\mu\nu\alpha\tau]}$ is antisymmetric in every pair of the indices μ, ν, α, τ. It is a sum of 24 terms with a positive sign in front of the terms having even permutations of the subscripts $\mu\nu\alpha\tau$ and a negative sign in front of the terms having odd permutations of $\mu\nu\alpha\tau$. The following identities are quite useful[10]:

$$A_{[\mu\nu\alpha\tau]} = \tfrac{1}{4}(A_{\mu[\nu\alpha\tau]} - A_{\nu[\mu\alpha\tau]} + A_{\alpha[\mu\nu\tau]} - A_{\tau[\mu\nu\alpha]}), \quad (A4)$$

$$A_{[\mu\nu\alpha]\tau} = \tfrac{1}{3}(A_{\mu[\nu\alpha]\tau} - A_{\nu[\mu\alpha]\tau} + A_{\alpha[\mu\nu]\tau}), \quad (A5a)$$

or

$$A_{[\mu\nu\alpha]\tau} = \tfrac{1}{3}(A_{\mu[\nu\alpha]\tau} + A_{\alpha[\mu\nu]\tau} + A_{\nu[\alpha\mu]\tau}). \quad (A5b)$$

Parentheses are used to denote symmetrization. For example,

$$C_{(\mu\nu)\alpha\tau} = (1/2!)(C_{\mu\nu\alpha\tau} + C_{\nu\mu\alpha\tau}). \quad (A6)$$

The Riemann tensor $R_{\mu\nu\alpha}{}^{\sigma}$ is antisymmetric in its first two indices

$$R_{(\mu\nu)\alpha}{}^{\sigma} = 0. \quad (A7)$$

In the case of a symmetric connection,

$$\Gamma_{[\mu\nu]}{}^{\alpha} = 0, \quad (A8)$$

the Riemann tensor satisfies

$$R_{[\mu\nu\tau]}{}^{\sigma} = 0. \quad (A9)$$

If the connection is not only symmetric but also expressible in terms of a metric, then one has the additional identities

$$R_{\mu\nu(\tau\sigma)} = 0, \quad (A10)$$

$$R_{\mu\nu\tau\sigma} = R_{\tau\sigma\mu\nu}. \quad (A11)$$

Note that in general, for any vector

$$2\nabla_{[\alpha}\nabla_{\beta]}v_{\tau} = -R_{\alpha\beta\tau}{}^{\sigma}v_{\sigma} \quad (A12)$$

and this implies that for any vector v_{τ}

$$2\nabla_{[\alpha}\nabla_{\beta}v_{\tau]} = -R_{[\alpha\beta\tau]}{}^{\sigma}v_{\sigma}. \quad (A13)$$

For a symmetric connection, one can substitute Eq. (A9) into Eq. (A13) and obtain for any vector v_{τ}

$$\nabla_{[\alpha}\nabla_{\beta}v_{\tau]} = 0. \quad (A14)$$

Note however that if $B_{\tau\omega\sigma\beta}$ is any tensor then even when (A9) holds one has

$$\nabla_{[\mu}\nabla_{\nu}B_{\tau]\omega\sigma\beta} \neq 0 \quad (A15)$$

and also

$$\nabla_{[\mu}\nabla_{\nu}B_{\tau\omega]\sigma\beta} \neq 0. \quad (A16)$$

Thus the identity (2.1) is by no means trivial.

For any tensor $B_{\tau\omega\sigma\beta}$ the commutator operator $\nabla_{[\mu}\nabla_{\nu]}$ acts as follows:

$$2\nabla_{[\mu}\nabla_{\nu]}B_{\tau\omega\alpha\beta} = -R_{\mu\nu\tau}{}^{\sigma}B_{\sigma\omega\alpha\beta} - R_{\mu\nu\omega}{}^{\sigma}B_{\tau\sigma\alpha\beta} - R_{\mu\nu\alpha}{}^{\sigma}B_{\tau\omega\sigma\beta} - R_{\mu\nu\beta}{}^{\sigma}B_{\tau\omega\alpha\sigma}. \quad (A17)$$

One also has for any tensor $B_{\tau\omega\alpha}{}^{\eta}$

$$2\nabla_{[\mu}\nabla_{\nu]}B_{\tau\omega\alpha}{}^{\eta} = -R_{\mu\nu\tau}{}^{\sigma}B_{\sigma\omega\alpha}{}^{\eta} - R_{\mu\nu\omega}{}^{\sigma}B_{\tau\sigma\alpha}{}^{\eta} - R_{\mu\nu\alpha}{}^{\sigma}B_{\tau\omega\sigma}{}^{\eta} + R_{\mu\nu\sigma}{}^{\eta}B_{\tau\omega\alpha}{}^{\sigma}. \quad (A18)$$

The Bianchi identities are usually written in the case of a symmetric connection

$$\nabla_{\mu}R_{\alpha\beta\tau}{}^{\sigma} + \nabla_{\beta}R_{\mu\alpha\tau}{}^{\sigma} + \nabla_{\alpha}R_{\beta\mu\tau}{}^{\sigma} = 0. \quad (A19)$$

However, since $R_{\alpha\beta\tau}{}^{\sigma}$ is antisymmetric in its first two indices, Eq. (A19) may also be written in the form

$$\nabla_{[\mu}R_{\alpha\beta]\tau}{}^{\sigma} = 0. \quad (A20)$$

A less familiar reformulation of Eq. (A19) is

$$\nabla_{[\mu}R_{|\alpha|\beta\tau]}{}^{\sigma} = 0. \quad (A21)$$

If the symmetric connection is derivable from a metric, one also has

$$\nabla_{[\mu}R_{|\alpha\beta|\tau\sigma]} = 0. \quad (A22)$$

[1] C. M. Pereira, J. Math. Phys. **13**, 1542 (1972). For related discussions, see C. M. Pereira, J. Math. Phys. **14**, 1498 (1973); C. M. Pereira, "Two Therorems on Energy-Momentum Conservation"; J. W. York, J. Math. Phys. **13**, 125 (1972).

[2] J. A. Schouten, *Ricci-Calculus* (Springer-Verlag, Berlin, 1954), 2nd ed., Chap. III, Eq. (5.18).

[3] J. F. Ritt, *Differential Algebra* (Dover, New York, 1966), Chap VIII [or (AMS Colloq. Publ., New York, 1950), Vol. 33, Chap. VIII].

[4] T. Y. Thomas, *The Differential Invariants of Generalized Spaces* (Cambridge U. P., Cambridge, 1934), Chap. 10.

[5] For example in Ref. 2, p. 146, Schouten describes the derivation of Eq. (3.1) and then proceeds to imply that since Eq. (3.1) is identically satisfied, it is not an integrability condition.

[6] See Ref. 2, p. 146.

[7] See Appendix, Eq. (A4).

[8] A. G. Walker, Proc. Lond. Math. Soc. **52**, 36 (1950).

[9] C. M. Pereira, "The Curvature and the Metric in Three Dimensions," (to be published).

[10] For more general algebraic identities,see Ref. 2, p. 16.

Lowest nonnegative solution of the Yukawa Hamiltonian*

Michael Martin Nieto

Los Alamos Scientific Laboratory, University of California, Los Alamos, New Mexico 87544
(Received 21 September 1973)

Two recent papers [J. Math. Phys. **13**, 1825 (1972); **14**, 1140 (1973)] obtained rough limits $1.68 \geq b_0 \geq 1.67$ for the quantity b_0 such that an energy eigenvalue of the Yukawa Hamiltonian $[-\hbar^2\nabla^2/2m - U_0 \exp(-\mu r c/\hbar)/r]$ must be zero or positive if $b \equiv (m/\mu)(2U_0/c\hbar) \leq b_0$. We point out that (along with other aspects of the Yukawa equation) the quantity b_0 has long been known to a much higher precision.

A few years ago Dyson and Lenard[1] had occasion to touch on the Yukawa potential Schrödinger equation

$$\left[-\frac{\hbar^2}{2m}\nabla^2 - \frac{U_0(e^{-\mu r c/\hbar})}{r}\right]\psi = E\psi, \tag{1}$$

which, by the change of variables

$$x = \mu r c/\hbar, \quad b = (m/\mu)(2U_0/c\hbar),$$

$$a = E(2m/\mu^2 c^2), \quad \phi(x) = x\psi(x), \tag{2}$$

can be written for the spherically symmetric case ($l = 0$) in the form

$$\left[\frac{d^2}{dx^2} + a + \frac{be^{-x}}{x}\right]\phi = 0, \quad \phi(0) = \phi(\infty) = 0. \tag{3}$$

Dyson and Lenard were discussing the stability of a many body system of charged particles, and it was useful to them to demonstrate quickly in their Lemma I that if $b \leq \sqrt{2}$, then the Yukawa Hamiltonian is nonnegative (i.e., all energy eigenvalues (E or a) ≥ 0. In other words, with a given m and U_0 there is a maximum Yukawa particle mass μ_0 (or a minimum Yukawa parameter b_0) such that for all b satisfying

$$b \leq b_0 \quad \text{for some } b_0 \geq \sqrt{2}, \tag{4}$$

one has

$$E \propto a \geq 0. \tag{5}$$

Quite recently, two papers have appeared[2,3] which were inspired by Ref. 1 and whose purpose was to show by approximation techniques that

$$1.68 \geq b_0 \geq 1.64 \quad \text{and} \quad b_0 \geq 1.67, \tag{6}$$

respectively.

In this note we wish to point out that the Yukawa equations (1) and (3) and many details of its solution have been studied extensively in the past. In particular, evaluations of b_0 as good as and orders of magnitude more precise than Eq. (6) have existed in the literature for many years.

To place the problem in perspective, recall that in 1935 Yukawa[4] proposed that the nucleon–nucleon force is intermediated by a particle of mass μ (what we now call the meson) which satisfies the wave equation

$$(\Box + \mu^2)\chi = 0. \tag{7}$$

Equation (7) implies the static potential of Eq. (1), and people[5,6] began to investigate Eq. (1) in hopes of solving the nucleon–nucleon interaction.

Negative, zero, and positive energies were studied since there was interest in the deutron bound state problem and the nucleon–nucleon scattering length and general scattering problems, respectively. During this period evaluations of b_0 of the caliber of Eq. (6) were discussed by many authors.[7-10]

In 1942 Hulthén[11] published a more detailed study of the Yukawa problem ($b_0 \cong 1.680$), this work being greatly extended in 1951 by Hulthén and Laurikainen.[12] It was observed that a solution to the Yukawa equation (3) can be written as

$$\phi(x) = \exp[-(-a)^{1/2}x]\sum_{n=1}^{\infty}\omega_n(1 - e^{-x})^n. \tag{8}$$

Arbitrarily precise solutions can be obtained by taking various numbers of terms in the Eq. (8) expansion and using standard approximation techniques.

Hulthén and Laurikainen[12] used Eq. (8) to solve Eq. (3) as an eigenvalue problem for b (or μ), given many individual energies (including zero total energy). For a given energy (quantity a) they found the three lowest b eigenvalues and the wavefunctions to high precision. They also obtained asymptotic expansion formulas. In particular, for $E = 0$, Hulthén and Laurikainen's iterations ($n = 1, 2, 3$) for the lowest eigenvalue b (i.e., b_0) were

$$b_0(1) = 1.679\,933, \quad b_0(2) = 1.679\,853, \quad b_0(3) = 1.679\,8195. \tag{9}$$

This exhibits the precision we described.

It is worthwhile to note that Hulthén[11] also devised the potential function

$$F = e^{-x}/(1 - e^{-x}). \tag{10}$$

This is useful since it admits of an exact solution and goes over to the Yukawa potential for small x. One can consult the literature for properties of this potential.[11-14] Also, Ref. 15 gives a review of the subject of Yukawa and Hulthén potential theory.

Finally, we mention that other aspects of the Yukawa potential have been studied. Examples are its use in massive-photon electrodynamics[16] and its connection to Cauchy–Riemann and Hilbert transform theory in the two-dimensional case.[17-18]

ACKNOWLEDGMENTS

For discussions on this problem I wish to thank Richard A. Krajcik, James D. Louck, and especially Mark Bolsterli, who directed me to Ref. 15.

* Work performed under the auspices of the U.S. Atomic Energy Commission.

[1]F. J. Dyson and A. Lenard, J. Math. Phys. 8, 423 (1967).
[2]J. Piepenbrink, J. Math. Phys. 13, 1825 (1972).
[3]P. C. Hemmer, J. Math. Phys. 14, 1140 (1973).
[4]H. Yukawa, Proc. Phys. Math. Soc. Japan 17, 48 (1935).
[5]E. C. G. Stueckelberg, Phys. Rev. 52, 41 (1937).
[6]N. Kemmer, Nature 141, 116 (1938); H. J. Bhabha, Nature 141, 117 (1938). A personal historical account of the impact of Yukawa's theory on European physicists is given in N. Kemmer, Suppl. Prog. Theor. Phys. (Kyoto), extra number, 602 (1965).
[7]A. H. Wilson, Proc. Cambridge Phil. Soc. 34, 365 (1938).
[8]R. G. Sachs and M. Goeppert-Meyer, Phys. Rev. 53, 991 (1938).
[9]F. J. Belinfante, *Theory of Heavy Quanta*, University of Leiden thesis, (Martinus Nijhoff, S'-Gravenhage, 1939). See p. 45.
[10]C. Møller and L. Rosenfeld, D. Kgl. Danske Vidensk. Selskab, Mat.-Fys. Medd. 25, No. 8 (1940). See pp. 57-58.
[11]L. Hulthén, Arkiv Mat., Astron. Fys. 28A, No. 5 (1942).
[12]L. Hulthén and K. V. Laurikainen, Rev. Mod. Phys. 23, 1 (1951).
[13]R. Jost, Helv. Phys. Acta 20, 256 (1947).
[14]S. T. Ma, Austral, J. Phys. 7, 365 (1954).
[15]L. Hulthén and M. Sugawara, in *Handbuck der Physik, Vol. 39*, edited by S. Flügge (Springer-Verlag, Berlin, 1957). See Secs. 12–16.
[16]A. S. Goldhaber and M. M. Nieto, Rev. Mod. Phys. 43, 277 (1971).
[17]R. J. Duffin, J. Math. Anal. Appl. 35, 105 (1971).
[18]R. J. Duffin, Proc. Natl. Acad. Sci. USA 69, 3677 (1972).

On ergodic limits of normal states in quantum statistical mechanics

E. Prugovečki*

Department of Mathematics, University of Toronto, Toronto 181, Canada

A. Tip†

FOM-Instituut voor Atoom-en Molecuulfysica, Amsterdam (Wgm), The Netherlands
(Received 7 February 1973; revised manuscript received 5 June 1973)

The asymptotic behavior for $t \to \pm \infty$ of $S(t) = \exp(-iHt)S\exp(+iHt)$ and its time average $\bar{S}(t) = t^{-1}\int_0^t du\ S(u)$ is discussed. Here is S an element of the Banach space $\mathfrak{B}_1$, constituted by the trace class of operators on the (separable or nonseparable) Hilbert space $\mathfrak{H}$, and H is the Hamiltonian, i.e., a bounded or unbounded self-adjoint operator on $\mathfrak{H}$. Necessary and sufficient conditions are given for the existence of the limits $\bar{S}(\pm\infty)$ and $S(\pm\infty)$ with respect to the weak topology on $\mathfrak{B}_1$, for the latter under the assumption that the continuous spectrum of H is absolutely continuous. In addition it is shown that if, for a normal state (density operator) ρ, $\bar{\rho}(t)$ has a weak limit, then the limit is again a normal state. This provides further insight in the nature of Emch's "first ergodic paradox" [G. G. Emch, J. Math. Phys. 7, 1413 (1966)].

I. INTRODUCTION

In the present paper we are concerned with the asymptotic temporal behavior of density operators (normal states) $\rho(t) = \exp[-iHt]\rho(0)\exp[+iHt]$, H being the Hamiltonian of the system under consideration.

In order to set the stage for the further discussion let us introduce some notation. Let $\mathfrak{H}$ be a (separable or nonseparable) Hilbert space with elements $x, y, \ldots$ and inner product (x,y) linear in x and antilinear in y. We denote by $\mathfrak{B} = \mathfrak{B}(\mathfrak{H})$ the C^*- algebra of all bounded operators on $\mathfrak{H}$ and by $|A|$ the operator bound of $A \in \mathfrak{B}$. $\mathfrak{B}_\infty = \mathfrak{B}_\infty(\mathfrak{H})$ is the minimal norm ideal of all completely continuous operators on $\mathfrak{H}$, whereas $\mathfrak{B}_1 = \mathfrak{B}_1(\mathfrak{H})$ and $\mathfrak{B}_2 = \mathfrak{B}_2(\mathfrak{H})$ denote the trace class and the Schmidt class of operators, respectively. The trace norm of $S \in \mathfrak{B}_1$ is denoted by $|S|_1$, whereas for $S, T \in \mathfrak{B}_2$ we have the inner product $(S,T)_2 = \mathrm{tr}\{ST^*\}$ and norm $|S|_2 = (S,S)_2^{1/2}$. $\mathfrak{B}_1, \mathfrak{B}_2$, and $\mathfrak{B}_\infty$ are familiar examples of minimal norm ideals I_α of completely continuous operators on $\mathfrak{H}$.[1] As is well known any such minimal norm ideal I_α, equipped with the α-norm topology is a Banach space and the set $\mathfrak{F}$ of all operators of finite rank is α-norm dense in I_α. Another important result is that the dual $\mathfrak{B}_\infty^*$ of $\mathfrak{B}_\infty$ is isometrically isomorphic to the trace class $\mathfrak{B}_1$, the dual of which, in its turn, is isometrically isomorphic to $\mathfrak{B}$. For this reason the notation $\mathfrak{B}_\infty = \mathfrak{B}_{**}$ and $\mathfrak{B}_1 = \mathfrak{B}_*$ is sometimes used.

A bounded linear functional (blf) on $\mathfrak{B}_1$ has the representation $\phi(S) = \mathrm{tr}\{SA\}$, $S \in \mathfrak{B}_1$ with a fixed $A \in \mathfrak{B}$ whereas a blf on $\mathfrak{B}$ can be written in a unique way as $\psi(A) = \mathrm{tr}\{SA\} + \psi^\perp(A)$, where $A \in \mathfrak{B}$, S is a fixed element of $\mathfrak{B}_1$ and $\psi^\perp(A)$ vanishes identically for each $A \in \mathfrak{B}_\infty$. We shall say that a blf ψ on $\mathfrak{B}$ is normal in case $\psi^\perp \equiv 0$. The normal *states* are the density operators, namely they are those elements ρ of $\mathfrak{B}_1$ which are positive (i.e, $\rho \geqslant 0$) and have unit trace (i.e, $\mathrm{tr}\rho = 1$).

The Hamiltonian H is a self-adjoint operator on $\mathfrak{H}$, i.e., $H = H^*$, so that, according to Stone's theorem, $U(t) = \exp(-iHt)$, $t \in \mathfrak{R}$ is a one-parameter, strongly continuous group of unitary operators on $\mathfrak{H}$. The equation of motion for the density operator is then given by $\rho(t) = U(t)\rho U^*(t) = \mathbf{U}(t)\rho$, ρ being the density operator of the system at $t = 0$. In the sequel we investigate the asymptotic behaviour for large $|t|$ of $S(t)$ and $\bar{S}(t) = t^{-1}\int_0^t ds\ S(s)$ [a precise definition of $S(t)$ is given in Sec. 3] for general $S \in \mathfrak{B}_1$ and with respect to various topologies. The emphasis, however, will be on the weak topology, i.e., on the asymptotic behaviour of $\mathrm{tr}\{S(t)A\}$ and $\mathrm{tr}\{\bar{S}(t)A\}$ with $A \in \mathfrak{B}$.

In recent years work along similar lines has been done by Emch[2] and by Moyal.[3] In fact, Emch has pointed out that nonnormal states may result as ergodic limits of normal ones. We show that this cannot be the case if $\mathrm{tr}\bar{S}(t)A$ has a limit for every $A \in \mathfrak{B}$. More precisely, if S is a normal blf on $\mathfrak{B}$ and $\bar{S}(t)$ has a weak limit [i.e., $S(t)$ converges weakly towards some blf $\psi \in \mathfrak{B}^*$] for t tending to $+\infty$ or $-\infty$, respectively (in short $t \to \pm\infty$), then ψ is again normal. A basic distinction between our work and Ref. 2 is that we consider a general self-adjoint Hamiltonian H, whereas Emch considers only bounded Hamiltonians. This makes it necessary to modify a number of definitions and proofs adopted from the case when H is bounded. For instance, the infinitesimal generator $-i\mathbf{H}$ of $\mathbf{U}(t)$, defined above, does no longer have the simple representation as the inner derivation $-i\mathbf{H} = -i[H, \cdot]$. The case of a general self-adjoint Hamiltonian H has been considered by Moyal,[3] but unfortunately there are some errors in his work. We correct these errors in Sec. II where we derive some properties of the point spectrum of the generator $\mathbf{H}$.

In Sec. III we prove the normality of weak ergodic limits, referred to above, whereas in Sec. IV we determine those $S \in \mathfrak{B}_1$ with $\mathrm{tr}S \neq 0$, for which $S(t)$ possesses an ergodic limit with respect to the norm topology on $\mathfrak{B}_1$. Denoting by E_p the projector upon the subspace of $\mathfrak{H}$ spanned by the eigenvectors of H and by $E_c = I - E_p$ its complement in $\mathfrak{H}$, we find that a sufficient condition for the existence of $\lim_{t\to\pm\infty}\bar{S}(t)$, with respect to either the weak or strong topology on $\mathfrak{B}_1$, is that $F_cSE_c \equiv 0$. For S with $S \geqslant 0$ or $S \leqslant 0$ this condition is both necessary and sufficient.

In case H has a pure point spectrum it is well known that $S(t)$ does not have a weak or strong limit for $t \to \pm\infty$, unless S is invariant, i.e., $S(t) = S$ for each $t \in \mathfrak{R}$. However, if the continuous spectrum of H is absolutely continuous, then the weak limit of $E_pS(t)E_c$ and $E_cS(t)E_p$ is zero. In that case sufficient conditions for the existence of the weak limit of $S(t)$ for $t \to \pm\infty$ are $E_cSE_c \equiv 0$ and E_pSE_p invariant. For S with $S \leqslant 0$ or $S \geqslant 0$ these conditions are both necessary and sufficient.

II. PROPERTIES OF THE POINT SPECTRUM OF H

In order to discuss some of the spectral properties of the infinitesimal generator $-i\mathbf{H}$ of $\mathbf{U}(t)$, it is convenient to make use of the concept of transformer.[4-6] Transformers are (linear) transformations which map operators into operators (the name superoperator is also often used, whereas within the present context H is often referred to as the Liouville operator).

Let us consider the three families of mappings of $S \in I_\alpha$ into

$$\mathbf{U}_l(t)S = U(t)S, \mathbf{U}_r(t)S = SU(t), \mathbf{U}(t)S = U(t)SU(-t), \quad (1)$$

respectively, as groups of transformers on a given minimal norm ideal I_α of $\mathfrak{B}(\mathfrak{H})$. The general theory of semigroups of transformers of the above type has been considered recently by the present authors (cf. Ref. 6). In the sequel we will only mention a few results pertaining to the problem at hand. Thus it can be shown that the three transformer groups defined in (1) are continuous in the norm (i.e., strong) topology of the Banach space I_α. Denoting their infinitesimal generators by $-i\mathbf{H}_l, -i\mathbf{H}_r$ and $-i\mathbf{H}$, respectively, we have $\mathbf{H}_l S = (HS)^{**}$ for $S \in \mathfrak{D}_{\mathbf{H}_l}$ and $\mathbf{H}_r S = (SH)^{**}$ for $S \in \mathfrak{D}_{\mathbf{H}_r}$ (here $\mathfrak{D}_A$ denotes the domain of A).

In addition, for $S \in \mathfrak{D}_{\mathbf{H}_l} \cap \mathfrak{D}_{\mathbf{H}_r} \subset \mathfrak{D}_{\mathbf{H}}$

$$\mathbf{H}S = (HS - SH)^{**}.$$

Although $(HS)^{**} = HS$, since H, being self-adjoint, is closed and densely defined, whereas S is bounded, such a reduction is not possible for $(SH)^{**}$, unless H is bounded. Thus an expression for H given by Moyal [Ref. 3, Eq. (5); Moyal uses D instead of H] is, in general, correct only for bounded H.

Our notation for the canonical representation of a completely continuous operator S will be $S = \sum_k \lambda_k (\cdot, u_k) v_k$ where $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ and $\{u_k\}$ and $\{v_k\}$ are two countable orthonormal systems in $\mathfrak{H}$. $\{(\cdot, x)y$ is defined according to $[(\cdot, x)y]z = (z, x)y, z \in \mathfrak{H}.\}$ Since each λ_k occurs only finitely many times we can also write

$$S = \sum_k \lambda_k \sum_{j_k=1}^{n_k} (\cdot, u_{j_k}) v_{j_k}, \text{where now } \lambda_1 > \lambda_2 > \cdots > 0$$

and n_k is finite. Using the latter notation we have $SS^* = \sum_k \omega_k P_k$ and $S^*S = \sum_k \omega_k Q_k$, where $\omega_k = \lambda_k^2$ while

$$\left\{P_k = \sum_{j_k=1}^{n_k} (\cdot, v_{j_k}) v_{j_k}\right\} \text{ and } \left\{Q_k = \sum_{j_k=1}^{n_k} (\cdot, u_{j_k}) u_{j_k}\right\}$$

are two families of mutually orthogonal projectors, the ranks of P_k and Q_k being given by the finite number n_k. In addition we have

$$\sum_{k=1}^{\infty} P_k = I - P_0 \text{ and } \sum_{k=1}^{\infty} Q_k = I - Q_0,$$

where P_0 and Q_0 are the projectors upon the null-spaces of SS^* and S^*S, respectively. In case $\mathfrak{H}$ is non-separable the latter null-spaces are necessarily infinite dimensional, nonseparable subspaces of $\mathfrak{H}$.

In the sequel use will be made of the following theorem:

Theorem 1: Suppose $S \in \mathfrak{D}_{\mathbf{H}} \subset I_\alpha$ is an eigenvector of the transformer H acting in a minimal norm ideal I_α of completely continuous operators, i.e., $\mathbf{H}S = \mu S, S \neq 0$. Then μ equals the difference of two eigenvectors of H and S has the form $S = \sum_k \lambda_k (\cdot, y_k) x_k$, where $\{x_k\}$ and $\{y_k\}$ are two orthonormal systems of eigenvectors of H, such that the difference of the eigenvalues, corresponding to x_k and y_k in each term $(\cdot, y_k)x_k$ equals μ. Conversely, if S has this structure, then it is an eigenvector of H.

Remark: The above theorem has been stated by Moyal[3] for the special case when I_α is either $\mathfrak{B}_1$ or $\mathfrak{B}_2$. Moyal's proof, however, contains a few errors. As mentioned before his expression for H is incorrect. Also it is not *a priori* evident that $S = \sum_k \lambda_k (\cdot, u_k) v_k \in D_{\mathbf{H}}$ implies that $u_k, v_k \in \mathfrak{D}_{\mathbf{H}}$ for $k = 1, 2, 3, \ldots$.

Proof: We write S in the form

$$S = \sum_k \lambda_k \sum_{j_k=1}^{n_k} (\cdot, u_{j_k}) v_{j_k}$$

so that (see above)

$$SS^* = \sum_k \omega_k P_k \text{ and } S^*S = \sum_k \omega_k Q_k.$$

Since $\mathbf{H}S = \mu S$ we have for $S(t) = \mathbf{U}(t)S$;

$$\partial_t S(t) = \partial_t \mathbf{U}(t)S = -i\mathbf{H}\mathbf{U}(t)S$$
$$= -i\mathbf{U}(t)\mathbf{H}S = -i\mu\mathbf{U}(t)S = -i\mu S(t).$$

and hence (see Ref. 7, pp. 68, 69) $S(t) = \exp[i\mu t]S$, i.e.,

$\mathbf{U}(t)S = \exp[-i\mu t]S$. It follows that

$\mathbf{U}(t)S^* = (\mathbf{U}(t)S)^* = \exp[+i\mu t]S^*$ and $\mathbf{U}(t)(SS^*) = SS^*$, whereas

$\mathbf{U}(t)(S^*S) = S^*S$. Thus restricting our attention to SS^*;

$$\exp[-iHt] \sum_k \omega_k P_k \exp[+iHt] = \sum_k \omega_k P_k.$$

Hence, multiplying this relation from the left with P_l and from the right with $\exp[-iHt]P_m$, admitting the values $l = 0$ and $m = 0, \omega_0$ being zero, we obtain;

$$(\omega_l - \omega_m) P_l \exp[-iHt] P_m = 0.$$

Since $\omega_l \neq \omega_m$ unless $l = m$ and

$$\sum_{k=0}^{\infty} P_k = I,$$

it follows that

$$P_k, k = 0, 1, 2, 3, \ldots, \text{reduces } \exp[-iHt].$$

Let $x \in \mathfrak{D}_H$. Then, since $(-it)^{-1}\{\exp[-iHt] - 1\}P_k x = P_k(-it)^{-1}\{\exp[-iHt] - 1\}x$ and the right-hand side converges strongly towards $P_k Hx$, it follows that $P_k x \in \mathfrak{D}_H$ and $P_k Hx = HP_k x$. For $k = 1, 2, \ldots$, (only non-zero k's are considered in the remaining part of the proof) P_k is a projector of finite rank n_k and, as H is closed and densely defined, it follows that HP_k is closed and everywhere defined, and consequently must be bounded. In particular $P_k H P_k$ is a bounded self-adjoint operator of finite rank n_k. In addition $\exp[-iHt]P_k = \exp[-iP_k H P_k t]P_k$. In the same way we obtain the corresponding results for Q_k. Next we turn to the equation

$$\mathbf{U}(t)S = \exp[-iHt]S \exp[+iHt] = \exp[-i\mu t]S.$$

Multiplication from the left with P_k and from the right with Q_k leads to

$$\exp[-iP_k H P_k t]P_k S Q_k \exp[+iQ_k H Q_k t] = \exp[-i\mu t]P_k S Q_k,$$

where, obviously, $P_k S Q_k = \lambda_k \sum_{j_k=1}^{n_k} (\cdot, u_{j_k}) v_{j_k}$.

Let $\{x_{l_k}\}_{l_k=1}^{n_k}$ and $\{y_{m_k}\}_{m_k=1}^{n_k}$ denote the n_k orthonormal eigenvectors of P_kHP_k in $P_k\mathcal{K}$ and of Q_kHQ_k in $Q_k\mathcal{K}$, respectively (they may be nonunique, but what matters here is that they span $P_k\mathcal{K}$ and $Q_k\mathcal{K}$, respectively). Then, dropping the subindex k for the moment,

$$\sum_j (\cdot, u_j)v_j = \sum_j \sum_l \sum_m c_{lm}(\cdot, y_l)x_m,$$

with

$$c_{lm} = \sum_j (v_j, x_m)(y_l, u_j),$$

and we have, denoting the eigenvalue of H, corresponding to x_m by ϵ_m and the one corresponding to y_l by ϵ_l:

$$\sum_{l,m} \exp[-i(\epsilon_m - \epsilon_l)t]c_{lm}(\cdot, y_l)x_m = \exp[-i\mu t]\sum_{l,m} c_{lm}(\cdot, y_l)x_m.$$

Hence $c_{lm} = 0$ unless $\epsilon_m - \epsilon_l = \mu$. Since $S \neq 0$ there is at least one c_{lm} in the expansion of some P_kSQ_k, $k = 1, 2, 3, \ldots$ which is nonvanishing and hence μ equals the difference of two eigenvalues of H. Consider now

$$\sum_l c_{lm}(\cdot, y_l)x_m = (\cdot, \sum_l \bar{c}_{lm}y_l)x_m$$

for some fixed m. Since any y_l must be an eigenvector of H at the eigenvalue $\epsilon_m - \mu$, this must be true for

$$w_m = \sum_l \bar{c}_{lm}y_l = \sum_j (x_m, v_j)u_j$$

[note that $(w_m, w_n) = \delta_{mn}$]. Thus we have

$$P_kSQ_k = \lambda_k \sum_{m_k=1}^{n_k} (\cdot, w_{m_k})x_{m_k},$$

where x_{m_k} and w_{m_k} are eigenvectors of H, the difference of the corresponding eigenvalues being μ. Since $\{w_{m_k}\}$ spans $Q_k\mathcal{K}$ and $\{x_{m_k}\}$ spans $P_k\mathcal{K}$ and $\{P_k\}$ and $\{Q_k\}$ are two sets of mutually orthogonal projectors, the expression for S, as announced in the theorem, follows.

The converse statement is obvious.

It is already implicit in theorem 1 that in case H has empty point spectrum, there is no nonvanishing $S \in I_\alpha$ such that $HS = 0$. Defining the commutant $\{H\}'$ of H as the set of elements of $\mathfrak{B}(\mathcal{K})$ which commute with all projections E_κ of the resolution of the identity of H, we in fact have the following result (ϕ denotes the empty set);

Lemma 1: Suppose H has empty point spectrum. Then $\mathfrak{B}_\infty \cap \{H\}' = \phi$.

Proof: Assume to the contrary that some nonzero $S \in \mathfrak{B}_\infty$ commutes with all E_κ from the resolution of the identity of H; i.e., $SE_\kappa = E_\kappa S$. Then it follows that $S^*E_\kappa = E_\kappa S^*$ and hence $SS^*E_\kappa = E_\kappa SS^*$.

Using the notation introduced before we have $SS^* = \sum_k \omega_k P_k, \omega_1 > \omega_2 > \cdots > 0$, where the P_k's are mutually orthogonal, and we obtain from the relation $P_kSS^*E_\kappa P_l = P_kE_\kappa SS^*P_l$ the equation $(\omega_k - \omega_l)P_kE_\kappa P_l = 0$, $k, l = 1, 2, \cdots$. As $P_0SS^* = SS^*P_0 = 0$ this relation remains true if k or l assumes the value zero ($\omega_0 = 0$).

Since

$$\sum_{k=0}^{\infty} P_k = I,$$

it follows that each P_k reduces each E_κ from the resolution of the identity of H. Hence P_k reduces H, i.e., $P_kH = HP_k = P_kHP_k$. Since for $k > 0$ each P_k has finite rank and since the index k assumes at least one value, it follows that H has at least one eigenvector and we have a contradiction.

We say that $S \in I_\alpha$ is *invariant* in case $\mathbf{U}(t)S = S$ for all $t \in \mathfrak{R}$. Thus obviously, an invariant $S \in I_\alpha$ is an eigenvector of H with the eigenvalue zero. If H has empty point spectrum we conclude from Theorem 1 that there exists no invariant $S \in I_\alpha$.

III. THE NATURE OF ERGODIC LIMITS ON THE TRACE CLASS $\mathfrak{B}_1(\mathcal{K})$

In this section we investigate the limits of $S(t)$ and $\overline{S}(t)$ in the weak* and weak topology on $\mathfrak{B}$, respectively, as t tends to $\pm\infty$. We recall that the weak* topology on $\mathfrak{B}_1$ is generated by the bounded linear functionals of the type $\phi(S) = \mathrm{tr}SA$, where $S \in \mathfrak{B}_1$, and A belongs to the predual $\mathfrak{B}_\infty = \mathfrak{B}_{**}$ of $\mathfrak{B}_1$. A neighborhood base at $S_0 \in \mathfrak{B}_1$ for this topology is given by all sets of the form

$$\{S \in \mathfrak{B}_1 \,|\, |\mathrm{tr}(S - S_0)A_j| < \epsilon, j = 1, \ldots k\},$$

where $A_1, \ldots, A_k \in \mathfrak{B}_\infty$. Similarly the weak topology on $\mathfrak{B}_1$ is generated by the set of all blf on $\mathfrak{B}_1$ and a neighborhood base at S_0 in this topology is given by sets of the above form except that now $A_1, \ldots, A_k \in \mathfrak{B}$. Thus we shall be concerned with the asymptotic behavior for large $|t|$ of the expression $\phi(t, S, A) = \mathrm{tr}[\mathbf{U}(t)A = \mathrm{tr}\{S(t)A\}$ in which S is contained in $\mathfrak{B}_1$ and A belongs to either $\mathfrak{B}_\infty$ or to $\mathfrak{B}$.

Suppose that for t tending to $\pm\infty$ the functional $\phi(t, S, A)$ has a limit for certain S and A. Then, since $\phi(t, S, A)$ is continuous in t,

$$\overline{\phi}(t, S, A) = t^{-1}\int_0^t ds\ \phi(s, S, A)$$

exists and tends to the same limit for $t \to \pm\infty$ as $\phi(t, S, A)$. However, since H generates a strongly continuous (semi-) group on $\mathfrak{B}_1$, $\overline{S}(t) = t^{-1}\int_0^t ds \exp[-i\mathbf{H}s]S$ exists in $\mathfrak{B}_1$ as a Riemann–Stieltjes integral with respect to the norm topology of $\mathfrak{B}_1$ (Ref. 7, p. 63, Theorem 3. 3. 2) and in fact we have $\overline{\phi}(t, S, A) = \mathrm{tr}\{\overline{S}(t)A\}, A \in \mathfrak{B}$. Thus $\overline{\phi}$ can be interpreted as a blf on $\mathfrak{B}$ and as such it is normal. Obviously, $\overline{\phi}$ is linear in S as well and since

$$|\overline{\phi}(t, S, A)| \leq t^{-1}\int_0^t ds\,|\mathrm{tr}[\mathbf{U}(t)S]A| \leq t^{-1} \times \int_0^t ds\,|\mathbf{U}(t)S|_1|A| = t^{-1}\int_0^t ds|S|_1|A| = |S|_1|A|,$$

it follows that $\overline{\phi}(S)$ is a blf on $\mathfrak{B}_1$ and has the representation $\overline{\phi}(t, S, A) = \mathrm{tr}\{S\overline{A}(-t)\}, \overline{A}(-t) \in \mathfrak{B}$. In fact, since

$$\begin{aligned}\overline{\phi}(t, S, A) &= t^{-1}\int_0^t ds\ \mathrm{tr}\{\exp[-iHs]S\exp[+iHs]A\} \\ &= t^{-1}\int_0^t ds\ \mathrm{tr}\{S\exp[+iHs]A\exp[-iHs]\} \\ &= t^{-1}\int_0^t ds\ \mathrm{tr}\{S\mathbf{U}(-s)A\},\end{aligned}$$

we can write

$$\overline{A}(t) = t^{-1}\int_0^t ds\mathbf{U}(s)A \text{ for } A \in \mathfrak{B},$$

where the integral is now defined with respect to the weak* topology on $\mathfrak{B}$.

Theorem 2: Let $S \in \mathfrak{B}_1$ and let $\{E_\kappa \,|\, \kappa \in K\}$ be the set of mutually orthogonal eigenprojectors of H. Then $\mathbf{P}_0S = \sum_\kappa E_\kappa SE_\kappa$ exists as an element of $\mathfrak{B}_1$ and

$$w^* - \lim_{t\to\pm\infty} \overline{S}(t) = \mathbf{P}_0 S$$

i.e., for any $A \in \mathfrak{B}_\infty$ we have $\lim_{t\to\pm\infty} \mathrm{tr}\overline{S}(t)A = \mathrm{tr}(\mathbf{P}_0 S)A$.

Remark: The above results were obtained earlier by Emch[2] (see also Ref. 8) for separable $\mathcal{H}$. Since we do not make this restrictive assumption and also since part of Emch's proof seems to have remained unpublished we give a complete proof.

Proof: (a) We show first that $\mathbf{P}_0 S \in \mathfrak{B}_1$ when $S \in \mathfrak{B}_1$. For separable $\mathcal{H}$ this follows at once from Ref. 4 p. 105, theorem III 8.7 (only separable Hilbert spaces are considered in Ref. 4). Hence we reduce the problem at hand to a separable subspace of $\mathcal{H}$ in case $\mathcal{H}$ is nonseparable. Since $S \in \mathfrak{B}_1, S$ is also contained in $\mathfrak{B}_2$ and we have for a finite number of mutually orthogonal eigenprojectors of H:

$$\sum_{k=1}^{n} |E_k S|_2^2 = \Big|\sum_{k=1}^{n} E_k S\Big|_2^2 \leq |E_p S|_2^2,$$

where E_p is the projector upon the subspace $E_p\mathcal{H} = \mathcal{H}_p$ of $\mathcal{H}$ spanned by eigenvectors of H. From the above relation it follows that $|E_\kappa S|_2$ can be nonzero only at most for a countably infinite subset $K_1 = K_1(S) \subset K$ of indices. Since $|E_\kappa S|_2 = 0$ implies $E_\kappa S = 0$ it follows that $E_\kappa S = 0$ unless $\kappa \in K_1(S)$. In the same way we conclude that SE_κ is nonzero only for an at most countably infinite subset $K_2(S)$ of K. Thus $\mathbf{P}_0 S$, if it exists, is of the form

$$\mathbf{P}_0 S = \sum_{k\in K_0} E_k S E_k,$$

where $K_0 = K_1 \cap K_2$. E_k may still project upon a nonseparable subspace of $\mathcal{H}$. If this is the case we can consider a basis for $E_k\mathcal{H}$ and associate an eigenprojector of H with each basis vector. Repeating the argument presented above we then conclude that there is an $E'_k \subset E_k$ such that $E_k S = E'_k S$, where E'_k projects upon a separable subspace of $\mathcal{H}$. Thus we may assume in the following that in the expression for $\mathbf{P}_0 S$ the E_k's are projectors upon separable subspaces of $\mathcal{H}$. Thus we may assume in the following that in the expression for $\mathbf{P}_0 S$ the E_k's are projectors upon separable subspaces of $\mathcal{H}$. Thus

$$E_1 = \sum_{k\in K_1} E_k \text{ and } E_2 = \sum_{k\in K_2} E_k,$$

being sums of a countable family of mutually orthogonal projectors, are well-defined projectors upon separable subspaces of $\mathcal{H}$. In fact we have $E_p S = E_1 S$ and $SE_p = SE_2$, so that $E_p S E_p = E_1 S E_2$. In order to show this, suppose there is an $x \in \mathcal{H}, x \neq 0$ such that $E_p S x \neq 0$ and $E_p S x \notin E_1\mathcal{H}$. Then there exists a $y \in E_p\mathcal{H} - E_1\mathcal{H}$ such that $(E_p S x, y) \neq 0$. This means that there is at least one $\kappa \in K - K_1$ with $(E_\kappa S x, y) \neq 0$. Hence $E_\kappa S \neq 0$ and we have a contradiction. We conclude that $E_p S = E_1 S$ and, in a similar way, that $SE_p = SE_2$. Writing

$$S = \sum_l \lambda_l (\cdot, u_l) v_l$$

and defining

$$P = \sum_l (\cdot, v_l) v_l$$

and

$$Q = \sum_l (\cdot, u_l) u_l$$

(which both project upon separable subspaces of $\mathcal{H}$) we see that both S and $E_p S E_p$ are contained in $\mathfrak{B}_1(\mathfrak{M})$, where $\mathfrak{M} = P\mathcal{H} \cup Q\mathcal{H} \cup E_1\mathcal{H} \cup E_2\mathcal{H}$ is a separable subspace of $\mathcal{H}$. Thus the theorem referred to above can be applied In particular $\mathbf{P}_0 S \in \mathfrak{B}_1(\mathcal{H})$ and $|\mathbf{P}_0 S|_1 \leq |S|_1$.

(b) It remains to show that

$$\lim_{t\to\pm\infty} \mathrm{tr}\overline{S}(t)A = \lim_{t\to\pm\infty} t^{-1}\int_0^t ds\ \mathrm{tr}S(t)A$$

exists and is equal to $\mathrm{tr}(\mathbf{P}_0 S)A$ for $S \in \mathfrak{B}_1$ and each $A \in \mathfrak{B}_\infty$. As Emch[2] pointed out this can be done conveniently by using von Neumann's mean ergodic theorem on the Hilbert space $\mathfrak{B}_2$. His argument runs as follows; Since $\mathfrak{F}$ is dense in $\mathfrak{B}_\infty$ there exists for $0 \neq S \in \mathfrak{B}_1, A \in \mathfrak{B}_\infty$ and $\epsilon > 0$ an $F \in \mathfrak{F}$ such that $|A - F| < \epsilon/(2|S|_1)$. Hence

$$\begin{aligned}
&|\mathrm{tr}(\overline{S}(t) - \mathbf{P}_0 S)A| \\
&\quad \leq |\mathrm{tr}(\overline{S}(t) - \mathbf{P}_0 S)F| + |\mathrm{tr}(\overline{S}(t) - \mathbf{P}_0 S)(A - F)| \\
&\quad \leq |\mathrm{tr}(\overline{S}(t) - \mathbf{P}_0 S)F| + (|\overline{S}(t)|_1 + |\mathbf{P}_0 S|_1)|A - F| \\
&\quad < |\mathrm{tr}(\overline{S}(t) - \mathbf{P}_0 S)F| + \epsilon.
\end{aligned}$$

Now, as $\overline{S}(t)$ and F are both contained in $\mathfrak{B}_2$, we have $\mathrm{tr}\{\overline{S}(t)F\} = t^{-1}\int_0^t ds\ (\mathbf{U}(s)S, F^*)_2$. Since $\mathbf{U}(t)$ is a continuous group of unitary operators on the Hilbert space $\mathfrak{B}_2$, it follows that H is self-adjoint on $\mathfrak{B}_2$ and von Neumann's mean ergodic theorem then yields

$$\lim_{t\to\pm\infty} (\overline{S}(t), F^*)_2 = (\mathbf{P}S, F^*)_2,$$

where P is the eigenprojector of H at the eigenvalue zero. It is then seen from the general expression

$$\mathbf{E}_\omega T = \int E_{\omega-\kappa} T d_\kappa E_\kappa, T \in \mathfrak{B}_2, H = \int \kappa d_\kappa E_\kappa$$

for the members of the resolution of the identity for H [see Ref. 9, Eq. (A33); the results stated in Theorem A. 2 of this reference are true even for nonseparable $\mathcal{H}$], that

$$\mathbf{P} = \lim_{\delta\to 0^+} (\mathbf{E}_\delta - \mathbf{E}_{-\delta})$$

(in the strong topology of $\mathfrak{B}_2$) equals $\mathbf{P}_0$ and the theorem is proven.

According to the above theorem $2\ \overline{\phi}(t.S,A)$ converges for t tending to $\pm\infty$ towards $\mathrm{tr}(\mathbf{P}_0 S)A$ for each $A \in \mathfrak{B}_\infty$. Suppose now that $\overline{\phi}(t,S,A)$ converges for t tending to $\pm\infty$ for each $A \in \mathfrak{B}$, i.e., that the limits

$$\overline{\psi}_\pm(A) = \lim_{t\to\pm\infty} \overline{\phi}(t,S,A)$$

exist for each $A \in \mathfrak{B}$. Obviously $\overline{\psi}_\pm(A)$ are blf on $\mathfrak{B}$, the boundedness being evident from the inequality $|\overline{\phi}(t,S,A)| \leq |S|_1|A|$. We show that $\overline{\psi}_\pm(A)$ are normal.

Theorem 3: Suppose $S(t)$ [or $\overline{S}(t)$] converges weakly towards the blf $\psi_\pm \in \mathfrak{B}^*$ (or $\overline{\psi}_\pm \in \mathfrak{B}^*$, respectively) for $t \to \pm\infty$. Then $\psi_\pm$ (or $\overline{\psi}_\pm$) is normal and has the unique representation $\psi_\pm = \mathbf{P}_0 S$ (or $\overline{\psi}_\pm = \mathbf{P}_0 S$ respectively).

Proof: Since the weak convergence of $S(t)$ implies the weak convergence of $\overline{S}(t)$, the limit being the same, we can restrict our attention to $\overline{S}(t)$.

We write $\overline{S}_n = \overline{S}(n)$ for $n = 1, 2, 3, \ldots$. Since $\overline{S}(t)$ converges weakly towards ψ_+ for t tending to infinity, it follows that $\{\overline{S}_n\}_{n=1}^\infty$ is a weak Cauchy sequence in $\mathfrak{B}_1$, i.e., for given $\epsilon > 0$ and $A \in \mathfrak{B}$ there exists an $n_0 = n_0(\epsilon, A)$ with $|\mathrm{tr}(\overline{S}_n - \overline{S}_m)A| < \epsilon$ for $n, m > n_0$. Since $\mathfrak{B}_1$, as the predual of the W^*-algebra $\mathfrak{B}(\mathcal{H})$, is weakly sequentially complete,[10] it follows that $\{\overline{S}_n\}$ has a weak limit

$\overline{S}_+$ in $\mathfrak{B}_1$. From the inequality $|\mathrm{tr}(\overline{S}(t) - \overline{S}_+)A| \leq |\mathrm{tr}(\overline{S}(t) - \overline{S}(n))A| + |\mathrm{tr}(\overline{S}_n - \overline{S}_+)A|$ it then follows that $\overline{S}(t)$ converges weakly towards $\overline{S}_+$ for t tending to infinity, since, by taking t and n sufficiently large, both terms on the right can be made arbitrarily small.

Since $\overline{S}(t)$ converges towards $\mathbf{P}_0 S$ in the weak* topology on $\mathfrak{B}_1$, and the weak topology is finer than the weak* topology, it follows that necessarily $\overline{S}_+ = \mathbf{P}_0 S$. The result for t tending to $-\infty$ follows in the same way.

In general $\overline{\psi}_\pm(A)$ will not exist for every $A \in \mathfrak{B}$. Let us therefore introduce the subsets $\mathfrak{D}_\pm \subset \mathfrak{B}$, consisting of those $A \in \mathfrak{B}$ for which $\overline{\psi}_\pm(A)$ exists. It is a routine matter to show that $\mathfrak{D}_\pm$ are closed in the uniform topology on $\mathfrak{B}$. Since $\mathfrak{B}_\infty \subset \mathfrak{D}_\pm$ and $\mathfrak{B}_\infty'' = \mathfrak{B}$ it follows[11] that the weak*-closure of $\mathfrak{D}_\pm$ is $\mathfrak{B}$ itself.

Let H have empty point spectrum and let $S = \rho$ be a density operator. Then $\mathfrak{D}_\pm$ do not coincide with $\mathfrak{B}$. This follows from Theorem 3 by remarking that if $\mathfrak{D}_\pm = \mathfrak{B}$ then $\overline{\psi}_\pm(A) = 0$ for every $A \in \mathfrak{B}$. On the other hand, $\overline{\psi}_\pm(I) = \mathrm{tr}\rho = 1$ and we have a contradiction. Emch[2] has pointed out that in this case $\overline{\psi}_\pm$ is no longer normal on $\mathfrak{D}_\pm$ (although ρ *does* represent a normal state on $\mathfrak{B}$). For separable $\mathfrak{H}$ this is directly evident from the inequality

$$0 = \sum_k \overline{\psi}_\pm(P_k) \neq \overline{\psi}_\pm\Big(\sum_k P_k\Big) = \overline{\psi}_\pm(I) = 1$$

for any set of mutually orthogonal one-dimensional projectors $\{P_k\}_{k=1}^\infty \subset \mathfrak{B}_\infty$ with the property

$$\sum_{k=1}^\infty P_k = I.$$

Thus $\overline{\psi}_\pm$ lacks the complete additivity property and hence cannot be normal on $\mathfrak{D}_\pm$. In particular $\overline{\psi}_\pm$ is not normal on the C^*-algebra $C = \{\alpha I + A \mid \alpha \in \mathfrak{R}, A \in \mathfrak{B}_\infty\} \subset \mathfrak{D}_\pm$. On the other hand $\overline{\psi}_\pm$ *is* normal on any von Neumann algebra $\mathfrak{N} \subset \mathfrak{D}_\pm$ for arbitrary $S \in \mathfrak{B}_1$. (A proof of this statement follows along the lines of the proof of Theorem 3.) As a nontrivial example consider the situation that H has empty point spectrum and that this is also the case for H_0, a second self-adjoint operator on $\mathfrak{H}$. Suppose further that the wave operators

$$\Omega_\pm = \underset{t\to\pm\infty}{s\text{-lim}} \exp[iHt]\exp[-iH_0t]$$

exist on $\mathfrak{H}$ as unitary operators, so that

$$\underset{t\to\pm\infty}{s\text{-lim}} \exp[iH_0t]\exp[-iHt] = \Omega_\pm^*$$

exist on $\mathfrak{H}$ as well (cf. the case of a simple scattering system with no bound states). Then it can be shown that

$$\underset{t\to\pm\infty}{s\text{-lim}} \exp[i\mathbf{H}_0t]\exp[-i\mathbf{H}t]S = \Omega_\pm^* S\Omega_\pm \equiv S_\pm$$

for any $S \in \mathfrak{B}_1$. Consequently, $\psi_\pm(A) = \bar{\psi}_\pm(A) = \mathrm{tr}S_\pm A$ for any A in the von Neumann algebra $\{H_0\}'$ {since $A = \exp[i\mathbf{H}_0t]A$ for such A}. The idea of considering the von Neumann algebra of observables that commute with the "free Hamiltonian" H_0 has recently been applied to the theory of scattering from long-range potentials[12] (see also Refs. 13 and 14).

Next we consider the question whether for H with empty point spectrum $\overline{S}(t)$ can have a weak limit for a general $S \in \mathfrak{B}_1$. Since, if this limit exists, $\overline{S}^*(t)$ also has a weak limit we restrict ourselves to Hermitian S. As the limit must be zero in the present case we can disprove its existence by exhibiting an $A \in \mathfrak{B}$ for which either $\overline{\psi}_\pm(A)$ does not exist or has a nonzero value. Since $\psi_\pm(A) = \mathrm{tr}SA$ for $A \in \{H\}'$ the above limit does indeed not exist if $\mathrm{tr}SA \neq 0$ for one such A, in particular if $\mathrm{tr}S \neq 0$. Unfortunately this criterion is not sufficient to rule out the existence of the above limit for every $S \in \mathfrak{B}_1$. As a counter example consider $S = (\cdot, x)x - (\cdot, y)y$ where x and y are orthonormal. Denoting by $\{E_\lambda\}$ the resolution of the identity for H we note that $W_\lambda = 1 - 2E_\lambda$ is unitary for every $\lambda \in \mathfrak{R}$ and that $f(\lambda) = (W_\lambda x, x)$ is a continuous function of λ which assumes both positive and negative values as λ runs through $\mathfrak{R}$. Thus there is a λ_0 with $f(\lambda_0) = 0$. But then $y = W_{\lambda_0} x$ is orthogonal to x and we have for this particular value of y, $\mathrm{tr}SA = (Ax, x) - (W_{\lambda_0}^* A W_{\lambda_0} x, x) = 0$ for every $A \in \{H\}'$.

IV. THE EXISTENCE OF ERGODIC LIMITS IN $\mathfrak{B}_1(\mathfrak{H})$

In Sec. III we showed that weak ergodic limits of elements of $\mathfrak{B}_1$ are again contained in $\mathfrak{B}_1$. Next we determine those $S \in \mathfrak{B}_1$ for which $\overline{S}(t)$ and $S(t)$ do have a weak ergodic limit in $\mathfrak{B}_1$. In fact we find that the set of $S \in \mathfrak{B}_1$ for which $\overline{S}(t)$ has a weak ergodic limit for $t \to \pm\infty$ coincides with the set for which $\overline{S}(t)$ possesses a strong limit in $\mathfrak{B}_1$.

It turns out to be convenient for the further discussion to separate H into its discrete and continuous part, i.e., $H = H_p + H_c$, where $H_p = E_p H E_p$ and $H_c = E_c H E_c$. Here is E_p the projector upon the subspace of $\mathfrak{H}$ spanned by the eigenvectors of H and $E_c = I - E_p$ is its complement in $\mathfrak{H}$. Thus H_p has a pure point spectrum and H_c a pure continuous spectrum on the respective subspaces $\mathfrak{H}_p = E_p\mathfrak{H}$ and $\mathfrak{H}_c = E_c\mathfrak{H}$ of $\mathfrak{H}$. Let $S \in \mathfrak{B}_1$. Then we may write $S = S_{pp} + S_{pc} + S_{cp} + S_{cc}$, where $S_{ij} = E_i S E_j$, $i, j \in \{p, c\}$. Furthermore, if we write $S_{ij}(t) = \mathrm{U}(t)S_{ij}$ then obviously $S_{ij}(t) = \exp[-iHt]E_i S E_j \exp[+iHt] = \exp[-iH_i t]S_{ij}\exp[+iH_j t]$. In the sequel

$$s_1 - \lim_{t\to\pm\infty} \overline{S}_{ij}(t)$$

denotes the limit for $t \to \pm\infty$ of $\overline{S}_{ij}(t)$ in the strong (i.e., norm) topology of $\mathfrak{B}_1$ and similarly

$$w_1 - \lim_{t\to\pm\infty} \overline{S}_{ij}(t)$$

stands for the same limit with respect to the weak topology on $\mathfrak{B}_1$.

Lemma 2: $s_1 - \lim_{t\to\pm\infty} \overline{S}_{pp}(t)$ exists and is equal to $\mathbf{P}_0 S_{pp} = \mathbf{P}_0 S$.

Proof: Let us denote [see part (a) of the proof of Theorem 2] $\{E_k \mid k \in K_1(S)\}$ by $\{F_k\}_{k=1}^\infty$ and similarly $\{E_k \mid k \in K_2(S)\}$ by $\{G_k\}_{k=1}^\infty$. Then

$$E_1 = \sum_{k=1}^\infty F_k \quad \text{and} \quad E_2 = \sum_{k=1}^\infty G_k.$$

By application of Lemma 4.1 in Ref. 6 it follows from the fact that

$$\sum_{k=1}^p F_k$$

converges towards E_1 and

$$\sum_{k=1}^p G_k$$

towards E_2 in the strong operator topology on $\mathfrak{H}$ when p tends to infinity, that for given $\epsilon > 0$ there exists an

$n_1 = n_1(\epsilon)$ such that $|S_{pp} - S_n|_1 = |E_1 S E_2 - S_n|_1 < \epsilon$ for $n > n_1$. Here S_n is defined as

$$S_n = \sum_{k=1}^{n} \sum_{l=1}^{n} F_k S_{pp} G_l .$$

Also for given $\epsilon > 0$ there is an $m_0 = m_0(\epsilon)$, such that in $S_{pp} = \sum_{j=1}^{m} \lambda_j(\cdot, u_j)v_j + \sum_{j=m+1}^{\infty} \lambda_j(\cdot, u_j)v_j = S_{pp}^m + \sum_{j=m+1}^{\infty} \lambda_j(\cdot, u_j)v_j$ the second part on the right has trace norm smaller than ϵ for $m > m_0$. Furthermore, if

$$S_n^m = \sum_{k=1}^{n} \sum_{l=1}^{n} F_k S_{pp}^m G_l$$

then there exists for given $\epsilon > 0$ an $n_2 = n_2(\epsilon)$ such that $|S_{pp}^m - S_n^m| < \epsilon$ for $n > n_2$ for the same reasons as discussed above in connection with S_{pp} and S_n.

Let us now take, for such an $\epsilon > 0$, a fixed $m > m_0(\epsilon)$ and $n > \max\{n_1(\epsilon), n_2(\epsilon)\}$. Then

$$\begin{aligned}
&|\overline{S}_{pp}(t) - \mathbf{P}_0 S_{pp}|_1 \\
&\quad \leq |\overline{S}_{pp}(t) - \overline{S}_{pp}^m(t)|_1 + |\overline{S}_{pp}^m(t) - \overline{S}_n^m(t)|_1 \\
&\quad\quad + |\overline{S}_n^m(t) - \mathbf{P}_0 S_n^m|_1 + |\mathbf{P}_0(S_n^m - S_{pp}^m)|_1 \\
&\quad\quad + |\mathbf{P}_0(S_{pp}^m - S_{pp})|_1 \\
&\quad \leq |S_{pp} - S_{pp}^m|_1 + |S_{pp}^m - S_n^m|_1 + |\overline{S}_n^m(t) - \mathbf{P}_0 S_n^m|_1 \\
&\quad\quad + |S_n^m - S_{pp}^m|_1 + |S_{pp}^m - S_{pp}|_1 < 4\epsilon + |\overline{S}_n^m - \mathbf{P}_0 S_n^m|_1 .
\end{aligned}$$

We write

$$S_n^m = \sum_{k=1}^{n} \sum_{l=1}^{n} \sum_{j=1}^{m} \lambda_j T_{jkl}$$

with

$$T_{jkl} = F_k(\cdot, u_j)v_j G_l = (\cdot, G_l u_j)F_k v_j$$

so that $(0 \leq \lambda_j \leq 1)$,

$$|\overline{S}_n^m(t) - \mathbf{P}_0 S_n^m|_1 \leq \sum_{k=1}^{n} \sum_{l=1}^{n} \sum_{j=1}^{m} |\overline{T}_{jkl}(t) - \mathbf{P}_0 T_{jkl}|_1 .$$

Denoting the eigenvalue of H_p corresponding to F_k by δ_k and the one corresponding to G_l by ϵ_l, we have

$$\mathbf{P}_0 T_{jkl} = \begin{cases} 0 , & \delta_k \neq \epsilon_l, \\ T_{jkl}, & \delta_k = \epsilon_l. \end{cases}$$

Also

$$\begin{aligned}
\overline{T}_{jkl}(t) &= t^{-1} \int_0^t ds\, \exp[-iHs]\{F_k(\cdot, u_j)v_j G_l\} \exp[+iHs] \\
&= t^{-1} \int_0^t ds\, \exp[-i\delta_k s]\{F_k(\cdot, u_j)v_j G_l\} \exp[+i\epsilon_l s]
\end{aligned}$$

so that

$$\overline{T}_{jkl}(t) = \begin{cases} \dfrac{\exp[-i(\delta_k - \epsilon_l)t] - 1}{-i(\delta_k - \epsilon_l)t}\, T_{jkl}, & \delta_k \neq \epsilon_l \\ T_{jkl}, & \delta_k = \epsilon_l. \end{cases}$$

Thus

$$|\overline{T}_{jkl}(t) - \mathbf{P}_0 T_{jkl}|_1 = 0 \text{ for } \delta_k = \epsilon_l$$

and

$$\begin{aligned}
|\overline{T}_{jkl}(t) - \mathbf{P}_0 T_{jkl}|_1 &= \left| \frac{\exp[-i(\delta_k - \epsilon_l)t] - 1}{-i(\delta_k - \epsilon_l)t} \right| |T_{jkl}|_1 \\
&\leq \left| \frac{\exp[-i(\delta_k - \epsilon_l)t] - 1}{-i(\delta_k - \epsilon_l)t} \right| \text{for } \delta_k \neq \epsilon_l .
\end{aligned}$$

If follows that $\overline{T}_{jkl}(t)$ converges strongly towards $\mathbf{P}_0 T_{jkl}$ on $\mathfrak{B}_1$ for any $j \in \{1, \ldots, m\}$ and $k, l \in \{1, \ldots, n\}$ and hence, since m and n are fixed, we conclude that

$$s_1\text{-}\lim_{t \to \pm\infty} \overline{S}_{pp}(t) = \mathbf{P}_0 S_{pp}.$$

Lemma 3: $s_1\text{-}\lim_{t \to \pm\infty} \overline{S}_{pc}(t) = s_1 - \lim_{t \to \pm\infty} \overline{S}_{cp}(t) = 0.$

Proof: The first part of the proof goes along the same lines as in the proof of Lemma 2. Thus there exist for given $\epsilon > 0$ a fixed n and m such that $|S_{pc} - S_n^m|_1 < \epsilon$, where S_n^m is now given by

$$S_n^m = \left(\sum_{k=1}^{n} \right) F_k\, S_{pc}^m E_c .$$

Here S_{pc}^m is defined in the same way as S_{pp}^m in the proof of Lemma 2. Hence:

$$|\overline{S}_{pc}(t)|_1 \leq |\overline{S}_{pc}(t) - \overline{S}_n^m(t)|_1 + |\overline{S}_n^m(t)|_1 < \epsilon + |\overline{S}_n^m(t)|_1 .$$

Next we write

$$S_n^m = \sum_{k=1}^{n} \sum_{j=1}^{m} F_k \lambda_j(\cdot, u_j)v_j E_c = \sum_{k=1}^{n} \sum_{j=1}^{m} \lambda_j T_{kj},$$

and we have

$$|\overline{S}_n^m(t)|_1 \leq \sum_{k=1}^{n} \sum_{j=1}^{m} |\overline{T}_{kj}(t)|_1 .$$

Since

$$\begin{aligned}
\overline{T}_{kj}(t) &= t^{-1} \int_0^t ds\, \exp[-iHs]F_k(\cdot, u_j)v_j E_c \exp[+iHs] \\
&= t^{-1} \int_0^t ds\, \exp[-i\delta_k s]F_k(\cdot, u_j)v_j E_c \exp[iH_c s] \\
&= t^{-1} \int_0^t ds(\cdot, \exp[-i(H_c - \delta_k)s]E_c u_j)F_k v_j \\
&= (\cdot, t^{-1} \int_0^t ds\, \exp[-i(H_c - \delta_k)s]E_c u_j)F_k v_j,
\end{aligned}$$

we obtain

$$|\overline{T}_{kj}(t)|_1 = |t^{-1} \int_0^t ds\, \exp[-i(H_c - \delta_k)s]Eu_j\,|\,|F_k v_j|.$$

Next we apply von Neumann's mean ergodic theorem to the expression in the first factor on the right. Since only quantities pertaining to $\mathcal{H}_c$ occur and $H_c - \delta_k$ has empty point spectrum on this subspace, it follows that the first factor on the right tends to zero for t tending to $\pm\infty$. This being true for each $k \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$ we conclude that

$$s_1\text{-}\lim_{t \to \pm\infty} \overline{S}_{pc}(t) = 0.$$

In the same way the result for $\overline{S}_{cp}(t)$ is obtained.

Theorem 4: Let $S \in \mathfrak{B}_1$ with $\mathrm{tr} S_{cc} \neq 0$. Then $\overline{S}(t)$ does not have a strong limit in $\mathfrak{B}_1$ for $t \to \pm\infty$. If $S_{cc} = 0$ then

$$s_1\text{-}\lim_{t \to \pm\infty} \overline{S}(t) = \mathbf{P}_0 S.$$

Proof: Since according to what has been proven in Lemma's 2 and 3, $\overline{S}_{pp}(t)$, $\overline{S}_{pc}(t)$, and $\overline{S}_{cp}(t)$ have strong limits for $t \to \pm\infty$, it follows that $\overline{S}(t)$ has a strong limit if and only if $\overline{S}_{cc}(t) = \overline{S}(t) - \overline{S}_{pp}(t) - \overline{S}_{pc}(t) - \overline{S}_{cp}(t)$ has a strong limit for $t \to \pm\infty$.

Now $\mathrm{tr}\{\overline{S}_{cc}(t)A\} = t^{-1} \int_0^t ds\, \mathrm{tr}\{\exp[-iH_c s]S_{cc}$

$\exp[+iH_c s]E_c A E_c\}$ so that only quantities pertaining to $\mathcal{H}_c$ occur. On this subspace H_c has empty point spectrum and according to Theorem 3 (cf. the remarks at the end of Sec. III) $\overline{S}_{cc}(t)$ does not have a weak limit on $\mathfrak{B}_1$ for $t \to \pm\infty$ if $\mathrm{tr}S_{cc} \neq 0$. This, in turn implies that $\overline{S}_{cc}(t)$ does not have a strong limit if $\mathrm{tr}S_{cc} \neq 0$ and hence the same holds true for $\overline{S}(t)$.

Corollary 1: Let ρ be a normal state. Then

$$s_1\text{-}\lim_{t\to\pm\infty}\bar{\rho}(t)$$

exists if and only if $\rho = \rho_{pp}$. In that case

$$s_1\text{-}\lim_{t\to\pm\infty}\bar{\rho}(t) = \mathbf{P}_0\rho.$$

Proof: Suppose $\bar{\rho}(t)$ has a strong limit. Then, necessarily $\rho_{cc} \equiv 0$. Thus, writing

$$\rho = \sum_{k=1}^{\infty} \lambda_k(\cdot, u_k)u_k, \qquad \lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant 0,$$

we obtain

$$\mathrm{tr}\rho_{cc} = \mathrm{tr}\left\{\sum_{k=1}^{\infty} \lambda_k(\cdot, E_c u_k)E_c u_k\right\} = \sum_{k=1}^{\infty} \lambda_k |E_c u_k|^2 = 0,$$

so that $E_c u_k = 0, k = 1, 2, 3, \cdots$. Hence $\rho = \rho_{pp}$. The converse statement is contained in Lemma 2.

As a second corollary we recover Theorem 5 of Ref. 3. Let $S \in \mathfrak{B}_1$ be invariant. Then, since $S = \overline{S}(t)$, we have $S = \mathbf{P}_0 S$, i.e., $S = S_{pp}$. Hence, for $A \in \mathfrak{B}$, we have $AS = (AS)_{pp} + (AS)_{cp}$ so that $\overline{AS}(t) = \overline{A}(t)S$ converges strongly towards $\mathbf{P}_0(AS) = (\mathbf{P}_0 A)S$ in $\mathfrak{B}_1$ [here $\overline{A}(t)$ and $\mathbf{P}_0 A$ are defined with respect to the weak* topology on $\mathfrak{B}$]. Thus:

Corollary 2: Let $S \in \mathfrak{B}_1$ be invariant and let $A \in \mathfrak{B}$. Then

$$s_1\text{-}\lim_{t\to\pm\infty}\overline{AS}(t) = s_1\text{-}\lim_{t\to\pm\infty}\overline{A}(t)S$$

and

$$s_1\text{-}\lim_{t\to\pm\infty}\overline{SA}(t) = s_1\text{-}\lim_{t\to\pm\infty}\overline{SA}(t)$$

exist and are equal to $\mathbf{P}_0(AS) = (\mathbf{P}_0 A)S$ and $\mathbf{P}_0(SA) = S(\mathbf{P}_0 A)$, respectively.

So far we have only considered the asymptotic behavior for large times of $S(t)$, the time average of $S(t)$. Since the existence of

$$w_1\text{-}\lim_{t\to\pm\infty}S(t)$$

implies the existence of

$$w_1\text{-}\lim_{t\to\pm\infty}\overline{S}(t)$$

it follows that $S(t)$ does not have a weak limit if $\mathrm{tr}S_{cc} \neq 0$.

Next we consider $S_{pp}(t)$. We have for $A \in \mathfrak{B}$

$$\mathrm{tr}\{S_{pp}(t)A\} = \mathrm{tr}\{\exp[-iH_p t]S_{pp}\exp[+iH_p t]E_p A E_p\}.$$

Since only quantities pertaining to $\mathcal{H}_p$ occur in this expression, it follows that the situation is equivalent to the one where the Hamiltonian has a pure point spectrum. As is well known there is no weak ergodic limit in that situation, unless S is invariant. This is easily verified by observing that if

$$w_1\text{-}\lim_{t\to\pm\infty}S(t)$$

exists, it must be equal to $\mathbf{P}_0 S$. By taking $A = (\cdot, u)v$, where u and v are eigenvectors of H with corresponding eigenvalues δ and ϵ, respectively, we obtain $\mathrm{tr}\{S(t)A\} = \exp[-i(\delta - \epsilon)t](Sv, u)$, whereas, on the other hand, $\mathrm{tr}(\mathbf{P}_0 S)A = 0$ unless $\delta = \epsilon$. In the latter case $\mathrm{tr}(\mathbf{P}_0 S)A = (Sv, u)$. It follows that S is not allowed to possess non-vanishing matrix elements between eigenvectors of H at different eigenvalues, i.e., S must be invariant. Thus we conclude that a second necessary condition for the existence of

$$w_1\text{-}\lim_{t\to\pm\infty}S(t)$$

is that S_{pp} is invariant. Since for a normal state ρ we have $\rho_{pc} = \rho_{cp} = 0$ the following result holds:

Theorem 5: Let ρ be a normal state. Then

$$w_1\text{-}\lim_{t\to\pm\infty}\rho(t)$$

exists if and only if ρ is invariant.

It remains to consider $S_{pc}(t)$ and $S_{cp}(t)$ for arbitrary $S \in \mathfrak{B}_1$. In general not much can be said concerning their ergodic behavior. However, in case H_c, when restricted to $\mathcal{H}_c$, has absolutely continuous spectrum only, then the following lemma holds:

Lemma 4: Let H_c, restricted to $\mathcal{H}_c$ have absolutely continuous spectrum only. Then

$$w_1\text{-}\lim_{t\to\pm\infty}S_{pc}(t) = w_1\text{-}\lim_{t\to\pm\infty}S_{cp}(t) = 0.$$

Proof: In the same way as in the proof of Lemma 3 we obtain for $A \in \mathfrak{B}$;

$$|\mathrm{tr}S_{pc}(t)A| < \epsilon\,|A| + \sum_{k=1}^{n}\sum_{j=1}^{m}|\mathrm{tr}T_{kj}(t)A|,$$

where

$$T_{kj}(t) = \exp[-iHt]F_k(\cdot, u_j)v_j E_c \exp[+iHt]$$
$$= \exp[-i\delta_k t]F_k(\cdot, u_j)v_j E_c \exp[iH_c t].$$

Thus

$$|\mathrm{tr}T_{kj}(t)A| = |(\exp[iH_c t]E_c A F_k v_j, u_j)|,$$
$$= \left|\int_{-\infty}^{\infty} \exp i\lambda t\, d\mu(\lambda)\right|,$$

where

$$\mu(\lambda) = (E_{c\lambda}E_c A F_k v_j, u_j),$$

$\{E_{c\lambda} | \lambda \in \mathfrak{R}\}$ being the resolution of the identity of H_c in $\mathcal{H}_c$. Since H_c has absolutely continuous spectrum only, the Radon–Nikodym derivative of $\mu(\lambda)$ with respect to the Lebesgue measure exists and we have according to the Riemann-Lebesgue lemma

$$\lim_{t\to\pm\infty}\int_{-\infty}^{+\infty}\exp[i\lambda t]d\mu(\lambda) = \lim_{t\to\pm\infty}\int_{-\infty}^{+\infty}\exp[i\lambda t]\frac{d\mu(\lambda)}{d\lambda}\cdot d\lambda = 0.$$

Hence $T_{kj}(t)$ converges weakly to zero for $t \to \pm\infty$ for each $k \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$ and we conclude that

$$w_1\text{-}\lim_{t\to\pm\infty}S_{pc}(t) = 0$$

The proof for $S_{cp}(t)$ is obtained in the same manner.

Comparing the above result with Lemma 3 the question rises whether Lemma 4 does hold with respect to

the strong topology on $\mathfrak{B}_1$ instead of the weak topology. That this cannot be the case follows immediately by considering $S = T_{kj}$. Then $|S(t)|_1 = |T_{kj}(t)| = |E_c u_j|.|F_k v_j|$, i.e., in general $S(t)$ does not converge strongly towards zero.

ACKNOWLEDGMENT

One of the authors (A. T.) is indebted to Dr. Th. Niemeyer for several useful discussion on the present subject.

* Supported in part by a grant from the National Research Council of Canada

† Supported by the Foundation for Fundamental Research on Matter (FOM) which is sponsored by the Netherlands Organization for Advancement of Pure Research (ZWO).

[1]An extensive discussion of minimal norm ideals of completely continuous operators can be found in: R. Schatten, *Norm Ideals of Completely Continuous Operators* (Springer-Verlag Berlin 1960), see also Refs. 4 and 5.

[2]G. Emch, J. Math. Phys. 7, 1413 (1966).
[3]J. E. Moyal, J. Math. Phys. 10, 506 (1968).
[4]I. C. Gohberg and M. G. Krein *Introduction to the Theory of Linear Non Self-Adjoint Operators* (American Mathematical Society, Providence, R.I., 1969).
[5]I. C. Gohberg and M. G. Krein *Theory and Applications of Volterra Operators in Hilbert Space* (American Mathematical Society, Providence, R.I., 1970).
[6]E. Prugovečki and A. Tip, "Semi-groups of rank-preserving transformers on minimal norm ideals in $B(H)$" (to be published).
[7]E. Hille and R. S. Phillips, *Functional Analysis and Semigroups* (American Mathematical Society, Providence, R.I., 1957).
[8]G. Emch, Lect. Theor. Phys. 8, 65 (1966).
[9]E. Prugovečki, J. Math. Phys. 13, 969 (1972).
[10]C. A. Akemann, Trans. Am. Math. Soc. 126 286 (1967), Corollary III 3.
[11]J. Dixmier, *Les algèbres d'opérateurs dans l'espace Hilbertien* (algèbres de von Neumann) (Ganthier-Villars, Paris, 1969), 2nd ed., p. 42, Corollary 1.
[12]W. O. Amrein, Ph. A. Martin, and B. Misra, Helv. Phys. Acta 43, 313 (1970).
[13]R. B. Lavine, J. Funct. Anal. 5, 368 (1970).
[14]E. Prugovečki, Nuovo Cimento B 4, 105 (1971).

Green's function for electromagnetic scattering from a random rough surface

John A. DeSanto

Naval Research Laboratory, Washington, D.C. 20375
(Received 4 May 1973)

The stochastic Green's matrix is calculated for a random rough surface with Gaussian statistics and a magnetic boundary condition. The techniques we use are similar to those developed for the scalar and elastic cases. The coupled surface integral equations which are derived are the Green's function version of the Franz formulas. These integral equations are represented in k-space in a certain gauge and a Feynman-diagram-like interpretation is given to each term in the equations. The diagram rules have many formal similarities with the scalar and elastic rules. By using partial summation techniques, the mean and second moment of this Green's function are shown to be solutions to Dyson and Bethe–Salpeter equations respectively. The Green's function is applied to a scattering problem. Some approximations and simple examples are presented. The lowest order approximations agree with the standard literature results. The main advantage of the diagram method, its systematic presentation of higher order approximations, is stressed.

I. INTRODUCTION

In previous papers[1,2] we presented a formal diagram expansion method analogous to that used in random media propagation problems[3] to calculate the Green's function for a random rough surface. In Ref. 1 we considered the scalar case of a hard (Neumann) boundary, and in Ref. 2 the elastic case of a stress free boundary. In both Refs. 1 and 2 we applied the Green's function to a scattering problem. We derived the Dyson equation for the mean of the Green's function, the Bethe–Salpeter equation for the second moment of the Green's function, and presented some examples of lowest order coherent and incoherent scattering. The diagram method gives a systematic approach to higher order corrections.

In this paper we extend our previous results to calculate the magnetic Green's dyadic for scattering from a perfectly conducting half-space bounded by a random rough surface. The surface has Gaussian height distribution and arbitrary correlation function. The dyadic Green's function approach is explained in the books by Morse and Feshbach[4] and Hauser,[5] the latter includes a discussion of the Kirchhoff approximation with dyadics. Previously, Levine and Schwinger[6] used dyadics to discuss scattering from a plane perfect conductor in a homogeneous medium. Saxon[7] discussed a tensor scattering matrix for electromagnetic scattering by a finite obstacle and derived reciprocity relations and the cross section theorem in a concise way. Twersky[8] used a dyadic formalism in multiple scattering of electromagnetic waves by arbitrary configurations of arbitrary scatterers, and Mitzner[9] used the Rayleigh perturbation method and dyadic notation on random interface problems. A general reference to electromagnetic scattering from random rough surfaces based on the Kirchhoff approximation can be found in the book by Beckmann and Spizzichino.[10]

Our approach has certain similarities with the above papers in that we work with the dyadic Green's function, but we are also interested in providing a systematic formalism to calculate higher order corrections. The diagram method provides such a formalism.

In Sec. II we present the basic formalism of the problem and derive, using Green's theorem, the coupled integral equations for what we define as the magnetic surface Green's dyadic (matrix). The integral equations are the Green's function generalization of the Franz formulas rather than the equivalent but more common Stratton–Chu formulas.[11]

We introduce the Fourier transforms of these Green's dyadics in Sec. III, and by using a gauge condition argument derive the integral equations in momentum or k space. An alternative Green's function is defined, the kernel of its integral equation is factored into propagator-vertex-interaction form, and a diagram piece is assigned to each of these parts. The problem is still, up to this point, deterministic. It is shown that the interaction term is the same as that of Refs. 1 and 2, and the propagator is the same scalar free space Green's function as in Ref. 1. In Ref. 2 the propagator was a 1-index object. In Ref. 1 the vertex was a scalar function, and in Ref. 2 was a 3-index object. In this paper the vertex is a 2-index (matrix) object. There is a strong formal similarity (save for the number of indices) between the diagram notation in Refs. 1 and 2 and in this paper.

In Sec. IV the statistical concepts are introduced and the Dyson and Bethe–Salpeter equations are derived for the mean and second moments, respectively, of the alternative Green's dyadic introduced in Sec. III. Additional diagram rules for the statistical case are shown. Some approximations of these integral equations are briefly discussed and examples of lowest order coherent and incoherent contributions to the intensity are presented. Their connection to the standard literature values is noted.

The summary and conclusions are in Sec. V, and there is an appendix on the calculation of the kernel terms for the integral equations in Sec. II.

II. INTEGRAL EQUATIONS

Our problem is to calculate the Green's dyadic (matrix) $\Gamma_{in}(\mathbf{x},\mathbf{x}'')$ which satisfies the inhomogeneous differential equation (harmonic time dependence $e^{-i\omega t}$ is assumed throughout)

$$\partial_l\partial_l\Gamma_{in}(\mathbf{x},\mathbf{x}'') - \partial_i\partial_m\Gamma_{mn}(\mathbf{x},\mathbf{x}'') + k_0^2\Gamma_{in}(\mathbf{x},\mathbf{x}'') = -\delta_{in}\,\delta(\mathbf{x}-\mathbf{x}'') \quad (1)$$

in the region $V[z \geq h(x_\perp)]$ indicated in Fig. 1. Here k_0 is the free space wavenumber, the abbreviation $\partial_i = \partial/\partial x_i$ is used ($i = 1, 2, 3$), δ_{ij} is the Kronecker delta, $\delta(\mathbf{x})$ is the three-dimensional Dirac delta function, and the convention of summing on repeated indices is assumed. Equation (1) is just the matrix version of the dyadic equation

$$\nabla x\nabla x\mathbf{\Gamma}(\mathbf{x},\mathbf{x}') - k_0^2\mathbf{\Gamma}(\mathbf{x},\mathbf{x}') = \mathbf{I}\,\delta(\mathbf{x}-\mathbf{x}')$$

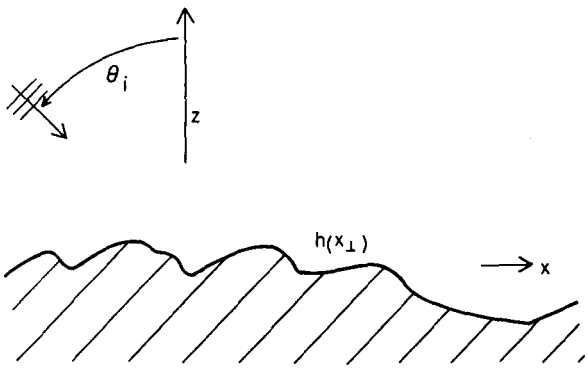

FIG. 1. Plane wave incident on a random rough surface $z = h(x_\perp)$. The region $z < h(x_\perp)$ is here considered to be a perfect conductor.

where I is the unit dyad. We use matrix notation throughout. The notation $\mathbf{x}$ is used for a 3-vector in V, $x_\perp$ for the transverse components of the 3-vector, $x_\perp = (x,y)$, and $\mathbf{x}_s$ for a 3-vector evaluated on the surface $z = h(x_\perp)$, $\mathbf{x}_s = (x_\perp, h(x_\perp))$. Here the height $h(x_\perp)$ is a Gaussian distributed random variable and the region $z < h(x_\perp)$ is considered to be a perfectly conducting half-space. In addition to satisfying the differential equation (1), $\Gamma_{in}(\mathbf{x},\mathbf{x}'')$ satisfies a boundary condition when $z = h(x_\perp)$ and asymptotic boundary conditions when its arguments approach infinity.

The free space Green's dyadics $\Gamma^{0\pm}_{in}$ both satisfy the same differential equation as Γ_{in}

$$\partial_l\partial_l\Gamma^0_{ij}(\mathbf{x}',\mathbf{x}) - \partial_i\partial_m\Gamma^0_{mj}(\mathbf{x}',\mathbf{x}) + k_0^2\Gamma^0_{ij}(\mathbf{x}',\mathbf{x}) = -\delta_{ij}\delta(\mathbf{x}'-\mathbf{x}) \quad (2)$$

and the asymptotic boundary conditions as $|\mathbf{x}'-\mathbf{x}| \to \infty$ of outgoing (+) or incoming (−) waves. The ± superscripts are temporarily dropped. Explicitly, Γ^0_{ij} is given by[6]

$$\Gamma^0_{ij}(\mathbf{x}',\mathbf{x}) = \delta_{ij}G^0(\mathbf{x}',\mathbf{x}) + k_0^{-2}\partial_i\partial_j G^0(\mathbf{x}',\mathbf{x}), \quad (3)$$

where G^0, the scalar free space Green's function satisfies the differential equation

$$(\partial_l\partial_l + k_0^2)G^0(\mathbf{x}',\mathbf{x}) = -\delta(\mathbf{x}'-\mathbf{x}) \quad (4)$$

and is given by

$$G^0(\mathbf{x}',\mathbf{x}) = \exp(ik_0|\mathbf{x}'-\mathbf{x}|)/|\mathbf{x}'-\mathbf{x}|4\pi. \quad (5)$$

Multiplying (1) from the left by $\Gamma^0_{ij}(\mathbf{x}',\mathbf{x})$ and subtracting from it (2) multiplied from the right by $\Gamma_{in}(\mathbf{x},\mathbf{x}'')$ yields the identity

$$\Gamma_{jn}(\mathbf{x},\mathbf{x}'')\delta(\mathbf{x}'-\mathbf{x}) - \Gamma^0_{jn}(\mathbf{x}',\mathbf{x})\delta(\mathbf{x}-\mathbf{x}'') = \partial_l K_{jln}(\mathbf{x}',\mathbf{x},\mathbf{x}'') \quad (6)$$

with

$$K_{jln}(\mathbf{x}',\mathbf{x},\mathbf{x}'') = \Gamma^0_{ij}(\mathbf{x}',\mathbf{x})[\partial_l\Gamma_{in}(\mathbf{x},\mathbf{x}'') - \partial_i\Gamma_{ln}(\mathbf{x},\mathbf{x}'')] - [\partial_l\Gamma^0_{ij}(\mathbf{x}',\mathbf{x}) - \partial_i\Gamma^0_{lj}(\mathbf{x}',\mathbf{x})]\Gamma_{in}(\mathbf{x},\mathbf{x}''). \quad (7)$$

Multiplying (6) by the step function $\theta(z - h(x_\perp))$, where

$$\theta(z) = \begin{cases} 1, & z > h(x_\perp), \\ 0, & z < h(x_\perp), \end{cases}$$

integrating over all space $\int d\mathbf{x}$, and doing a partial integration on the resulting integral term (with the neglect of the infinite surface term due to the radiation condition on the Green's functions) yields, for $z'' > h(x''_\perp)$,

$$\Gamma_{jn}(\mathbf{x}',\mathbf{x}'')\theta(z' - h(x'_\perp)) = \Gamma^0_{jn}(\mathbf{x}',\mathbf{x}'') - \int dx_\perp N_l(x_\perp)K_{jln}(\mathbf{x}',\mathbf{x}_s,\mathbf{x}''), \quad (8)$$

where

$$N_l(x_\perp) = \delta_{l3} - \partial_{l\perp}h(x_\perp)$$

is the normal into V. It can be shown that (8) is just the Green's dyadic version of the Franz diffraction formulas. The latter is equivalent to but much less than the Stratton–Chu formulas.[11] Since we wish to calculate the magnetic (μ) Green's function we choose the boundary condition

$$(\mathbf{N}(x_\perp)x\nabla x\Gamma(\mathbf{x}_s,\mathbf{x}''))_{in} = N_l(x_\perp)\partial_l\Gamma_{in}(\mathbf{x}_s,\mathbf{x}'') - N_j(x_\perp)\partial_i\Gamma_{jn}(\mathbf{x}_s,\mathbf{x}'') = 0. \quad (9)$$

Further, using (3), we can write

$$\partial_l\Gamma^0_{ij}(\mathbf{x}',\mathbf{x}) - \partial_i\Gamma^0_{lj}(\mathbf{x}',\mathbf{x}) = (\delta_{lj}\partial'_i - \delta_{ij}\partial'_l)G^0(\mathbf{x}',\mathbf{x}). \quad (10)$$

Substituting (9) and (10) into (8) and differentiating by ∂'_j, it is obvious that for $z' > h(x'_\perp)$ and $\mathbf{x}' \neq \mathbf{x}''$

$$\partial'_j\Gamma_{jn}(\mathbf{x}',\mathbf{x}'') = 0. \quad (11)$$

Hence the Green's dyadic is transverse as is usual in these problems.[9]

Next, let $\mathbf{x}' \to \mathbf{x}'_s$ through positive z' values. Equation (8) becomes, using (7), (9), (10) and the results in the Appendix

$$1/2(\delta_{ji} + N_j(x'_\perp)\delta_{i3})\Gamma^\mu_{in}(\mathbf{x}'_s,\mathbf{x}'') = \Gamma^0_{jn}(\mathbf{x}'_s,\mathbf{x}'') + \int dx_\perp P_{ji}(\mathbf{x}'_s,\mathbf{x}_s)\Gamma^\mu_{in}(\mathbf{x}_s,\mathbf{x}''), \quad (12)$$

where the μ superscript has to do with the fact that Γ_{in} satisfies (9), and with

$$P_{ji}(\mathbf{x}'_s,\mathbf{x}_s) = N_l(x_\perp)P_{ilj}(\mathbf{x}'_s - \mathbf{x}_s), \quad (A4)$$

$$P_{ilj}(\mathbf{x}_s) = \frac{1}{(2\pi)^3}\int d\mathbf{k}G^0(k)e^{i\mathbf{k}\cdot\mathbf{x}_s}R_{ilj}(\mathbf{k}), \quad (A5)$$

$$R_{ilj}(\mathbf{k}) = i\{[k_{i\perp} + \delta_{i3}P(K^2/k_3)]\delta_{lj} - [k_{l\perp} + \delta_{l3}P(K^2/k_3)]\delta_{ij}\}, \quad (A6)$$

where P stands for the Cauchy principle value distribution and $K^2 = k_0^2 - k_\perp^2$. Here $G^0(k) = (k^2 - k_0^2)^{-1}$ is the Fourier transform of $G^0(\mathbf{x})$. Further, defining the magnetic (μ) surface (s) Green's function as

$$\Gamma^{\mu s}_{jn}(\mathbf{x}_s,\mathbf{x}'') = Q_{jm}(x'_\perp)\Gamma^\mu_{mn}(\mathbf{x}'_s,\mathbf{x}''), \quad Q_{jm}(x'_\perp) = 1/2(\delta_{jm} + N_j(x'_\perp)\delta_{m3}), \quad (13)$$

it is possible to write (12) as

$$\Gamma^{\mu s}_{jn}(\mathbf{x}'_s,\mathbf{x}'') = \Gamma^0_{jn}(\mathbf{x}'_s,\mathbf{x}'') + \int dx_\perp P_{ji}(\mathbf{x}'_s,\mathbf{x}_s)U_{ip}(x_\perp)\Gamma^{\mu s}_{pn}(\mathbf{x}'_s,\mathbf{x}''), \quad (14)$$

where

$$\Gamma^\mu_{mn}(\mathbf{x}_s,\mathbf{x}'') = U_{mp}(x_\perp)\Gamma^{\mu s}_{pn}(\mathbf{x}_s,\mathbf{x}''),$$

$$Q_{jm}(x_\perp)U_{mp}(x_\perp) = \delta_{jp},$$

and

$$U_{mp}(x_\perp) = 2\delta_{mp} - N_m(x_\perp)\delta_{p3}. \quad (15)$$

Equation (14) is the surface integral equation for the magnetic surface Green's function which we will use.

We finally wish to make two remarks. The first remark is that for a flat surface ($h = 0$), $N_l(x_\perp) = \delta_{l3}$, and it is possible to show that $P_{ji} \sim P_{i3j} \sim \delta_{j3}$ so that the transverse ($j = 1, 2$) components of $\Gamma^{\mu}_{jn}(x'_\perp, \mathbf{x}'') = 2\Gamma^0_{jn}(x'_\perp, \mathbf{x}'')$.

That is, for a flat surface, the transverse components are double the value of the free space Green's function. This is the standard result.[6]

The second remark has to do with an integral representation for the magnetic Green's function $\Gamma^{\mu}_{jm}(\mathbf{x}', \mathbf{x})$ for $z' > h(x'_\perp)$ (i.e., evaluated in the field) in terms of $\Gamma^{\mu s}$. The derivation is similar to the derivation of (12) (but we do not let $\mathbf{x}' \to \mathbf{x}'_s$) and is sketched in the Appendix. The result is

$$\Gamma^{\mu}_{jm}(\mathbf{x}', \mathbf{x}'') = \Gamma^0_{jm}(\mathbf{x}', \mathbf{x}'') + \int dx_\perp P^0_{ji}(\mathbf{x}', \mathbf{x}_s) U_{in}(x_\perp) \times \Gamma^{\mu s}_{nm}(\mathbf{x}_s, \mathbf{x}''), \quad (16)$$

where P^0_{ji} follows from P_{ji} by setting $K^2 = k_3^2$ in R_{ilj} in (A6) and using this result in (A5) and hence in (A4). Setting $K^2 = k_3^2$ is called "going on the energy shell" and was discussed in Refs. 1 and 2. Equation (16) will be useful in the scattering results later.

III. FOURIER TRANSFORMS AND DIAGRAMS

Introducing the Fourier transforms

$$\Gamma^{\mu s}_{jn}(\mathbf{x}'_s, \mathbf{x}'') = (2\pi)^{-6} \iint d\mathbf{k}' d\mathbf{k}'' e^{i\mathbf{k}'\cdot\mathbf{x}'_s} \Gamma^{\mu s}_{jn}(\mathbf{k}', \mathbf{k}'') e^{-i\mathbf{k}''\cdot\mathbf{x}''}, \quad (17)$$

$$\Gamma^0_{jn}(\mathbf{x}'_s, \mathbf{x}'') = (2\pi)^{-6} \iint d\mathbf{k}' d\mathbf{k}'' e^{i\mathbf{k}'\cdot\mathbf{x}'_s} \times (2\pi)^3 \delta(\mathbf{k}' - \mathbf{k}'') \Gamma^0_{jn}(\mathbf{k}') e^{-i\mathbf{k}''\cdot\mathbf{x}''}, \quad (18)$$

as well as (A4) and (A5) into (14), yields the result

$$\iint d\mathbf{k}' d\mathbf{k}'' e^{i\mathbf{k}'\cdot\mathbf{x}'_s} F_{jn}(\mathbf{k}', \mathbf{k}'') e^{-i\mathbf{k}''\cdot\mathbf{x}''} = 0, \quad (19)$$

where F_{jn} is defined by

$$F_{jn}(\mathbf{k}', \mathbf{k}'') = \Gamma^{\mu s}_{jn}(\mathbf{k}', \mathbf{k}'') - (2\pi)^3 \delta(\mathbf{k}' - \mathbf{k}'') \Gamma^0_{jn}(\mathbf{k}') - \int d\mathbf{k} L_{jp}(\mathbf{k}', \mathbf{k}) \Gamma^{\mu s}_{pn}(\mathbf{k}, \mathbf{k}'') \quad (20)$$

and

$$L_{jp}(\mathbf{k}', \mathbf{k}) = (2\pi)^{-3} G^0(k') R_{ilj}(\mathbf{k}') \int dx_\perp N_l(x_\perp) U_{ip}(x_\perp) e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{x}_s}. \quad (21)$$

If, in (19), $\mathbf{x}'_s$ were replaced by $\mathbf{x}'$, i.e., if we were introducing the corresponding Fourier transforms in (16) for the magnetic Green's function in the *field* rather than in (14) for the magnetic Green's function on the *surface*, then the only solution of (19) would be

$$F_{jn}(\mathbf{k}', \mathbf{k}'') \equiv 0. \quad (22)$$

Since ultimately we will be interested solely in field calculations, we choose the "gauge" condition (22) to hold here. Further, integrating (21) by parts and neglecting the resulting surface terms (see Ref. 1) permits us to factor L_{jp} into propagator-vertex-interaction form

$$L_{jp}(\mathbf{k}', \mathbf{k}) = G^0(k') V_{jp}(\mathbf{k}', \mathbf{k}) A(\mathbf{k}' - \mathbf{k}), \quad (23)$$

where $G^0(k)$ is the scalar free space Green's function in k space (propagator) which also was our propagator in the scalar case, V_{jp} is the (matrix) vertex term given by

$$V_{jp}(\mathbf{k}, \mathbf{k}) = \frac{R_{ilj}(\mathbf{k}')}{(2\pi)^3} \frac{k'_l - k_l}{k'_3 - k_3} \left(2\delta_{ip} - \delta_{p3} \frac{k'_i - k_i}{k'_3 - k_3} \right), \quad (24)$$

where R_{ilj} is defined by (A6), and $A(\mathbf{k})$ is the interaction term

$$A(\mathbf{k}) = \int d\mathbf{x} e^{-ik_\perp \cdot x_\perp} e^{-ik_3 h(x_\perp)} \quad (25)$$

which appeared in both scalar[1] and elastic[2] cases. Using (22) and (23) in (20) yields

$$\Gamma^{\mu s}_{jn}(\mathbf{k}', \mathbf{k}'') = (2\pi)^3 \delta(\mathbf{k}' - \mathbf{k}'') \Gamma^0_{jn}(\mathbf{k}') + G^0(k') \times \int V_{jp}(\mathbf{k}', \mathbf{k}) A(\mathbf{k}' - \mathbf{k}) \Gamma^{\mu s}_{pn}(\mathbf{k}, \mathbf{k}'') d\mathbf{k}. \quad (26)$$

Defining away the delta function term by introducing the auxiliary Green's function G^s_{jl} as

$$\Gamma^{\mu s}_{jn}(\mathbf{k}', \mathbf{k}'') = (2\pi)^3 \delta(\mathbf{k}' - \mathbf{k}'') \Gamma^0_{jn}(\mathbf{k}) + (2\pi)^3 G^0(k') G^s_{jl}(\mathbf{k}', \mathbf{k}'') \Gamma^0_{ln}(\mathbf{k}'') \quad (27)$$

and substituting (27) into (26) yields an integral equation for G^s_{jl}

$$G^s_{jl}(\mathbf{k}', \mathbf{k}'') = V_{jl}(\mathbf{k}', \mathbf{k}'') A(\mathbf{k}' - \mathbf{k}'') + \int V_{jp}(\mathbf{k}', \mathbf{k}) A(\mathbf{k}' - \mathbf{k}) G^0(k) G^s_{pl}(\mathbf{k}, \mathbf{k}'') d\mathbf{k} \quad (28)$$

whose diagram rules are listed in Fig. 2. The integral equation using diagrams is shown in Fig. 3. An analogous procedure can be developed for the magnetic Green's function in the field. Defining an additional auxiliary Green's function (without superscript) G_{jl} as

$$\Gamma^{\mu}_{jn}(\mathbf{k}', \mathbf{k}'') = (2\pi)^3 \delta(\mathbf{k}' - \mathbf{k}'') \Gamma^0_{jn}(\mathbf{k}') + (2\pi)^3 G^0(k') G_{jl}(\mathbf{k}', \mathbf{k}'') \Gamma^0_{ln}(\mathbf{k}''), \quad (29)$$

we can write the integral relation

$$G_{jl}(\mathbf{k}', \mathbf{k}'') = V^0_{jl}(\mathbf{k}', \mathbf{k}'') A(\mathbf{k}' - \mathbf{k}'') + \int V^0_{jp}(\mathbf{k}', \mathbf{k}) A(\mathbf{k}' - \mathbf{k}) G^0(k) G^s_{pl}(\mathbf{k}, \mathbf{k}'') d\mathbf{k} \quad (30)$$

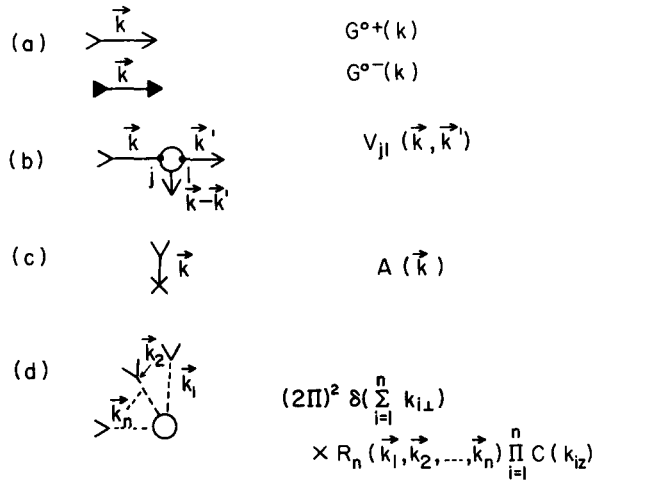

FIG. 2. Diagram rules associated with the auxiliary Green's function G^s_{jl}. The propagators (a) and vertex (b) are used for both deterministic and stochastic rules whereas the single deterministic interaction term (c) generalizes for multiple random interactions and with a cluster decomposition to the purely stochastic rule (d). When the rules are combined, internal momenta are integrated over, and repeated indices summed.

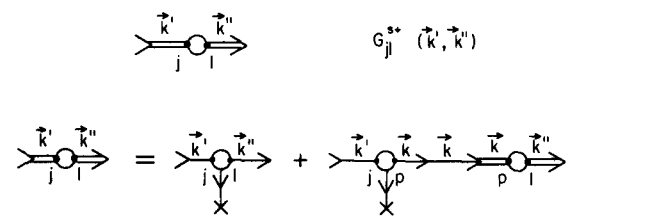

FIG. 3. Diagram rule for G^{s+}_{jl} itself, (a), and the integral equation (28) in diagram form, (b).

with the on-shell vertex V^0_{jp} similar to (24)

$$V^0_{jp}(\mathbf{k}',\mathbf{k}) = \frac{R^0_{ilj}(\mathbf{k}')}{(2\pi)^3}\frac{k'_l - k_l}{k'_3 - k_3}\left(2\delta_{ip} - \delta_{p3}\frac{k'_i - k_i}{k'_3 - k_3}\right) \tag{31}$$

where R^0_{ilj} is defined by (A11). Clearly, on shell,

$$G^s_{jl}(\mathbf{k}',\mathbf{k}'') = G_{jl}(\mathbf{k}',\mathbf{k}''). \tag{32}$$

Now (29) expresses the Green's function in the field Γ^μ in terms of "incident" or free space Green's function Γ^0. In coordinate space, and specifying the outgoing (+) wave boundary condition, (29) is

$$\Gamma^{\mu+}_{jn}(\mathbf{x}',\mathbf{x}'') = \Gamma^{0+}_{jn}(\mathbf{x}',\mathbf{x}'') + (2\pi)^3 \iint d\mathbf{x}_1 d\mathbf{x}_2 G^{0+}(\mathbf{x}'-\mathbf{x}_1) \times G^+_{jl}(\mathbf{x}_1,\mathbf{x}_2)\Gamma^{0+}_{ln}(\mathbf{x}_2,\mathbf{x}'') \tag{33}$$

and Γ^μ is thus written as an incident Green's function Γ^0 plus a scattered Green's function arising from an integral transformation. The outgoing scattered *field* $\psi^{(0)}_j$ can be written in terms of the incident *field* $\psi^{(i)}_j$ using this same integral transformation.

The result is

$$\psi^{(0)}_j(\mathbf{x}') = (2\pi)^3 \iint d\mathbf{x}_1 d\mathbf{x}_2 G^{0+}(\mathbf{x}'-\mathbf{x}_1) G^+_{jl}(\mathbf{x}_1,\mathbf{x}_2)\psi^{(i)}_l(\mathbf{x}_2). \tag{34}$$

In (34) use the two-dimensional representation

$$G^{0+}(\mathbf{x}) = \frac{\pi i}{(2\pi)^3}\int dk_\perp \frac{e^{ik_\perp\cdot x_\perp}}{K} e^{iK|z|} \quad (K = \sqrt{k_0^2 - k_\perp^2}) \tag{35}$$

and expand the outgoing field in terms of plane waves $\phi^{(0)}_j$

$$\psi^{(0)}_j(\mathbf{x}) = \int dk_\perp e^{i\mathbf{k}\cdot\mathbf{x}}\phi^{(0)}_j(k_\perp) \quad (k_z = +K). \tag{36}$$

Here the specification $k_z = +K$ is to insure that the field is outgoing. Then it is possible to write in the far field

$$\phi^{(0)}_j(k_\perp) = \frac{\pi i}{K}\iint d\mathbf{x}_1 d\mathbf{x}_2 e^{-i\mathbf{k}\cdot\mathbf{x}_1} G^+_{jl}(\mathbf{x}_1,\mathbf{x}_2)\psi^{(i)}_l(\mathbf{x}_2) \quad (k_z = +K). \tag{37}$$

Similarly, expanding the incident field in plane waves $\phi^{(i)}_l$

$$\psi^{(i)}_l(\mathbf{x}) = \int dk'_\perp e^{i\mathbf{k}'\cdot\mathbf{x}}\phi^{(i)}_l(k'_\perp) \quad k'_z = -K' \tag{38}$$

and inserting (38) in (37), yields

$$\phi^{(0)}_j(k_\perp) = \int dk'_\perp T^+_{jl}(k_\perp,k'_\perp)\phi^{(i)}_l(k'_\perp) \tag{39}$$

with

$$T^\pm_{jl}(k_\perp,k'_\perp) = \frac{\pi i}{k_z}\iint d\mathbf{x}_1 d\mathbf{x}_2 e^{-i\mathbf{k}\cdot\mathbf{x}_1} G^\pm_{jl}(\mathbf{x}_1,\mathbf{x}_2)e^{i\mathbf{k}'\cdot\mathbf{x}_2} \quad (k'_z = \pm K, k'_z = \mp K'), \tag{40}$$

where we have included the incoming (−) boundary value and its on-shell value. The latter can be derived in a similar manner as T^+. The T^-_{jl} matrix is needed for the complex conjugate fields below. The T^+_{jl} matrix is analogous to the linear connective introduced by Saxon[7]

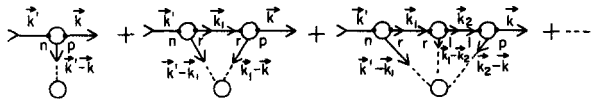

FIG. 4. Mass operator $M_{n\rho}(\mathbf{k}',\mathbf{k})$ used in the Dyson equation.

who also derived some general properties of it. The right-hand side of (40) is just the Fourier transform of G_{jl} and since we are on shell we can write

$$T^\pm_{jl}(k_\perp,k'_\perp) = \frac{\pi i}{k_z}G^\pm_{jl}(\mathbf{k},\mathbf{k}') = \frac{\pi i}{k_z}G^s_{jl}(\mathbf{k},\mathbf{k}') \quad (k_z = \pm K, k'_z = \mp K'). \tag{41}$$

Combining (36) and (39) we can write the outgoing field in terms of incident plane waves as

$$\psi^{(0)}_j(\mathbf{x}) = \iint dk_\perp dk'_\perp e^{i\mathbf{k}\cdot\mathbf{x}}T^+_{jl}(k_\perp,k'_\perp)\phi^{(i)}_l(k'_\perp) \quad (k_z = K) \tag{42}$$

A similar discussion can be given for the complex conjugate fields which can be expressed as

$$\psi^{(0)*}_j(\mathbf{x}) = \int dk_\perp e^{-i\mathbf{k}\cdot\mathbf{x}}\phi^{(0)*}_j(k_\perp) \quad (k_z = +K), \tag{43}$$

$$\psi^{(i)*}_l(\mathbf{x}) = \int dk'_\perp e^{-i\mathbf{k}'\cdot\mathbf{x}}\phi^{(i)*}_l(k'_\perp) \quad (k'_z = -K'), \tag{44}$$

$$\phi^{(0)*}_j(k_\perp) = \int dk'_\perp T^-_{jl}(-k_\perp,-k'_\perp)\phi^{(i)*}_l(k'_\perp), \tag{45}$$

$$\psi^{(0)*}_j(\mathbf{x}) = \iint dk_\perp dk'_\perp e^{-i\mathbf{k}\cdot\mathbf{x}}T^-_{jl}(-k_\perp,-k'_\perp)\phi^{(i)*}_l(k'_\perp) \quad (k_z = K). \tag{46}$$

These latter equations are used later in the discussion of the intensity.

IV. RANDOM ROUGH SURFACE

We have presented the deterministic results in the previous sections in some detail. The purely statistical aspects of the problem however have been thoroughly discussed in Ref. 1. A brief discussion of the statistical aspects, suitable as an introduction to this section, was presented in an introduction to Sec. IV of Ref. 2, and we refer to this latter discussion rather than reproduce it here. In fact, the entire discussion of this section closely parallels that of Sec. IV of Ref. 2. The only additional diagram notation we need is shown in Fig. 2d, where some of the R_n functions are explicitly listed in Ref. 1 (see Ref. 2 for the misprint in R_2). Here $C(k_z) = \exp[-\frac{1}{2}\Gamma(0)k_z^2]$is the characteristic function and $\Gamma(x)$ the two point correlation function. Using partial summation techniques,[3] the mean of G^s_{jl}, $\langle G^s_{jl}\rangle$ can be written as

$$\langle G^s_{jl}(\mathbf{k}',\mathbf{k}'')\rangle = M_{jl}(\mathbf{k}',\mathbf{k}'') + \int M_{jn}(\mathbf{k}',\mathbf{k})G^0(k) \times \langle G^s_{nl}(\mathbf{k},\mathbf{k}'')\rangle d\mathbf{k}, \tag{47}$$

where the sum of connected diagrams or "mass operator" M_{jl} is shown in Fig. 4. Note the formal similarity of Fig. 4 here with Fig. 3 of Ref. 2. Equation (47) is the Dyson equation of our problem. A transverse delta function can be factored out of M_{jl} and hence out of $\langle G^s_{jl}\rangle$ as

$$M_{jl}(\mathbf{k}',\mathbf{k}'') = \delta(k'_\perp - k''_\perp)\overline{M}_{jl}(k'_z,k''_z), \tag{48}$$

$$\langle G^s_{jl}(\mathbf{k}',\mathbf{k}'')\rangle = \delta(k'_\perp - k''_\perp)g^s_{jl}(k'_z,k''_z), \tag{49}$$

where the transverse momentum dependence of $\overline{M}_{jl}$ and g^s_{jl} has been suppressed. Substituting (48) and (49) in (47) yields

$$g^s_{jl}(k'_z,k''_z) = \overline{M}_{jl}(k'_z,k''_z) + \int \overline{M}_{jn}(k'_z,k_z)G^0(k) \times g^s_{nl}(k_z,k''_z)dk_z. \tag{50}$$

This is called the reduced Dyson equation. The full Dyson equation, i.e., the Dyson equation for $\langle\Gamma^{\mu s}_{jl}\rangle$, can be written as

$$\langle \Gamma^{\mu s}_{jl}(\mathbf{k}', \mathbf{k}'')\rangle = (2\pi)^3\delta(\mathbf{k}' - \mathbf{k}'')\Gamma^0_{jl}(\mathbf{k}') + G^0(k') \int d\mathbf{k} M_{jn}(\mathbf{k}', \mathbf{k})\langle \Gamma^{\mu s}_{nl}(\mathbf{k}, \mathbf{k}'')\rangle. \quad (51)$$

As a simple example, choose for M_{jl} the first term in the expansion in Fig. 4. Then, using (48), (29), and the diagram rules, we can write

$$\overline{M}_{jl}(k'_z, k''_z) = V^{(1)}_{jl}(\mathbf{k}')C(k'_z - k''_z) \quad (52)$$

with

$$V^{(1)}_{jp}(\mathbf{k}') \equiv (2\pi)^2 V_{jp}(\mathbf{k}', \mathbf{k}'')\,|_{k'_\perp = k''_\perp} = \frac{i}{\pi}\left\{\delta_{j3}\left[k'_{p\perp} + \delta_{p3}P\left(\frac{K'^2}{k'_3}\right)\right] - \delta_{jp}P\left(\frac{K'^2}{k'_3}\right)\right\}, \quad (53)$$

so that (50) becomes

$$g^s_{jl}(k_z, k''_z) = C(k'_z - k''_z)V^{(1)}_{jl}(\mathbf{k}') + V^{(1)}_{jn}(\mathbf{k}') \int C(k'_z - k_z)G^0(k)g^s_{nl}(k_z, k''_z)dk_z \quad (54)$$

which is similar to Equation (42) in Ref. 1 and Equation (4. 14) of Ref. 2. (Also see footnote 9 of Ref. 2).

In order to calculate the second moments, use the mutual coherence function C_{mn} given by

$$C_{mn}(\mathbf{x}, \mathbf{x}') = \langle \psi^{(0)}_m(\mathbf{x})\psi^{(0)*}_n(\mathbf{x}')\rangle. \quad (55)$$

Substituting (36) and (43) in (55) yields ($k_z = K$, $k'_z = K'$)

$$C_{mn}(\mathbf{x}, \mathbf{x}') = \iint dk_\perp dk'_\perp e^{i[\mathbf{k}\cdot\mathbf{x} - \mathbf{k}'\cdot\mathbf{x}']}\langle \phi^{(0)}_m(k_\perp)\phi^{(0)*}_n(k'_\perp)\rangle \quad (56)$$

and, using (39), (41), and (45), we have

$$\langle \phi^{(0)}_m(k_\perp)\phi^{(0)*}_n(k'_\perp)\rangle = \iint dk''_\perp dk'''_\perp \langle T^+_{ml}(k_\perp, k''_\perp)T^-_{np}(-k'_\perp, -k'''_\perp)\rangle \phi^{(i)}_l(k''_\perp)\phi^{(i)*}_p(k'''_\perp) = \iint dk''_\perp dk'''_\perp \frac{\pi^2}{KK'}\langle G^{s+}_{ml}(\mathbf{k}, \mathbf{k}'')G^{s-}_{np}(-\mathbf{k}, -\mathbf{k}''')\rangle \phi^{(i)}_l(k''_1)\phi^{(i)*}_p(k'''_1), \quad (57)$$

where $k_z = K$, $k''_z = -K''$, $k'_z = K'$, and $k'''_z = -K'''$.

Again using partial summation techniques, the reduced Bethe–Salpeter equation is

$$\begin{aligned}\langle G^{s+}_{ml}(\mathbf{k}, \mathbf{k}_1)G^{s-}_{np}(\mathbf{k}', \mathbf{k}'_1)\rangle &= \langle G^{s+}_{ml}(\mathbf{k}, \mathbf{k}_1)\rangle\langle G^{s-}_{np}(\mathbf{k}', \mathbf{k}'_1)\rangle + K_{ml,np}(\mathbf{k}, \mathbf{k}'\,|\,\mathbf{k}_1, \mathbf{k}'_1)\\ &+ \int d\mathbf{k}_2 K_{mr,np}(\mathbf{k}, \mathbf{k}'\,|\,\mathbf{k}_2, \mathbf{k}'_1)G^{0+}(k_2)\langle G^{s+}_{rl}(\mathbf{k}_2, \mathbf{k}_1)\rangle\\ &+ \int d\mathbf{k}'_2 K_{ml,nq}(\mathbf{k}, \mathbf{k}'\,|\,\mathbf{k}_1, \mathbf{k}'_2)G^{0-}(k'_2)\langle G^{s-}_{qp}(\mathbf{k}'_2, \mathbf{k}'_1)\rangle\\ &+ \iint d\mathbf{k}_2 d\mathbf{k}'_2 K_{mr,nq}(\mathbf{k}, \mathbf{k}'\,|\,\mathbf{k}_2, \mathbf{k}'_2)\\ &\times G^{0+}(k_2)G^{0-}(k'_2)\langle G^{s+}_{rl}(\mathbf{k}_2, \mathbf{k}_1)G^{s-}_{qp}(\mathbf{k}'_2, \mathbf{k}'_1)\rangle, \end{aligned} \quad (58)$$

where the function $K_{ml,np}$, corresponding to the intensity operator of random volume scattering theory,[3] is given by the sum of connected diagrams in Fig. 5. The full Bethe–Salpeter equation can be written using (27) and (58) as

$$\begin{aligned}\langle \Gamma^{s+}_{mn}(\mathbf{k}, \mathbf{k}_1)\Gamma^{s-}_{pq}(\mathbf{k}', \mathbf{k}_1)\rangle &= \langle \Gamma^{s+}_{mn}(\mathbf{k}, \mathbf{k}_1)\rangle\langle \Gamma^{s-}_{pq}(\mathbf{k}', \mathbf{k}'_1)\rangle + G^{0+}(k)G^{0-}(k')\\ &\times \iint d\mathbf{k}_2 d\mathbf{k}'_2 K_{mr,pl}(\mathbf{k}, \mathbf{k}'\,|\,\mathbf{k}_2, \mathbf{k}'_2)\langle \Gamma^{s+}_{rn}(\mathbf{k}_2, \mathbf{k}_1)\Gamma^{s-}_{lq}(\mathbf{k}'_2, \mathbf{k}'_1)\rangle. \end{aligned} \quad (59)$$

As previously mentioned,[1,2] it is possible to reduce the dimensionality of the integral equations by using the translational invariance of $K_{ml,nq}$ so that (58) and (59) are four-dimensional integral equations. It is simpler however to discuss some examples, and that is how we proceed.

If the incident field is a plane wave

$$\phi^{(i)}_l(k_\perp) = \delta(k_\perp - k_{i\perp})\chi_l, \quad (60)$$

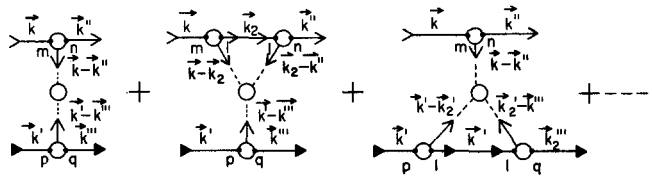

FIG. 5. Intensity operator $K_{mn,pq}(\mathbf{k}, \mathbf{k}'\,|\,\mathbf{k}'', \mathbf{k}''')$ used in the Bethe–Salpeter equation.

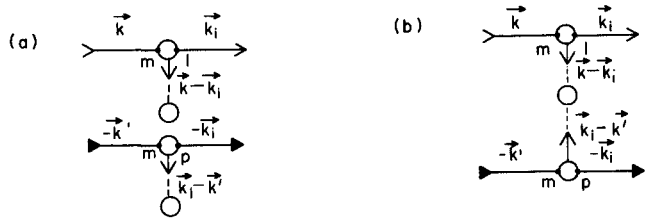

FIG. 6. Lowest order coherent (a) and incoherent (b) contributions to the intensity.

where χ_l indicates the direction of the wave (polarization factor), then (57) yields a definition of intensity I if we substitute (60) into (57), multiply by δ_{mn}, and sum to yield

$$\delta(k_\perp - k'_\perp)I(k_\perp, k_{i\perp}) = \frac{\pi^2}{KK'}\langle G^{s+}_{ml}(\mathbf{k}, \mathbf{k}_i)G^{s-}_{mp}(-\mathbf{k}', -\mathbf{k}_i)\rangle \chi_l\chi^*_p. \quad (61)$$

Here $I(k_\perp, k_{i\perp})$ is the intensity in the $k_\perp$ direction due to an incident plane wave in the $k_{i\perp}$ direction, $k_z = K$, $k'_z = K'$ and $k_{iz} = -K_i$.

The lowest order coherent contribution (lc) to the intensity is shown in Fig. 6a, with the additional on-shell conditions above. Using the diagram rules we get

$$I^{(lc)}(k_\perp, k_{i\perp}) = \delta(k_\perp - k_{i\perp})|f_m(k_{i\perp})|^2, \quad (62)$$

where the delta function indicates specular scattering and where the scattering function f_m is defined by

$$f_m(k_{i\perp}) = \frac{C(2K_i)}{2K_i}\left[(k_{qi\perp} + K_i\delta_{q3})\delta_{m3} - K_i\delta_{qm}\right](2\chi_q - \delta_{q3}\chi_3). \quad (63)$$

Some further simplifications of these results are possible. For a horizontally (H) polarized incident field ($\chi_1 = \chi_3 = 0$, $\chi_2 = 1$), we get ($k_{i2} = 0$ and $k_{i\perp} = k_{i1} = k_0\sin\theta_i$, θ_i the incident angle)

$$|f^H_m(k_{i\perp})|^2 = C^2(2K_i) = \exp[-4k_0^2\Gamma(0)\cos^2\theta_i]$$

which is the standard result,[12] and for a vertically (V) polarized incident field ($\chi_2 = 0$, $\chi_1 = \cos\theta_i$, $\chi_3 = \sin\theta_i$), we get

$$|f^V_m(k_{i\perp})|^2 = [(K_i^2 + k_{i1}^2)/K_i^2]C^2(2K_i)\cos^2\theta_i = C^2(2K_i)$$

which is the same as the horizontal polarization for this lowest order term. Note that for $\Gamma(0) = \sigma^2 \to 0$, both $|f^H_m|$ and $|f^V_m| \to 1$ which is the usual result for flat surface scattering.

As a final example, the lowest order incoherent (li) contribution to the intensity is shown in Fig. 6b. Using the diagram rules and the on-shell conditions $k_z = K = k'_z$ and $k_{iz} = -K_i$, the result is

$$I^{(li)}(k_\perp, k_{i\perp}) = R_2(k_\perp - k_{i\perp}, K + K_i, -K - K_i)\,|\,g_m(k_\perp, k_{i\perp})|^2,$$

where

$$g_m(k_\perp, k_{i\perp}) = \frac{2\pi^2}{K}V^0_{ml}(\mathbf{k}, \mathbf{k}_i)\chi_l C(K + K_i)$$

with V^0_{ml} given by (31) and the function R_2 given by[1,13]

$$R_2(\mathbf{k}_1, \mathbf{k}_2) = \int dy_\perp e^{-ik_{1\perp}\cdot y_\perp}(e^{-\Gamma(y_\perp)k_{1\perp}k_{2z}} - 1)$$

and the remark that

$$R_2(\mathbf{k}_1, \mathbf{k}_2) = R_2(\mathbf{k}_1, k_{2z}) = R_2(k_{1\perp}, k_{1z}, k_{2z})$$

and $k_z - k_{iz} = K + K_i$.

V. SUMMARY AND CONCLUSIONS

The diagram expansion method of calculating the stochastic Green's function for scattering from a random rough surface has been extended to Electromagnetic problems. The particular problem here concerns a Gaussian distributed surface bounding a semi-infinite perfectly conducting medium. The methods we discussed were mainly in keeping with a systematic and formal presentation of the problem although examples of lowest order coherent and incoherent contributions to the intensity were discussed and an approximation for the Dyson equation was presented. Using the diagram rules, higher order corrections are trivial in principle to write down, although in practice many difficulties remain as far as approximation methods and computational difficulties are concerned.

Finally, it was previously remarked in this paper that the diagram notation in the scalar[1] and elastic[2] cases was similar to that for the electromagnetic problem here. The functional forms for the propagators, vertices, and interaction for the three cases are summarized in Table I. Generally, the interaction term is the same for all three cases, the vertex terms are considerably different in each case, and the propagators are the same in the scalar and electromagnetic cases. The elastic propagator is a vector with components $G_0^{t,l}$ having the functional form of the free space Green's function G_0, but with the free space wavenumber replaced by the transverse (t) and longitudinal (l) wavenumber, respectively.

ACKNOWLEDGMENTS

I am grateful to Dr. George Georges for several conversations on this problem and to Mrs. Jean Krause for typing the manuscript.

APPENDIX: KERNELS OF THE INTEGRAL EQUATIONS

In this appendix we calculate the kernel functions which arose in the integral equations of Sec. II. Using (7), (9), and (10), Eq. (8) can be written as

$$\Gamma^{\mu}_{jn}(\mathbf{x}', \mathbf{x}'')\Theta(z' - h(x'_\perp)) = \Gamma^0_{jn}(\mathbf{x}', \mathbf{x}'') + \int dx_\perp N_l(x_\perp)(\delta_{lj}\partial'_i - \delta_{ij}\partial'_l)G^0(\mathbf{x}', \mathbf{x}_s) \times \Gamma^{\mu}_{in}(\mathbf{x}_s, \mathbf{x}''). \quad (A1)$$

From Ref. 1, Appendix A, we can write

$$\partial'_m G^0(\mathbf{x}', \mathbf{x}_s) = \frac{i}{(2\pi)^3}\int d\mathbf{k} G^0(k) e^{i\mathbf{k}\cdot[\mathbf{x}'-\mathbf{x}_s]}\left[k_{m\perp} + \delta_{m3}P\left(\frac{K^2}{k^3}\right)\right] - \tfrac{1}{2}\delta_{m3}\epsilon(z' - h(x'_\perp))\delta(x_\perp - x_\perp), \quad (A2)$$

where the symbol P stands for the Cauchy principle value, and $\epsilon(z) = \theta(z) - \theta(-z)$. Hence we can write

TABLE I. Functional forms for the diagram pieces in the scalar (Ref. 1), electromagnetic (here), and elastic (Ref. 2) cases.

	Propagator	Vertex	Interaction
Scalar	$G_0(k)$	$V(\mathbf{k}, \mathbf{k}')$	$A(\mathbf{k})$
Electro-Magnetic	$G_0(k)$	$V_{mn}(\mathbf{k}, \mathbf{k}')$	$A(\mathbf{k})$
Elastic	$P_n(k) = (G_0^t(k), G_0^l(k), 1)$	$V_{mnp}(\mathbf{k}, \mathbf{k}')$	$A(\mathbf{k})$

$$N_l(x_\perp)(\delta_{lj}\partial'_i - \delta_{ij}\partial'_l)G^0(\mathbf{x}', \mathbf{x}_s) = P_{ji}(\mathbf{x}', \mathbf{x}_s) + \tfrac{1}{2}\epsilon(z' - h(x'_\perp))\delta(x'_\perp - x_\perp)(\delta_{ji} - N_j(x'_\perp)\delta_{i3}), \quad (A3)$$

where

$$P_{ji}(\mathbf{x}', \mathbf{x}_s) = N_l(x_\perp)P_{ilj}(\mathbf{x}' - \mathbf{x}_s), \quad (A4)$$

$$P_{ilj}(\mathbf{x}) = (2\pi)^{-3}\int d\mathbf{k} G^0(k) e^{i\mathbf{k}\cdot\mathbf{x}} R_{ilj}(\mathbf{k}), \quad (A5)$$

and

$$R_{ilj}(\mathbf{k}) = i\left\{\left[k_{i\perp} + \delta_{i3}P\left(\frac{K^2}{k_3}\right)\right]\delta_{lj} - \left[k_{l\perp} + \delta_{l3}P\left(\frac{K^2}{k_3}\right)\right]\delta_{ij}\right\}. \quad (A6)$$

If we substitute (A3) into (A1) and let $x' \to x'_s$ through positive z' values ($\epsilon = 1$) we get the result shown in Eq. (12) of Sec. II.

Now the singularity in (A2) arose because the point z' was to be evaluated on the surface $h(x'_\perp)$. If we wish to calculate the Green's dyadic in the field $[z' > h(x'_\perp)]$ this is not the case. No singularity arises. Hence we do not need to go through the procedure of Appendix A of Ref. 1. We can write the Fourier representation of G^0

$$G^0(\mathbf{x}', \mathbf{x}_s) = (2\pi)^{-3}\int d\mathbf{k} G^0(k) e^{i\mathbf{k}\cdot(\mathbf{x}'-\mathbf{x}_s)}$$

and differentiate it directly to get

$$\partial'_m G^0(\mathbf{x}', \mathbf{x}_s) = i(2\pi)^{-3}\int d\mathbf{k} G^0(k) k_m e^{i\mathbf{k}\cdot(\mathbf{x}'-\mathbf{x}_s)}. \quad (A7)$$

We can get (A7) from (A2) if we drop the singular (delta function) term and set $K^2 = k_3^2$ in (A2). The latter is called "going on the energy shell" and was discussed previously.[1,2] Hence, by direct differentiation or by going on the energy shell we can write for $z' > h(x'_\perp)$

$$N_l(x_\perp)(\delta_{lj}\partial'_i - \delta_{ij}\partial'_l)G^0(\mathbf{x}', \mathbf{x}_s) = P^0_{ji}(\mathbf{x}', \mathbf{x}_s), \quad (A8)$$

where

$$P^0_{ji}(\mathbf{x}', \mathbf{x}_s) = N_l(x_\perp)P^0_{ilj}(\mathbf{x}' - \mathbf{x}_s), \quad (A9)$$

$$P^0_{ilj}(\mathbf{x}) = (2\pi)^{-3}\int d\mathbf{k} G^0(k) e^{i\mathbf{k}\cdot\mathbf{x}} R^0_{ilj}(\mathbf{k}), \quad (A10)$$

$$R^0_{ilj}(\mathbf{k}) = i(k_i\delta_{lj} - k_l\delta_{ij}). \quad (A11)$$

Substituting (A8) and (15) into (A1) yields the integral relation (16).

[1] G. G. Zipfel and J. A. DeSanto, J. Math. Phys. 13, 1903 (1972).
[2] J. A. DeSanto, J. Math. Phys. **14**, 1566 (1973).
[3] U. Frisch, "Wave Propagation in Random Media", in *Probabilistic Methods in Applied Mathematics*, edited by A. T. Bharucha-Reid (Academic, New York, 1968).
[4] P. M. Morse and H. Feshback, *Methods of Theoretical Physics*, (McGraw-Hill, New York, 1953), Chap. 13. (See the corrections of this approach in Ref. 9).
[5] W. Hauser, *Introduction to the Principles of Electromagnetism*, (Addison-Wesley, Reading Mass., 1971), Chap. 17.
[6] H. Levine and J. Schwinger, Comm. Pure Appl. Math. 3, 361 (1950) which is reproduced in *The Theory of Electromagnetic Waves*, edited by M. Kline (Dover, New York, 1965).
[7] D. S. Saxon, Phys. Rev. **100**, 1771 (1955).
[8] V. Twersky, J Math. Phys. 8, 589 (1967).
[9] K. M. Mitzner, J. Math. Phys. 5, 1776 (1964).
[10] P. Beckmann and A. Spizzichino, *The Scattering of Electromagnetic Waves from Rough Surfaces* (Macmillan, New York, 1963)
[11] C. T. Tai, IEEE Trans. **AP-20**, 114 (1972).
[12] Reference 10, p. 81, Eq. (6).
[13] Reference 2, footnote 9.

An initial value method for an inverse problem in wave propagation

R. Kalaba

Department of Economics, University of Southern California, Los Angeles, California 90007

E. Zagustin

Department of Civil Engineering, California State University, Long Beach, California 90801

(Received 13 September 1973)

Kay and Balanis have reduced various inverse problems in wave propagation to the solution of Fredholm integral equations. These integral equations can be further reduced to Cauchy systems.

1. INTRODUCTION

In two basic papers Kay[1] and Balanis[2] have shown that certain inverse problems in wave propagation reduce to solving the Fredholm integral equation

$$0 = R(x + t) + U(x, t) + \int_{-x}^{+x} R(t + y)\, U(x, y)dy, \qquad -x \leq t \leq x, \quad 0 \leq x,$$

where the function R is an observed reflected wave, and the function U is to be determined. More precisely, it is the function

$$V(x) = \frac{d}{dx} U(x, x), \qquad 0 \leq x,$$

that we wish to find, for it is directly related to the density of the medium. In this paper we shall show how this may be done directly, i.e., without solving the integral equation for all values of t in the interval $[-x, +x]$. Other inverse problems involving the fitting of differential equations to experimental data are considered in the book by Bellman and Kalaba.[3]

2. DERIVATION OF THE INITIAL VALUE METHOD FOR THE INVERSE PROBLEM IN WAVE PROPAGATION

Consider the family of Fredholm integral equations

$$0 = R(x + t) + U(x, t) + \int_{-x}^{x} R(t + y)U(x, y)dy, \qquad -x \leq t \leq x, \quad 0 \leq x, \tag{1}$$

for the unknown function U. The function R has the form

$$R(t + y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iz(t+y)} r(z)dz, \tag{2}$$

where the function r is known, and we desire to determine the physically meaningful function V given by

$$V(x) = \frac{d}{dx} U(x, x), \qquad x > 0. \tag{3}$$

We shall show how this may be done without actually solving the integral equations in Eq. (1) for all $x > 0$.

We introduce the auxiliary function J,

$$J = J(t, x, z), \qquad -x \leq t \leq x, -\infty < z < \infty, \tag{4}$$

to be the solution of the integral equation

$$0 = e^{-izt} + J(t, x, z) + \int_{-x}^{x} R(t + y)J(y, x, z)dy. \tag{5}$$

In view of Eqs. (1) and (2) we see that the function U may be expressed in terms of the function J as

$$U(x, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} J(t, x, z)e^{-izx} r(z)dz. \tag{6}$$

Now we consider determining the auxiliary function J as the solution of a Cauchy system. We differentiate both sides of Eq. (5) to obtain the relation

$$0 = J_x(t, x, z) + R(t + x)J(x, x, z) - R(t - x)J(-x, x, z) + \int_{-x}^{x} R(t + y)J_x(y, x, z)dy, \tag{7}$$

which is viewed as an integral equation for the unknown function J_x. Its solution is

$$J_x(t, x, z) = \frac{J(x, x, z)}{2\pi} \int_{-\infty}^{\infty} J(t, x, z')e^{-iz'x} r(z')dz' - \frac{J(-x, x, z)}{2\pi} \int_{-\infty}^{\infty} J(t, x, z')e^{iz'x} r(z')dz'. \tag{8}$$

Now let us rewrite the integral equation in Eq. (5) in the form

$$0 = e^{-izt} + J(t, x, z) + \int_{-x}^{x} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iz'(t+y)} r(z')dz' J(y, x, z)dy. \tag{9}$$

From this it follows that

$$J(t, x, z) = -e^{-izt} - \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iz't} \rho(z', z, x) r(z')dz', \tag{10}$$

where we have introduced the basic function ρ through the definition

$$\rho(v, z, x) = \int_{-x}^{x} e^{-ivy} J(y, x, z)dy, \qquad -\infty < v, z < \infty, \quad 0 \leq x. \tag{11}$$

In particular, it follows that $J(x, x, z)$ and $J(-x, x, z)$ can be expressed in terms of ρ as

$$J(x, x, z) = -e^{-izx} - \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iz'x} \rho(z', z, x) r(z')dz' \tag{12}$$

and

$$J(-x, x, z) = -e^{+izx} - \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{iz'x} \rho(z', z, x) r(z')dz'. \tag{13}$$

Thus the problem of finding J, and hence U, is reduced to that of finding the function ρ. We shall now obtain the Cauchy system which the function ρ satisfies.

We differentiate both sides of Eq. (11) to obtain the equality

$$\rho_x(v, z, x) = e^{-ivx} J(x, x, z) - e^{+ivx} J(-x, x, z) + \int_{-x}^{x} e^{-ivy} J_x(y, x, z)dy. \tag{14}$$

We can evaluate the integral in the last equation by using Eq. (8). The result is

$$\int_{-\infty}^{\infty} e^{-ivy} J_x(y,x,z)dy = \int_{-x}^{x} e^{-ivy}\left(\frac{J(x,x,z)}{2\pi}\int_{-\infty}^{\infty} J(y,x,u)e^{-iux}r(u)du - \frac{J(-x,x,z)}{2\pi}\int_{-\infty}^{\infty} J(y,x,u)e^{iux}r(u)du\right)dy. \quad (15)$$

But making use of Eq. (11) this becomes

$$\int_{-\infty}^{\infty} e^{ivy} J_x(y,x,z)dy = \frac{J(x,x,z)}{2\pi}\int_{-\infty}^{\infty}\rho(v,u,x)e^{-iux}r(u)du - \frac{J(-x,x,z)}{2\pi}\int_{-\infty}^{\infty}\rho(v,u,x)e^{iux}r(u)du. \quad (16)$$

Thus Eq. (14) becomes

$$\rho_x(v,z,x) = J(x,x,z)\left(e^{-ivx} + \frac{1}{2\pi}\int_{-\infty}^{\infty}\rho(v,u,x)e^{-iux}r(u)du\right) - J(-x,x,z)\left(e^{ivx} + \frac{1}{2\pi}\int_{-\infty}^{\infty}\rho(v,u,x) \times e^{iux}r(u)du\right). \quad (17)$$

Finally, by using the results in Eqs. (12) amd (13) we find that the function ρ satisfies the differential equation

$$\rho_x(v,z,x) = -\left(e^{-izx} + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iz'x}\rho(z',z,x)r(z')dz'\right) \times \left(e^{-ivx} + \frac{1}{2\pi}\int_{-\infty}^{\infty}\rho(v,u,x)e^{-iux}r(u)du\right) + \left(e^{+ixz} + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{iz'x}\rho(z',z,x)r(z')dz'\right) \times \left(e^{ivx} + \frac{1}{2\pi}\int_{-\infty}^{\infty}\rho(v,u,x)e^{iux}r(u)du\right) \quad (18)$$

$$-\infty < v,z < \infty, \quad 0 \leq x.$$

And from the definition in Eq. (11) we see that the initial condition on the function ρ at $x = 0$ is

$$\rho(v,z,0) = 0, \quad -\infty < v,z < +\infty. \quad (19)$$

Once the function ρ has been determined, the function J is given by Eq. (10), and then Eq. (6) yields the function U. As was pointed out earlier, though, in Eq. (3), we are not interested in the function U for all values of t in the interval $(-x,x)$, but only in the value of the total derivative of U with respect to x at $t = x$. We shall now show how this may be determined directly, that is, without determing U.

We observe that

$$\frac{d}{dx}U(x,x) = \frac{\partial}{\partial x}U(x,t)\Big|_{t=x} + \frac{\partial}{\partial t}U(x,t)\Big|_{t=x}. \quad (20)$$

From Eq. (6) we see that

$$\frac{\partial}{\partial x}U(x,t)\Big|_{t=x} = \frac{1}{2\pi}\int_{-\infty}^{\infty}[J_x(x,x,z)e^{-izx} - izJ(x,x,z)e^{-izx}]r(z)dz. \quad (21)$$

Keeping Eqs. (12), (13), and (8) in mind, we see that the right-hand side of Eq. (21) can be expressed in terms of the function ρ. Also we have

$$\frac{\partial}{\partial t}U(x,t)\Big|_{t=x} = \frac{1}{2\pi}\int_{-\infty}^{\infty} J_t(x,x,z')e^{-iz'x}r(z')dz'. \quad (22)$$

We can evaluate the function $J_t(t,x,z)\big|_{t=x}$ in the following way. From Eq. (5) we find that

$$0 = -ize^{-zt} + J_t(t,x,z) + \int_{-x}^{x}\frac{1}{2\pi}\int_{-\infty}^{\infty}(-iu)e^{-iu(t+y)}r(u)du J(y,x,z)dy. \quad (23)$$

This implies that

$$J_t(t,x,z) = ize^{-izt} + \frac{1}{2\pi}\int_{-\infty}^{\infty} iue^{-iut}\rho(u,z,x)r(u)du, \quad (24)$$

which expresses the function J_t in terms of the basic function ρ. This completes the demonstration.

3. DISCUSSION

In this paper we have presented the analytical aspects of a new approach to the integral equation of Kay and Balanis. The numerical aspects remain to be considered, but much previous experience with related initial value problems[4] indicates the efficacy of the approach.

[1] I. Kay, Commun. Pure Appl. Math. 12, 371 (1960).
[2] G. Balanis, J. Math. Phys. 13, 1001 (1972).
[3] R. Bellman and R. Kalaba, *Quasilinearization and Nonlinear Boundary Value Problems* (American Elsevier, New York, 1965).
[4] J. Casti and R. Kalaba, *Imbedding Methods in Applied Mathematics* (Addison-Wesley, New York, 1973)

Time-dependent multichannel Coulomb scattering theory*

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 1 March 1973; final revised manuscript received 20 August 1973)

The formulation by Mulherin and Zinnes of two-particle Coulomb scattering theory is extended to the multichannel case. The wave operators so obtained are proved by a direct method to be identical with those of Dollard.

1. INTRODUCTION

Several years ago Dollard[1,2] established a time-dependent multichannel Coulomb scattering theory that is mathematically coherent and physically plausible. Subsequently, proceeding along quite different lines, Mulherin and Zinnes[3] attempted to develop an alternative formalism. They were successful in the two-particle case and proved that their wave operators were the same as those of Dollard. They were not, however, successful in the general multichannel case, the problem thus remaining open.

The purpose of this paper is to complete the program of Mulherin and Zinnes by providing a multichannel generalization of their two-particle theory. Our interest in the matter is twofold. First, apart from rather convincing plausibility arguments advanced by Dollard, there is no evidence, theoretical or experimental, that we have seen supporting the physical validity of his multichannel theory. Since the somewhat different approach of Mulherin and Zinnes is equally plausible, a proof that the two multichannel theories are equivalent would be important evidence that the theories are physically sound and therefore worthy of further study. Second, the form of the generalized Mulherin-Zinnes theory would provide a concrete example of a mathematical formulation currently being advanced[4-6] for the general problem of scattering by long range forces.

This paper is thus devoted to extending the Mulherin-Zinnes theory to the multichannel case and to proving, by a method somewhat more direct than the original one, that the channel wave operators exist and are equal to those of Dollard.

2. STATEMENT OF THE THEOREM

Consider a system of N spinless distinguishable particles scattering into a channel β in which the particles are arranged into n clusters. At least two of the particles are assumed to be charged. The positions[7] of the centers of mass of the clusters are denoted by vectors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in $\mathbf{R}^{3n}$ and their corresponding momenta by $\mathbf{K} = (\mathbf{k}_1, \ldots, \mathbf{k}_n)$. The masses of the clusters are denoted by $m_1, \ldots, m_n$ and a diagonal mass matrix M is defined by the equation $M\mathbf{K} = (m_1\mathbf{k}_1, \ldots, m_n\mathbf{k}_n)$. The charges of the clusters are denoted by the charge numbers $z_1, \ldots, z_n$.

Because of the long range character of the Coulomb force the asymptotic clusters cannot be assumed to be freely moving, in contrast to the usual scattering theory for short range forces. A weak residual interaction must be built into the asymptotic time-dependent wave functions. The difference between the theories of Dollard[1,2] and of Mulherin-Zinnes[3] lies in how this is done.

In the theory of Dollard the asymptotic wave functions for the channel β are constructed as follows. Auxiliary functions are defined:

$$\phi(\mathbf{K}, \mathbf{X}, t) \equiv \sum_{s=1}^{n} [\mathbf{k}_s \cdot \mathbf{x}_s - (t/2m_s)|\mathbf{k}_s|^2], \tag{2.1}$$

$$\psi^{\pm}(\mathbf{K}, \mathbf{X}) \equiv -\alpha \sum_{1 \le r < s \le n} z_r z_s |(\mathbf{k}_r/m_r) - (\mathbf{k}_s/m_s)|^{-1} \ln \Delta^{\pm}_{rs}, \tag{2.2}$$

where α is the fine structure constant; and

$$\Delta^{\pm}_{rs} \equiv (m_r + m_s)^{-1}[|m_r\mathbf{k}_s - m_s\mathbf{k}_r| |\mathbf{x}_s - \mathbf{x}_r| \pm (m_r\mathbf{k}_s - m_s\mathbf{k}_r) \cdot (\mathbf{x}_s - \mathbf{x}_r)]. \tag{2.3}$$

The function

$$f^{\pm}_D(\mathbf{X}, t) \equiv (2\pi)^{-3n/2} \int d\mathbf{K} \hat{f}(\mathbf{K}) \exp[i\phi(\mathbf{K}, \mathbf{X}, t) \pm i\psi^{\pm}(\mathbf{K}, M^{-1}\mathbf{K}t)], \tag{2.4}$$

where $\hat{f}$ is a suitably chosen square integrable function, is used to describe the motion of the cluster centers of mass. Finally, the asymptotic wave functions for channel β are written in the form

$$F^{\pm}_{D,\beta}(t) = f^{\pm}_D(\mathbf{X}, t) g_\beta e^{-iE_\beta t}, \tag{2.5}$$

where g_β is stationary bound state wave function and E_β is the total bound state energy appropriate to the clusters of the channel.

The distinctive feature of Dollard's theory is that the spatial variation of the asymptotic wave function $f^{\pm}_D$ is of a plane wave type, represented by the exponent ϕ, just as in the short range theory. The long range nature of the Coulomb force is introduced by altering the usual time dependence, also contained in ϕ, by inclusion of the exponent $\psi^{\pm}(\mathbf{K}, M^{-1}\mathbf{K}t)$.

One might argue, on the other hand, that one should leave the time dependence alone and instead represent the motion of the clusters by a superposition of distorted plane waves. This approach, suggested by the well known time-independent theory for two particles, is the one adopted by Mulherin and Zinnes.[3] Their theory is generalized to the multichannel case by representing the system asymptotically with wave functions of the form

$$F^{\pm}_{MZ,\beta}(t) = f^{\pm}_{MZ}(\mathbf{X}, t) g_\beta e^{-iE_\beta t}. \tag{2.6}$$

Here $f^{\pm}_{MZ}$ is defined by

$$f^{\pm}_{MZ}(\mathbf{X}, t) \equiv (2\pi)^{-3n/2} \int d\mathbf{K} \hat{f}(\mathbf{K}) \exp[i\phi(\mathbf{K}, \mathbf{X}, t) \pm i\psi^{\pm}(\mathbf{K}, \mathbf{X})], \tag{2.7}$$

and $\hat{f}, g_\beta$, and E_β have the same meanings as before. The function $\psi^\pm(\mathbf{K}, \mathbf{X})$ clearly provides the advertisied distortion of the plane wave structure of ϕ.

One easily verifies that the formula (2.6) reduces in the two-particle case to the corresponding one of Mulherin and Zinnes. Since there are no bound state wave functions g_β appearing in the two-particle problem, one identifies $F^\pm_{MZ}$ of Eq. (2.6) with $f^\pm_{MZ}$ of Eq. (2.7). Variables appropriate to the center of mass coordinate system are introduced:

$$\mathbf{R} = (m_1\mathbf{x}_1 + m_2\mathbf{x}_2)/(m_1 + m_2), \qquad \mathbf{r} = \mathbf{x}_2 - \mathbf{x}_1, \tag{2.8}$$

$$\mathbf{p} = \mathbf{k}_1 + \mathbf{k}_2, \qquad \mathbf{k} = (m_1\mathbf{k}_2 - m_2\mathbf{k}_1)/(m_1 + m_2). \tag{2.9}$$

In terms of these coordinates one has

$$\phi(\mathbf{K}, \mathbf{X}, t) = \{\mathbf{p}\cdot\mathbf{R} - [t/2(m_1 + m_2)]|\mathbf{p}|^2\} + [\mathbf{k}\cdot\mathbf{r} - (t/2m)|\mathbf{k}|^2], \tag{2.10}$$

$$\psi^\pm(\mathbf{K}, \mathbf{X}, t) = -\alpha z_1 z_2 m|\mathbf{k}|^{-1}\ln(|\mathbf{k}|\,|\mathbf{r}| \pm \mathbf{k}\cdot\mathbf{r}), \tag{2.11}$$

where $m = m_1m_2/(m_1 + m_2)$ is the reduced mass of the system. If $\hat{f}(\mathbf{K})$ is assumed to be a product $\hat{f}(\mathbf{K}) = \hat{f}_0(\mathbf{p})\hat{f}_1(\mathbf{k})$, then the function $f^\pm_{MZ}$ factors into a product of two functions. The first represents the free motion of the center of mass and hence, because the paper of Mulherin and Zinnes deals only with the relative motion of the particles, is suppressed in their formulas. The second factor,

$$(2\pi)^{-3/2}\int d\mathbf{k}\hat{f}_1(\mathbf{k})\exp\{i[\mathbf{k}\cdot\mathbf{r} - (t/2m)|\mathbf{k}|^2] \mp i\alpha z_1 z_2 m|\mathbf{k}|^{-1}\ln(|\mathbf{k}|\,|\mathbf{r}| \pm \mathbf{k}\cdot\mathbf{r})\}, \tag{2.12}$$

is precisely of the form used by Mulherin and Zinnes to represent the asymptotic motion of the system (cf. Eqs. (13) and (19) of Ref. 3].

Now that the asymptotic wave functions have been defined, one can speak of the channel wave operators. In Dollard's theory these operators $\Omega^\pm_\beta$ are defined on product wave functions fg_β by

$$\lim_{t\to\pm\infty}\|\Omega^\pm_\beta fg_\beta - e^{iHt}F^\pm_{D,\beta}(t)\| = 0. \tag{2.13}$$

Here $\|\cdot\|$ denotes L^2 norm, and f denotes the Fourier transform of $\hat{f}$. The self-adjoint operator H is the Hamiltonian, including Coulomb interactions, of the system. Existence of the channel wave operators in the Mulherin-Zinnes theory is established in the following theorem.

Theorem: Let the notation of the previous paragraph be adopted. Assume that $f(\mathbf{X})$ is the Fourier transform of a Schwartz test function $\hat{f}(\mathbf{K})$ with compact support.[8] Assume further that $\hat{f}(\mathbf{K})$ vanishes in neighbourhoods of the hyperplanes $m_r\mathbf{k}_s = m_s\mathbf{k}_r$, $1 \le r < s \le n$. Assume that $\Omega^\pm_\beta fg_\beta$ exists. Then the equation

$$\lim_{t\to\pm\infty}\|\Omega^\pm_\beta fg_\beta - e^{iHt}F^\pm_{MZ,\beta}(t)\| = 0 \tag{2.14}$$

is true.

The theorem states not only that the Mulherin-Zinnes channel wave operators exist on functions f of the type specified, but also that they are equal to the Dollard operators. Since functions f of the type demanded by the theorem are dense in the space of square integrable functions, one concludes that the two theories are equivalent.

A proof of this theorem, by a method different from and somewhat more direct than the two-particle proof of Mulherin and Zinnes,[3] is given in the next section.

3. PROOF OF THE THEOREM

A proof is given for $t \to \infty$, that for $t \to -\infty$ being essentially the same.

The triangle inequality, Eq. (2.13), and the unitarity of the operator e^{iHt} imply that the theorem is true if and only if $\|F^\pm_{D,\beta}(t) - F^\pm_{MZ,\beta}(t)\|$ vanishes in the limit $t \to \infty$. From this fact, and from the definitions of $F^\pm_{D,\beta}$ and $F^\pm_{MZ,\beta}$, it follows immediately that the theorem is true if and only if

$$\lim_{t\to\pm\infty}\int d\mathbf{X}|f^\pm_D(\mathbf{X}, t) - f^\pm_{MZ}(\mathbf{X}, t)|^2 = 0, \tag{3.1}$$

where $f^\pm_D$ and $f^\pm_{MZ}$ are defined by Eq. (2.4) and Eq. (2.7), respectively. A necessary and sufficient condition for Eq. (3.1) to be true is that

$$\lim_{t\to\infty}\int d\mathbf{X}|R_D(\mathbf{X}, t) - R_{MZ}(\mathbf{X}, t)|^2 = 0. \tag{3.2}$$

Here R_D is defined by the equation

$$R_D(\mathbf{X}, t) \equiv f^+_D(\mathbf{X}, t) - \prod_{s=1}^{n}(m_s/it)^{3/2}\hat{f}(M\mathbf{X}/t) \times \exp[i\phi(M\mathbf{X}/t, \mathbf{X}, t) + i\psi^+(M\mathbf{X}/t, \mathbf{X})], \tag{3.3}$$

and R_{MZ} is defined by the same equation with the subscript D replaced by MZ. To prove the theorem, therefore, it is necessary and sufficient to prove Eq. (3.2).

Dollard has already proved[9] that for every set of integers $\{p_1, \ldots, p_n\}$ there exist constants C and μ such that the inequality

$$|R_D(\mathbf{X}, t)| \le Ct^{-(3n+1)/2}(\ln t)^\mu \prod_{j=1}^{n}[1 + (m_j|\mathbf{x}_j|/t)]^{-p_j} \tag{3.4}$$

is true for $t > 1$. Since R_D is bounded and continuous, and hence measurable, it follows that it is square integrable in $\mathbf{X}$ for all fixed $t > 1$ and that

$$\lim_{t\to\infty}\int d\mathbf{X}|R_D(\mathbf{X}, t)|^2 = 0. \tag{3.5}$$

By the triangle inequality, therefore, a necessary and sufficient condition for the theorem to be true is that

$$\lim_{t\to\infty}\int d\mathbf{X}|R_{MZ}(\mathbf{X}, t)|^2 = 0. \tag{3.6}$$

For notational convenience, the subscript MZ will now be dropped for the remainder of the proof.

The function $R(\mathbf{X}, t)$, and hence also $|R(\mathbf{X}, t)|^2$, is clearly a measurable function of $\mathbf{X}$ for each fixed $t > 0$ since it is bounded and continuous at all points $\mathbf{X}$ in the complement of

$$E \equiv \{\mathbf{X} \mid \mathbf{x}_r = \mathbf{x}_s \text{ for some } r \text{ and } s, 1 \le r < s \le n\}, \tag{3.7}$$

a set of measure zero in $\mathbf{R}^{3n}$. Therefore if there exists a positive integrable function $S(\mathbf{X}, t)$ such that for $t > 1$ the inequality $|R(\mathbf{X}, t)|^2 \le S(\mathbf{X}, t)$ holds, and if

$$\lim_{t\to\infty}\int d\mathbf{X}S(\mathbf{X}, t) = 0, \tag{3.8}$$

then R is square integrable for each $t > 1$ and Eq. (3.6) holds.

In establishing the existence of the function $S(\mathbf{X}, t)$ and thus completing the proof of the theorem, one may assume without loss of generality that the (compact) support of $\hat{f}$ lies in the interior of a cone

$$C(\mathbf{N},\theta) \equiv \{\mathbf{K}|\ (m_r\mathbf{k}_s - m_s\mathbf{k}_r)\cdot(\mathbf{n}_s - \mathbf{n}_r) \geq \cos\theta|\ m_r\mathbf{k}_s - m_s\mathbf{k}_r|\ |\ \mathbf{n}_s - \mathbf{n}_r|,\ 1 \leq r < s \leq n\}. \quad (3.9)$$

Here $\mathbf{N} = (\mathbf{n}_1, \ldots, \mathbf{n}_n)$ is an arbitrary fixed vector and $0 < \theta < (\pi/2)$. Were the support of $\hat{f}$ not contained in such a cone, use of a suitable partition of unity would permit $\hat{f}$ to be written as a finite sum of functions $\hat{f}_j$, each with supports in cones $C(\mathbf{N}_j, \theta)$. Linearity then allows Eq. (3.6) to be proved for each $\hat{f}_j$ separately. Hence the additional support property can be assumed at the outset.

The remaining calculation of S is divided into two steps.

Case A: $M\mathbf{X} \notin C(\mathbf{N}, \theta)$

In this case $\hat{f}(M\mathbf{X}/t) = 0$ so that R and f^+ are identical.

To calculate S it is convenient to define

$$\tau \equiv \left(t^2 + \sum_{s=1}^{n} |\mathbf{x}_s|^2\right)^{1/2} \quad (3.10)$$

and consider the function $(1+\tau^2)^N R$, where N is an arbitrary integer.

Trivial algebra yields

$$(1+\tau^2)^N R = (2\pi)^{-3n/2} Q \exp[i\phi(M\mathbf{X}/t, \mathbf{X}, t)], \quad (3.11)$$

where

$$Q(\mathbf{X},t) \equiv (1+\tau^2)^N \int d\mathbf{K}\hat{f}(\mathbf{K}) \exp[-(i\tau^2/2t)\chi(\mathbf{K},\mathbf{X},t) + i\psi^+(\mathbf{K},\mathbf{X})], \quad (3.12)$$

$$\chi(\mathbf{K},\mathbf{X},t) \equiv \tau^{-2} \sum_{s=1}^{n} m_s^{-1} |t\mathbf{k}_s - m_s\mathbf{x}_s|^2. \quad (3.13)$$

Note that (t/τ) and $(\mathbf{X}/\tau)$ are bounded quantities. Hence χ is actually a function of variables $\{\mathbf{K}, (t/\tau), (\mathbf{X}/\tau)\}$ that range over compact sets. It follows from the continuity of χ that there exists $\delta > 0$ such that $\chi \geq \delta$ uniformly for $\mathbf{K}$ in the support of $\hat{f}$ and $\mathbf{X}$ such that $M\mathbf{X} \notin C(\mathbf{N}, \theta)$.

The problem now is to compute a bound for Q. First define the operators

$$D_0 \equiv \tau^{-1} \sum_{a=1}^{n} (t\mathbf{k}_a - m_a\mathbf{x}_a)\cdot\nabla_a, \quad (3.14)$$

where ∇_a denotes the gradient with respect to $\mathbf{k}_a$, and

$$D_1 \equiv \chi^{-1} D_0. \quad (3.15)$$

One now easily verifies that

$$Q(\mathbf{X},t) = \int d\mathbf{K}\hat{f} e^{i\psi^+}(1 - D_1^2)^N e^{-i(\tau^2/2t)\chi}. \quad (3.16)$$

The required bound now follows from partial integration of Eq. (3.16).

This partial integration is justified by an appeal to the divergence theorem. The usual statement of this theorem requires, however, that derivatives of the integrand be continuous. This requirement is unfortunately not met in the present case because of the singular nature of the function ψ^+. The support properties of $\hat{f}$ allow one to ignore singularities of the type $|m_s\mathbf{k}_r - m_r\mathbf{k}_s|^{-m}$, where m is some integer. The other singularities generated by application of powers of D_0 to $e^{i\psi^+}$ are at worst logarithmic in $\mathbf{K}$, a clear implication of the explicitly calculated equation

$$D_0 \ln\Delta_{rs}^+ = \tau^{-1}\{t - |\mathbf{x}_r - \mathbf{x}_s|\ |\ (\mathbf{k}_r/m_r) - (\mathbf{k}_s/m_s)|^{-1}\}. \quad (3.17)$$

One generalizes the divergence theorem to accommodate the weak logarithmic singularities encountered here by applying the usual theorem in the region exterior to an ϵ-neighborhood of the singular set and letting ϵ go to zero.

Since $\hat{f}$ has compact support the surface terms vanish and the partial integrations yield

$$Q = \int d\mathbf{K} e^{-(i\tau^2/2t)\chi}\{1 - [(D_0 + (3nt/\tau))\chi^{-1}]^2\}^N e^{i\psi^+}\hat{f}. \quad (3.18)$$

One now calculates the derivatives in Eq. (3.18) to obtain a bound for Q and hence for R. The result is a bound of the form

$$|R(\mathbf{X},t)| \leq \left(1 + t^2 + \sum_{s=1}^{n} |\mathbf{x}_s|^2\right)^{-N} \sum_{\{\nu_{rs}\}} A(\{\nu_{rs}\}) \prod_{r<s} |\ln|\mathbf{x}_r - \mathbf{x}_s||^{\nu_{rs}}, \quad (3.19)$$

where the sum is over all sequences $\{\nu_{rs}\}$ of integers ν_{rs} such that $\sum \nu_{rs} \leq 2N$, and where the $A(\{\nu_{rs}\})$ are constants.

The bounding function $S(\mathbf{X},t)$ for $M\mathbf{X} \notin C(\mathbf{N},\theta)$ is the square of the right-hand side of Eq. (3.19).

Case B: $M\mathbf{X} \in C(\mathbf{N}, \theta)$

To deal with this case one writes the function $f^+(\mathbf{X},t)$ as

$$f^+(\mathbf{X},t) = g^+(\mathbf{X},\mathbf{X}',t)|_{\mathbf{X}=\mathbf{X}'}, \quad (3.20)$$

where

$$g^+(\mathbf{X},\mathbf{X}',t) \equiv (2\pi)^{-3n/2} \int d\mathbf{K}\hat{f}(\mathbf{K}) \exp[i\phi(\mathbf{K},\mathbf{X},t) + i\psi^+(\mathbf{K},\mathbf{X}')]. \quad (3.21)$$

By considering the variable $\mathbf{X}'$ as fixed and following Dollard (Chap. III, Sec. II of Ref. 1; Lemma 2 of Ref. 2) in the treatment of the variable $\mathbf{X}$ one obtains

$$R(\mathbf{X},t) = \prod_r (m_r/2\pi i t)^{3/2} P(\mathbf{X},\mathbf{X}',t)|_{\mathbf{X}=\mathbf{X}'} \exp[i\phi(M\mathbf{X}/t,\mathbf{X},t)], \quad (3.22)$$

where [compare Eq. (68), p. 129 of Ref. 1; Eq. (45) of Ref. 2]

$$P(\mathbf{X},\mathbf{X}',t) \equiv \int d\mathbf{Y} g^+(\mathbf{Y},\mathbf{X}',0)\Delta(\mathbf{Y},t) \exp\left((i/t)\sum_{s=1}^{n} m_s\mathbf{x}_s\cdot\mathbf{y}_s\right), \quad (3.23)$$

$$\Delta(\mathbf{Y},t) \equiv \exp[i\phi(M\mathbf{Y}/t,\mathbf{Y},t)] - 1. \quad (3.24)$$

The problem now is to compute a bound for $P(\mathbf{X},\mathbf{X}',t)$.

First, however, one must consider the function $g^+(\mathbf{Y},\mathbf{X}',0)$. It is clear from the assumed properties of $\hat{f}$ that all derivatives of $g^+(\mathbf{Y},\mathbf{X}',0)$ with respect to $\mathbf{Y}$ are bounded. One also needs, for arbitrary N, a bound for

$$h(\mathbf{Y},\mathbf{X}) \equiv \left(1 + \sum_{s=1}^{n} |\mathbf{y}_s|^2\right)^N g^+(\mathbf{Y},\mathbf{X}',0). \quad (3.25)$$

To obtain this define the operator

$$D_2 \equiv \left(\sum_{s=1}^{n} |\mathbf{y}_s|^2\right)^{-1/2} \sum_{a=1}^{n} \mathbf{y}_a\cdot\nabla_a \quad (3.26)$$

where ∇_a denotes the gradient with respect to $\mathbf{k}_a$. Then

$$h(\mathbf{Y},\mathbf{X}') = (2\pi)^{-3n/2}\int d\mathbf{K}\hat{f}e^{i\psi^+}[1-D_2^2]^N \exp\left(i\sum_{s=1}^{n}\mathbf{k}_s\cdot\mathbf{y}_s\right). \tag{3.27}$$

Integration by parts, using the divergence theorem, is justified by the assumed properties of $\hat{f}$ and the fact that for $M\mathbf{X}'$ in $C(\mathbf{N},\theta) - E$ and $\mathbf{K}$ in the support of $\hat{f}$ the function ψ^+ has continuous derivatives with respect to $\mathbf{K}$ of all orders. By explicit calculation one verifies that

$$D_2\ln\Delta^{\pm}_{rs} = \left(\sum|\mathbf{y}_s|^2\right)^{-1/2}[|\mathbf{x}'_s-\mathbf{x}'_r|(m_r\mathbf{y}_s - m_s\mathbf{y}_r)\cdot\mathbf{z}_{sr}] \times[|m_r\mathbf{k}_s - m_s\mathbf{k}_r|(\mathbf{x}'_s-\mathbf{x}'_r)\cdot\mathbf{z}_{sr}]^{-1} \tag{3.28}$$

$$\mathbf{z}_{sr} = |\mathbf{x}'_s-\mathbf{x}'_r|(m_r\mathbf{k}_s - m_s\mathbf{k}_r) + |m_r\mathbf{k}_s - m_s\mathbf{k}_r|(\mathbf{x}'_s-\mathbf{x}'_r). \tag{3.29}$$

Using the fact that both $\mathbf{K}$ and $M\mathbf{X}'$ belong to $C(\mathbf{N},\theta)$, where $0<\theta<\pi/2$, one easily computes the inequality

$$|D_2\ln\Delta^{\pm}_{rs}| \le 2(m_r+m_s)(1+\cos2\theta)^{-1}|m_r\mathbf{k}_s - m_s\mathbf{k}_r|^{-1}. \tag{3.30}$$

The implication of Eq. (3.30) and similarly computed bounds for higher powers of D_2 applied to $\ln\Delta^{\pm}_{rs}$, is that application of powers of D_2 to $e^{i\psi^+}$ can give at worst logarithmic behavior in $\mathbf{X}'$. Performance of the partial integration in Eq. (3.27) thus leads to a bound of the form

$$|g^+(\mathbf{Y},\mathbf{X}',0)| \le \left(1+\sum_{s=1}^{n}|\mathbf{y}_s|^2\right)^{-N}\sum_{\{\nu_{rs}\}}B(\{\nu_{rs}\}) \times\prod_{r<s}|\ln|\mathbf{x}'_r-\mathbf{x}'_s||^{\nu_{rs}}, \tag{3.31}$$

where the symbolism has the same meaning as in Eq. (3.19). It is also clear from the preceding discussion that a similar bound holds for derivatives of g^+ with respect to $\mathbf{Y}$.

One now slavishly imitates the technique of Dollard (Chap. III, Sec. II of Ref. 1; Lemma 2 of Ref. 2) to obtain for $t \ge 1$, and arbitrary N, a bound of the form

$$|P(\mathbf{X},\mathbf{X}',t)| \le t^{-1/2}\left(1+t^{-2}\sum_{s=1}^{n}|\mathbf{x}_s|^2\right)^{-N}\sum_{\{\nu_{rs}\}}C(\{\nu_{rs}\}) \times\prod_{r<s}|\ln|\mathbf{x}'_r-\mathbf{x}'_s||^{\nu_{rs}}. \tag{3.32}$$

The bounding function $S(\mathbf{X},t)$ for $M\mathbf{X}\in C(\mathbf{N},\theta)$ is now easily derived from Eqs. (3.22) and (3.32). The proof of the theorem is thus completed.

4. SUMMARY

In summary, the two-particle Coulomb scattering theory of Mulherin and Zinnes[3] has been extended to the general multichannel situation. In addition, the wave operators of the generalized theory have been shown to be identical on a dense subset of functions to those proposed earlier by Dollard.[1,2]

The proof given here appears to be more direct than the original one of Mulherin and Zinnes, and in fact we have found no obvious generalization of their argument. The present proof takes advantage of the distinction (Cases A and B) between vectors $\mathbf{X}$ outside the velocity cone[10] of $\hat{f}$ and those inside, a distinction known to be useful in relativistic scattering problems[10-12] as well as in the physical interpretation of nonrelativistic scattering theory.[13] Whether this distinction is useful in other nonrelativistic problems is a question for further research.

Finally, note that the generalized Mulherin-Zinnes theory can be written in a slightly different form. Define the mapping $K^{\pm}_{\beta}$ from the subspace of product wave functions fg_{β} into the full Hilbert space $L^2(\mathbf{R}^{3N})$ by the equation

$$K^{\pm}_{\beta}fg_{\beta} \equiv F^{\pm}_{MZ,\beta}(0). \tag{4.1}$$

The operator $K^{\pm}_{\beta}$ thus represents the effect of the residual long range interaction between the asymptotic clusters. Although no specific properties, such as boundedness, of the operator $K^{\pm}_{\beta}$ are known, Eq. (4.1) does imply that the wave operators $\Omega^{\pm}_{\beta}$ can be written

$$\Omega^{\pm}_{\beta} = \lim_{t\to\pm\infty} e^{iHt}K^{\pm}_{\beta}e^{-iH_{\beta}t} \tag{4.2}$$

on a dense set of vectors. Here the operator H_{β} is the channel Hamiltonian[14] appropriate to channel β. The possibility of such a representation for the wave operators was already noticed by Mulherin and Zinnes in the two-particle case [Ref. 3, Eq. (39)]. This same type of representation, although with different operators $K^{\pm}_{\beta}$, has recently been investigated by several authors in the more general context of long range scattering.[4-6] The generalized Mulherin-Zinnes theory can be considered therefore, as a concrete example, the physics and mathematics of which are well understood, of these more general formulations.

Note added in proof: The result of this paper has been independently reported by L. Rosenberg [Phys. Rev. D 8, 1833 (1973)]. He does not give a proof but refers instead to unpublished work of P. J. Redmond.

*Supported in part by Sandia Laboratories, SURP Contract 51-3475.

[1] J. D. Dollard, Thesis, Princeton University, 1963.
[2] J. D. Dollard, J. Math. Phys. 5, 729 (1964).
[3] D. Mulherin and I. I. Zinnes, J. Math. Phys. 11, 1402 (1970).
[4] E. Prugovečki, J. Math. Phys. 14, 957 (1973).
[5] E. Prugovečki and J. Zorbas, J. Math. Phys. 14, 1398 (1973).
[6] V. B. Matveev and M. M. Skriganov, Teor. Mat. Fyz. 10, 238 (1972).
[7] By a whim of the authors, variables in this paper are dimensionless. To this end time is measured in units of $(\hbar/mc^2)$, where m is the electron mass. Distance is measured in units of $(\hbar/mc)$, momentum in units of mc, mass in units of m. Charge is measured in units of the electron charge.
[8] L. Schwartz, *Theories des distributions* (Hermann, Paris, 1957) Vol. I.
[9] Chapter III, Sec. II of Ref. 1. See in particular Eqs. (40), (52), (67), and (68). The variables $\mathbf{k}_j$ in Eq. (52) and $\mathbf{y}_j$ in Eqs. (67) and (68) are to be replaced in the present application by $(m_j\mathbf{x}_j/t)$. The subscript m_0 in Eq. (52) is to be replaced by f. Dollard's function R_f is related to R_D by $R_D = \Pi(m_s/it)^{3/2}\exp[i\phi(M\mathbf{X}/t,\mathbf{X},t)]R_f$. The extra factors of 2π in these equations arise from a different normalization of the Fourier transform. The treatment of the variable t in these equations is slightly different from that of this paper, but is easily seen to reduce to the same thing in the limit $t\to\infty$, a fact Dollard himself notes on p. 121.
[10] D. Ruelle, Helv. Phys. Acta 35, 147 (1962).
[11] K. Hepp in *Axiomatic Field Theory, 1965 Brandeis Lectures* (Gordon & Breach, New York, 1966), Vol. I, and references cited therein.
[12] C. Chandler, Phys. Rev. 174, 1749 (1968); C. Chandler and H. P. Stapp, J. Math. Phys. 10, 826 (1969).
[13] J. D. Dollard, Commun. Math. Phys. 12, 193 (1969) and J. Math. Phys. 14, 708 (1973).
[14] Reference 2, Sec. III.

Some differential-difference equations containing both advance and retardation*

L. S. Schulman

Physics Department, Technion, Haifa, Israel

Physics Department, Indiana University, Bloomington, Indiana 47401

(Received 13 July 1973)

An explicit solution is given to the boundary value problem for certain linear differential-difference equations. The solution is well behaved even in the presence of advanced interactions. Interest in these equations arises from study of time symmetric electrodynamics.

Time symmetric electrodynamics contradicts experience because we do not observe advanced interactions. Wheeler and Feynman[1] proposed that the universe acts to absorb the advanced interactions leaving a residual radiation reaction. In this article we study equations modeled on those of time symmetric electrodynamics, namely difference-differential equations, in the hope of determining whether the absorber theory and various other theories which arose from it[2,3] are indeed justified.

A great deal is known about differential equations with retarded arguments[4,5,6] including the retarded equations of electrodynamics.[7] For advanced interactions many of the theorems do not hold. Generally speaking, this is because unless one gives very special initial data (for a system with advanced arguments) the solution soon ceases to exist. We illustrate this with an example. Let

$$D^2x(t) + \omega^2 x(t) = \tfrac{1}{2}\alpha x(t-\tau) + \tfrac{1}{2}\beta x(t+\sigma) + \psi(t), \qquad (1)$$

where $D = d/dt$; $\alpha, \beta, \tau, \sigma$ are given constants, $\tau, \sigma > 0$ and $\psi(t)$ is a given function. For $\beta = 0$, if $x(t)$ is given on the interval $0 \le t \le \tau$ then it can be obtained at all subsequent times and becomes smoother on successive intervals. For $\beta \neq 0$ (and $\alpha \neq 0$) data must be given on an interval of length $\tau + \sigma$ and as the equation is iterated forward in time the solution becomes successively less differentiable and singularities develop at $t =$ (integer multiples of τ and σ) unless special initial conditions are given.

These problems seem to have hampered study of TSE (time symmetric electrodynamics). In this article we take a different approach which makes equations such as Eq. (1) far more tractable. The approach is simply to give boundary values rather than initial values for the solutions. Mathematically this will be seen to "neutralize" the equations. Physical justification is more serious and must be presented in the context of absorber theory and the kind of results one can hope to get from such a theory. Discussion of this point may be found in Ref. 8. In Fig. 1 is an indication of why Eq. (1) may have something to do with the problem of the absorber theory of Wheeler and Feynman.

In this article we present a solution of the boundary value problem for Eq. (1). First, however, we show how the boundary value problem arises naturally for time symmetric equations. Equation (1) with $\tau = \sigma$, $\alpha = \beta$, $\psi \equiv 0$ can be derived by variation of the action[9]

$$S = \tfrac{1}{2}\int dt\,[(Dx(t))^2 - \omega^2 x(t)^2 + \alpha x(t+\tau/2)x(t-\tau/2)]. \qquad (2)$$

Let the t integration run from a to b and consider $\delta S = S[x+\delta x] - S[x]$ where $x(t)$ is not yet restricted. Then

$$\begin{aligned}\delta S = \int_{a+\tau/2}^{b-\tau/2} dt\,\delta x(t)\{&-D^2x(t) - \omega^2 x(t) \\ &+ \tfrac{1}{2}\alpha[x(t+\tau) + x(t-\tau)]\} \\ &+ D x\delta x\,|_a^b - \left(\int_a^{a+\tau/2} + \int_{b-\tau/2}^{b}\right) dt\,\delta x\,(D^2x + \omega^2 x) \\ &+ \frac{\alpha}{2}\int_{a-\tau/2}^{a+\tau/2} x(t+\tau)\,\delta x(t)\,dt + \frac{\alpha}{2}\int_{b-\tau/2}^{b+\tau/2} x(t-\tau)\,\delta x(t)\,dt. \end{aligned} \qquad (3)$$

In order for δS to vanish we can require the term in the brackets in the first integral to vanish—this is the difference-differential equation, Eq. (1). In addition, the other terms in Eq. (3) must vanish. This can be accomplished by taking $\delta x(t) \equiv 0$ for $a - \tau/2 \le t \le a + \tau/2$ and $b - \tau/2 \le t \le b + \tau/2$. But the statement that $\delta x(t) \equiv 0$ for certain values of t (say $t \in I$) simply means that all $x(t)$ in the class of functions relative to which we wish S to be minimum must have the same value for $t \in I$. The variational problem is thus defined by giving $x(t)$ for $t \in I$. In the usual classical mechanics, I just consists of the initial and final times. For S of Eq. (2)

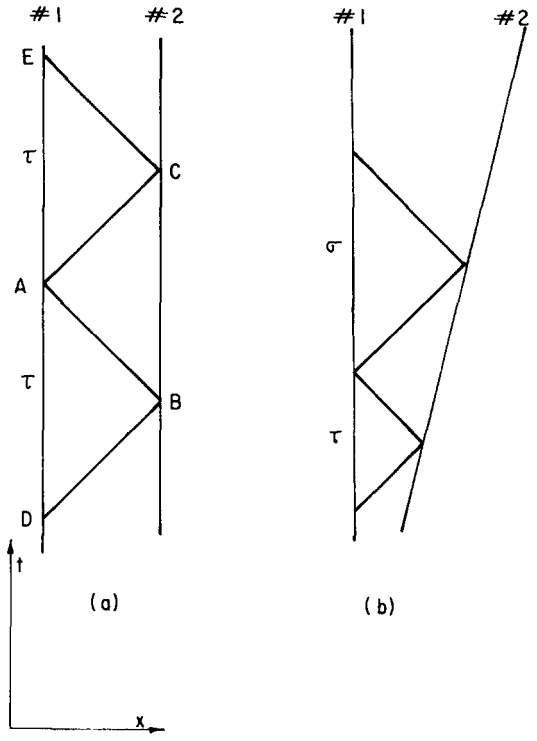

FIG. 1. Reaction with the "universe" reduced to self interaction. Both figures are space-time diagrams with t vertical. In (a), motion of particle # 1 at A affects # 2 at C and B which in turn affect # 1 at E, A and D. In (b) the particles are moving apart (on some cosmological scale) and the advanced deviation (σ) is greater than the retardation (τ). The inclusion of all absorbers into a single absorber is undoubtedly physically unreasonable.

I is $[a - \tau/2, a + \tau/2] \cup [b - \tau/2, b + \tau/2]$. We shall prove below by explicit construction that specification of $x(t)$ in these intervals does in fact provide adequate boundary conditions for a unique solution of Eq. (1) for almost all values of α, etc. Nevertheless, the formal manipulations just given suggest the naturalness of the boundary value problem.

The foregoing result could have been anticipated from the iteration scheme for solution of retarded interactions (sometimes call the "method of steps"). The initial condition problem for Eq. (1)—to the extent that blowups in the solution can be controlled—requires data on an interval of length $2\tau (\sigma = \tau)$. All we have done is break this interval into two pieces. The consequences of this are that for some special parameter values there is ambiguity in the solution and also as we shall see, to make the solution in the interval $[a + \tau/2, b - \tau/2]$ far smoother than that which is generated by the method of steps.

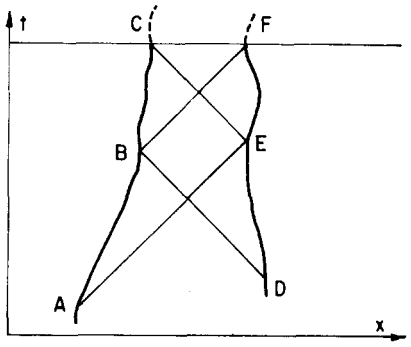

FIG. 2. Space–time diagram for initial conditions for two particles interacting via time symmetric electrodynamics. $c = 1$, so that AE, EC, BD, and BF are at 45° to the axes. Trajectories for the particles are given as some curves DF, AC and iteration forward in t from the points C and F (dotted lines) is done using the equations of motion.

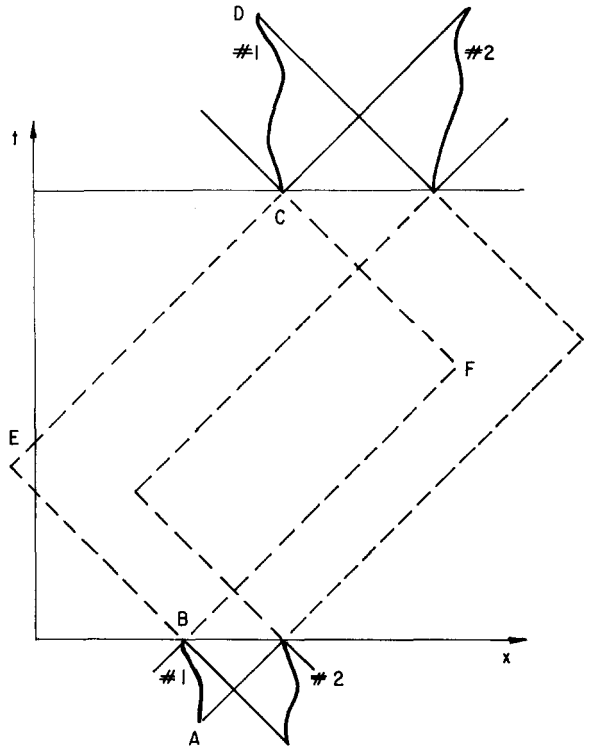

FIG. 3. Boundary conditions for time symmetric electrodynamics. The data that are given are the indicated paths of particles # 1 (AB and CD) and # 2. This is sufficient, for, e.g., # 1 does not depart from the rectangle $BECF$ (light velocity is unity) and its light cones either intersect the boundary data for # 2 or the path of # 2 in its associated rectangle. For additional space dimensions the rectangle becomes an intersecting pair of cones.

For two particles interacting via TSE the method of steps suggests initial conditions on an interval defined by a double light cone as illustrated in Fig. 2. Although we have not established any existence proof there does not seem to be any physical reason for failure of these initial conditions. In fact, Hui[10] has done computer studies iterating the equations of TSE forward in time. The only problem he encountered is the tendency of solutions to these equations to blow up at the discontinuity that occurs at the end of what corresponds to intervals of length τ for this variable deviation problem, namely the point where light cones out of C and F (in Fig. 2) meet the curves produced by the iteration, and at analogous points later on.

The boundary value problem for TSE can be similarly defined, as can be seen in Fig. 3. For two particles each must have its trajectory given in the corresponding light cone of the other. Thus instead of insisting on trajectories which go to infinity as many authors do, we have a well-defined variational principle with a finite action. We observe, however, that while Fig. 3 suggests necessary data for the boundary value problem, it does not guarantee the existence of a solution. Very likely to get a smooth solution some conditions will have to be placed on the data.

EXPLICIT SOLUTION OF THE BOUNDARY VALUE PROBLEM

We now justify the heuristic remarks above concerning existence of solutions for the boundary value problem by actually constructing such a solution. Rather than turning to Eq. (1) immediately we consider first the symmetric case $\sigma = \tau$, $\alpha = \beta$. Then

$$(D^2 + \omega^2)\, x(t) = (\alpha/2)\,[x(t + \tau) + x(t - \tau)\,] + \psi(t) \tag{4}$$

is to hold in the interval $0 \le t \le T$ with boundary condition $x(t) = \phi(t)$ for $-\tau \le t \le 0$ and $T \le t \le T + \tau$. The only condition on ϕ will be its integrability. For simplicity we assume $T = N\tau$ and take $\tau = 1$ (scaling of other quantities with τ can easily be recovered). Define

$$\begin{aligned} &\psi_n(t) = \psi(t + n - 1), \quad x_n(t) = x(t + n - 1), \quad n = 1, \ldots, N, \\ &x_0(t) = \phi(t - 1), \quad x_{N+1}(t) = \phi(t + N), \\ &[\,\xi(t)\,]_n = x_n(t), \quad n = 1, \ldots, N \end{aligned} \tag{5}$$

for $0 \le t \le 1$. Then Eq. (4) reads

$$(D^2 + \omega^2)\, x_n(t) = (\alpha/2)\,[x_{n+1}(t) + x_{n-1}(t)] + \psi_n(t) \tag{6}$$

for $n = 1, \ldots, N$. Letting R be the $N \times N$ raising operator $R_{ij} = \delta_{i+1,j}$ and L its transpose the lowering operator, Eq. (6) takes the form

$$(D^2 + \omega^2)\, \xi(t) = (\alpha/2)\,(R + L)\, \xi(t) + f(t), \tag{7}$$

where

$$f(t) = \begin{pmatrix} \phi(t-1) \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ \phi(t+N) \end{pmatrix} + \begin{pmatrix} \psi_1(t) \\ \cdot \\ \cdot \\ \cdot \\ \psi_N(t) \end{pmatrix}.$$

With the definition

$$\Omega^2 = \omega^2 1 - (\alpha/2)\,(R + L), \tag{8}$$

Eq. (7) becomes

$$(D^2 + \Omega^2)\,\xi = f. \tag{9}$$

A particular solution of Eq. (9) can easily be obtained using a Green's function; which of the various possible Green's functions (retarded, advanced, etc.) is used is irrelevant as the resulting particular solutions will differ only by solution of the homogeneous equation.

Solution of Eq. (9) still need not provide a solution to Eq. (4). Given a solution $\xi(t)$, the value of $(\xi(t))_n$ at $t = 1$ has no special relation to the value of $(\xi(t))_{n+1}$ at $t = 0$. On the other hand, for the function $x(t)$ we may demand continuity or more stringent conditions at integer values of t. We thus tailor the homogeneous solutions to Eq. (9) so as to relate $x_n(1)$ to $x_{n+1}(0)$. The space of homogeneous solutions is $2N$-dimensional. Continuity at $t = 0, 1, \ldots, N$, namely

$$x_{n+1}(0) = x_n(1), \quad n = 0, 1, \ldots, N, \tag{10}$$

involves $N + 1$ conditions. As the remaining $N - 1$ conditions take

$$Dx_n(1) = Dx_{n+1}(0), \quad n = 1, \ldots, N-1. \tag{11}$$

This is obviously an arbitrary albeit symmetric choice. Equation (10) can be written

$$\begin{aligned} \xi(1) &= R\xi(0) + v, \\ \xi(0) &= L\xi(1) + u, \end{aligned} \tag{12}$$

where $(u)_1 = \phi(0)$, $(v)_N = \phi(N)$ with other components of u and v zero. Equation (11) takes the form

$$LD\xi(1) = LRD\xi(0). \tag{13}$$

Equations (12) and (13) are $3N$ equations for $2N$ unknowns of which N equations are either redundant or identically zero. The function $\xi(t)$ has the form

$$\xi(t) = F(t) + H(t)\,\gamma, \tag{14}$$

where

$$F(t) = \tfrac{1}{2}\int_0^1 ds\ \frac{\sin[\Omega\,|t-s\,|]}{\Omega}\ f(s)$$

is an $N \times 1$ matrix, the "time symmetric" Green's function of Eq. (9) has been chosen, and $H(t)$ is an $N \times 2N$ matrix satisfying

$$\begin{aligned} &(D^2 + \Omega^2)\,H = 0, \\ &H(0) = (1 \mid 0), \quad H(1) = (0 \mid 1), \end{aligned} \tag{15}$$

(each block in H is an $N \times N$ matrix) and γ is a $2N \times 1$ matrix. $H\gamma$ is thus a homogeneous solution of Eq. (9) and by (15) H is

$$H(t) = \left(\frac{\sin\Omega(1-t)}{\sin\Omega} \;\middle|\; \frac{\sin\Omega t}{\sin\Omega}\right), \tag{16}$$

Defining projections $(P)_{ij} = \delta_{ij}$, $i = 1$, $j = 1, \ldots, N$ (a $1 \times N$ matrix) and $(P')_{ij} = \delta_{i+1,j}$, $i = 1, \ldots, N-1$, $j = 1, \ldots, N$ (an $(N-1) \times N$ matrix) Eqs. (12) and (13) become

$$\begin{pmatrix} P(-\,1 \mid L) \\ (-\,R \mid 1) \\ P'[LDH(1) \\ -\,LRDH(0)] \end{pmatrix} \gamma = \begin{pmatrix} -\,\phi(0) + PF(0) - PLF(1) \\ v + RF(0) - F(1) \\ P'[LRDF(0) - LDF(1)] \end{pmatrix} \equiv \chi \tag{17}$$

providing $2N$ equations for $2N$ unknowns. For computational purposes some further simplification can be made if γ and χ are broken into two column vectors of N components each

$$\gamma = \begin{pmatrix} \gamma' \\ \gamma'' \end{pmatrix}, \quad \chi = \begin{pmatrix} \chi' \\ \chi'' \end{pmatrix},$$

so that the first N of Eq. (17) allow a simple substitution for γ' in the second N equations, which now take the form

$$\begin{pmatrix} 0 \cdots & \cdots\, 0\ 1 \\ P'[LDH(1) - LRDH(0)] & \end{pmatrix} \gamma'' = \chi''.$$

The simplicity of the first row of this matrix allows the reduction of one more dimension and we introduce the $N \times (N-1)$ matrix $(P'')_{ij} = \delta_{ij}$, $i = 1, \ldots, N$, $j = 1, \ldots, N-1$, which when right multiplying an $N \times N$ matrix simply removes its last column.

Using the fact that

$$DH(t) = \frac{\Omega}{\sin\Omega}\,(-\cos\Omega(1-t) \mid \cos\Omega t),$$

we obtain finally that the solution γ involves inversion of the $(N-1) \times (N-1)$ matrix

$$Q = P'L(1 - R\cos\Omega)\frac{\Omega}{\sin\Omega}(1 - L)P''. \tag{18}$$

We note that Q is an analytic function of Ω^2 [except when $(\sin\Omega/\Omega)$ vanishes] and its determinant is an analytic function of ω^2 and α, except at those poles. For $\Omega = (\pi/2)\,1$ the determinant of Q is easily seen to be different from zero and hence, using analytic continuation this determinant can only be zero on a set of measure zero in the space of ω^2 and α. This proves that the solution of the boundary value problem for the variational principle of Eq. (2) (if it exists) is unique except for special values of α and ω^2. The values of α and ω^2 for which $\Omega/\sin\Omega$ has poles correspond to cases where there are homogeneous solutions of Eq. (9) which have period one and hence are of no use in adjusting $\xi(0)$ and $\xi(1)$ so as to satisfy Eqs. (10) and (11). The system in this case is overdetermined, as opposed to the underdetermination that corresponds to $\det Q = 0$.

Whether or not the solution to the differential equation that we have just obtained is a solution for the variational problem, depends on how one handles derivatives at the integer values of t in the variational problem.

A great deal is known about the matrix Ω^2, in particular all its eigenvalues and eigenfunctions. Unfortunately we have not succeeded—despite all this information—in inverting Q, except on the computer.

One check of our formalism is to set $\phi(t) = e^{st}$, $\psi(t) \equiv 0$. If one now demands that $x(t) = e^{st}$ for $0 \le t \le T$ also, then it is quite satisfying to see how the maze of matrix equations succeeds in forcing the condition $s^2 + \omega^2 = \alpha\,\cosh s\,(\tau = 1)$ on s.[4] [Compare the effect of substituting $x = e^{st}$ directly into Eq. (4) with $\psi \equiv 0$].

It may be noted that in contrast to the method of steps for advanced interactions the method described here produces a solution which is an integral of the boundary value function, and hence possesses one more derivative on the interior of the intervals $[j, j+1]$, $j = 0, \ldots, N-1$; at the points $t = 0$ and N, $x(t)$ is continuous and at the points $t = 1, \ldots, N-1$ possesses a continuous derivative.

Our method can be generalized to situations where the retardation and advance are not equal. Physically this is supposed to bear some relation to an expanding or contracting universe, as illustrated in Fig. 1. The equation under consideration is

$$D^2x(t) + \omega^2x(t) = \tfrac{1}{2}\alpha x(t-\tau) + \tfrac{1}{2}\beta x(t+\sigma) + \psi(t).$$

We assume the ratio of σ to τ to be rational, so that without loss of generality σ and τ can be taken to be integers. Appropriate data for the method of steps for this equation are the values of $x(t)$ on an interval of length $\sigma + \tau$. As above we use two intervals, $[-\tau, 0]$ and $[N, N+\sigma]$. The quantities $x_n(t), \xi(t), \psi_n, R$, and L are defined in exactly the same way as before but the equations of motion now take a different form. Letting $x_n(t) = \phi(t+n-1)$, where $\phi(t)$ is the boundary value function given on the intervals $[-\tau, 0]$ and $[N, N+\sigma]$ we have $(0 \le t \le 1, n = 1, \ldots, N)$

$$(D^2 + \omega^2)\, x_n(t) = \tfrac{1}{2}\alpha x_{n-\tau}(t) + \tfrac{1}{2}\beta x_{n+\sigma}(t) + \psi_n(t). \qquad (19)$$

Defining

$$f(t) = \begin{pmatrix} \phi_{1-\tau} \\ \cdot \\ \cdot \\ \cdot \\ \phi_0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ \phi_{N-1} \\ \cdot \\ \cdot \\ \cdot \\ \phi_{N+\sigma} \end{pmatrix} + \begin{pmatrix} \psi_1 \\ \cdot \\ \cdot \\ \cdot \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \psi_N \end{pmatrix}, \quad \Omega^2 = \omega^2 1 - \frac{\alpha}{2}L^\tau - \frac{\beta}{2}R^\sigma,$$

we once again obtain the equation

$$(D^2 + \Omega^2)\,\xi(t) = f(t), \qquad 0 \le t \le 1.$$

The conditions ensuring continuity of $x(t)$ at $t = 0, \ldots, N$ and of $Dx(t)$ at $t = 1 \ldots, N-1$ are now *exactly* the same as before, namely Eqs. (12) and (13). The appropriate homogeneous solution is therefore obtained as above.

Having developed a method for solution of Eq. (1) we mention that there is another problem to which this can be applied. Dirac[11] obtains the radiation reaction term from a difference of retarded and advanced self interactions of an electron evaluated at the position of the electron. This difference of interaction is presumably a force felt by the electron, and the evaluation of the effective force, which is essentially what Dirac has done, requires solution of a differential equation with deviating arguments. Dirac instead writes $x(t-\sigma) = x(t) - \dot{x}(t)\sigma + \cdots$. A proper check of Dirac's deviation and this expansion might involve Eq. (1) with $\tau = \sigma$, $\beta = -\alpha$, $\omega = 0$ and $\tau \to 0$, $\alpha \to \infty$.

DISCUSSION

Although the mathematical results presented here are still some way from providing a test for the absorber theory of Wheeler and Feynman, we would like to mention just how such a test can be performed in the language of our paper. The function $\psi(t)$ of Eq. (1) is taken to be $\delta(t - t_0)$ for some $t_0 \in (0, T)$. The solution $x(t)$ then exhibits the effect of the perturbation ψ and in general x is disturbed both before and after t_0. If for some reason the disturbance of x due to ψ is restricted to $t \ge t_0$ (for some range of t_0) then causal behavior for the system is established and the advanced interaction could be said to have been eliminated. Then the strength of the retarded interaction should be checked for enhancement (as in the absorber theory). A possible origin of the suppression of the advanced interaction could be differences in the values of σ and τ, α and β due to expansion of the universe as suggested by Fig. 1.

Some preliminary numerical work has in fact been done along the lines described above. However, we feel that to obtain the desired absorption effect it will be necessary to introduce many individually acting absorbers, rather than replacing them all by a single effective interaction.

We have thus solved a boundary value problem for a linear system with advanced and retarded interactions. The author is aware of another problem in the theory of differential equations with deviating arguments where it has been found useful to use data on two intervals, namely in Titchmarsh's computation[12] of the coefficients in the Fourier type expansion $[x(t) \sim \sum a_n e^{s_n t}, s_n$ a solution of, e.g., $s^2 + \omega^2 = \alpha \cosh\tau s$ for Eq. (4)]. In that case, however, although use is made of data on two intervals, in fact each of these intervals alone would fix a solution. Thus the method of Titchmarsh implicitly assumes that the function in one of these intervals is the solution for the data in the other. Therefore this method is not an operational way for getting a solution from boundary values.

In Refs. 13 and 14 existence theorems are to be found for various boundary value problems for differential equations with deviating arguments. While no explicit solutions are given, these authors do not limit themselves to linear systems.

ACKNOWLEDGMENTS

The author is grateful to K. R. Subba Swami for performing many numerical computations, and to D. Lichtenberg, A. Mann, R. Newton, A. Peres, M. Revzen, and R. Weiner for interesting and useful conservations. He also wishes to thank the referee for his remarks on the electrodynamic boundary value problem and for making him aware of Refs. 13 and 14.

*Supported in part by U. S. Army Research Office (Durham)

[1]J. A. Wheeler and R. P. Feynman, Rev. Mod. Phys. **17**, 157 (1945); **21**, 425 (1949).
[2]J. E. Hogarth, Proc. Roy. Soc. **267A**, 365 (1962).
[3]F. Hoyle and J. V. Narliker, Proc. Roy. Soc. **277A**, 1 (1964).
[4]R. Bellman and K. L. Cooke, *Differential-Difference Equations* (Academic, New York, 1963).
[5]L. E. Elsgolts, *Introduction to the Theory of Differential Equations with Deviating Arguments* (Holden-Day, San Francisco, 1966).
[6]J. Hale, *Functional Differential Equations* (Springer-Verlag, Berlin, 1971).
[7]R. D. Driver, Arch. Rat. Mech. Anal. **19**, 149 (1965); Ann. Phys. **21**, 122 (1963); Phys. Rev. **178**, 2051 (1969).
[8]L. S. Schulman, Phys. Rev. D **7**, 2868 (1973).
[9]R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).
[10]Paul Hui, private communiation.
[11]P. A. M. Dirac, Proc. Roy Soc. London **A167**, 148 (1938).
[12]See Bellman and Cooke (Ref. 4) for exposition and references.
[13]P. Waltman and J. S. W. Wong, Trans. Amer. Math. Soc. **164**, 39 (1972).
[14]L. J. Grimm and K. Schmitt, Aeq. Math. **4**, 176 (1970).

Solutions of the steady, one-speed neutron transport equation for small mean free paths*

Edward W. Larsen

Courant Institute of Mathematical Sciences, New York University, New York, New York 10012
(Received 10 September 1973)

The solution of the steady, one-speed neutron transport equation with isotropic scattering and small mean free path is obtained. The solution is asymptotic with respect to a small parameter ϵ, defined as the mean free path in terms of a unit length of the same order of magnitude as a typical dimension of the domain. The solution, the leading two terms of which are given, consists of a boundary layer solution plus an interior solution. The boundary layer solution decays exponentially with distance from the boundary, the decay rate being proportional to ϵ^{-1}, and it shows the effects of boundary curvature and variations in the incoming flux along the boundary. The interior solution is a multiple of the source for subcritical domains, and depends on a diffusion equation for near critical domains. The boundary condition for the diffusion equation and an asymptotic criticality condition are derived.

1. INTRODUCTION

In a recent paper, Larsen and Keller[1] obtained solutions of general energy-dependent and one-speed neutron transport problems. The domains in these problems are assumed to be large compared to the mean free path of a neutron. The solutions are asymptotic with respect to a small parameter ϵ, which is a typical mean free path in terms of a unit length of the same order of magnitude as a typical dimension of the domain. The solutions consist of the sum of an interior solution which is either a multiple of the source or satisfies a diffusion equation, a boundary layer solution which decays exponentially with distance from the boundary, and an initial layer solution which decays exponentially with time from the initial time. The decay rates for the boundary and initial layer solutions are proportional to ϵ^{-1}.

In this paper we solve the transport problem considered in Ref. 1 for the special case of time independence and one-speed, isotropic scattering. For this case, we use a representation derived by Gibbs[2] for the neutron density to obtain more accurate results than those obtained in Ref. 1. We show that the solution obtained here is, to lowest order, asymptotically equivalent to that obtained in Ref. 1, and we explain why the new solution is more accurate. We also derive a boundary condition for the diffusion equation and obtain an explicit formula for the critical value of c. This critical value is shown to be equivalent to that obtained from a formula which uses an "extrapolated end point."

An outline of this paper follows. In Sec. 2 we state certain preliminary results and then give the neutron density to leading order in Theorems 1 and 2. In Sec. 3 we give the angular density to leading order. In Sec. 4 we obtain a criticality condition, and in Secs. 5 and 6 we prove Theorems 1 and 2. In Appendix A we outline Gibbs' method and in Appendix B we give the $O(\epsilon)$ correction terms to the neutron density.

2. THE ASYMPTOTIC REPRESENTATION OF ρ

Let D be an open convex domain with a smooth boundary. We assume that a steady, one-speed, isotropic neutron transport process occurs within D. Then the boundary value problem governing the angular distribution $\psi(\mathbf{r},\Omega)$ is:

$$\Omega\cdot\nabla\psi(\mathbf{r},\Omega) + (1/\epsilon)\psi(\mathbf{r},\Omega) = (c/4\pi\epsilon)\int\psi(\mathbf{r},\Omega')\,d\Omega' + q_0(\mathbf{r}), \quad \mathbf{r}\in D, \tag{2.1}$$

$$\psi(\mathbf{r}_0,\Omega) = \psi_0(\mathbf{r}_0,\Omega), \quad \mathbf{r}_0\in\partial D, \quad \Omega\cdot\mathbf{n}(\mathbf{r}_0) < 0. \tag{2.2}$$

In (2.1), ϵ is the mean free path in terms of a unit length of the same order of magnitude as a typical dimension of D. In (2.2), $\mathbf{n}(\mathbf{r}_0)$ is the unit outer normal and ψ_0 is prescribed.

The boundary value problem (2.1), (2.2) can be rewritten as an integral equation for the neutron density, $\rho(\mathbf{r}) = (4\pi)^{-1}\int\psi(\mathbf{r},\Omega')\,d\Omega'$. To do this, we define $d(\mathbf{r},\Omega)$ for $\mathbf{r}\in D$ to be the distance from $\mathbf{r}$ to ∂D in the direction $-\Omega$. Then $\mathbf{r} - d(\mathbf{r},\Omega)\Omega\in\partial D$ and the line $\mathbf{r} - t\Omega$ lies in D for $0\le t < d(\mathbf{r},\Omega)$. This line is a characteristic curve of (2.1), along which we integrate (2.1) to obtain

$$\psi(\mathbf{r},\Omega) = \psi_0[\mathbf{r} - d(\mathbf{r},\Omega)\Omega,\Omega]e^{-d(\mathbf{r},\Omega)/\epsilon} + \int_0^{d(\mathbf{r},\Omega)}\frac{e^{-t/\epsilon}}{\epsilon}[c\rho(\mathbf{r}-t\Omega) + \epsilon q_0(\mathbf{r}-t\Omega)]dt. \tag{2.3}$$

This equation defines ψ in terms of ρ. To obtain the integral equation for ρ, we integrate (2.3) over Ω and get

$$\rho(\mathbf{r}) = q_1(\mathbf{r}) + \frac{1}{4\pi}\int_{|\Omega|=1}\int_{t=0}^{d(\mathbf{r},\Omega)}\frac{e^{-t/\epsilon}}{\epsilon} \times [c\rho(\mathbf{r}-t\Omega) + \epsilon q_0(\mathbf{r}-t\Omega)]dt\,d\Omega, \tag{2.4}$$

where q_1 is defined by

$$q_1(\mathbf{r}) \equiv (1/4\pi)\int_{|\Omega|=1}\psi_0[\mathbf{r} - d(\mathbf{r},\Omega)\Omega,\Omega]e^{-d(\mathbf{r},\Omega)/\epsilon}\,d\Omega. \tag{2.5}$$

The solution ρ of (2.4) can be represented in the form

$$\rho(\mathbf{r}) = q_1(\mathbf{r}) + \int_\nu g(c,\nu)[Q_0(\epsilon\nu,\mathbf{r}) + Q_1(\epsilon\nu,\mathbf{r})]d\nu + \int_\nu H(\epsilon\nu,\mathbf{r})\,d\nu. \tag{2.6}$$

This is shown in Appendix A by Gibbs' procedure. In (2.6) we have used the notation

$$\int_\nu h(\nu)\,d\nu = \int_{\nu=0}^1 h(\nu)\,d\nu + h(\nu_0). \tag{2.7}$$

Here ν_0, with $\arg\nu_0 = 0$ or $\pi/2$, is the zero of the function $\lambda(z)$ defined by

$$\lambda(z) \equiv 1 - \frac{cz}{2}\int_{-1}^1\frac{ds}{z-s} = 1 - \frac{cz}{2}\ln\left(\frac{z+1}{z-1}\right). \tag{2.8}$$

The function $g(c,\nu)$ is defined by (A7) in Appendix A. The functions Q_j are particular solutions of the equation

$$(-\Delta + 1/\epsilon^2\nu^2)Q_j(\epsilon\nu,\mathbf{r}) = (c^j/\epsilon^{j+1})q_j(\mathbf{r}), \quad j = 0,1, \tag{2.9}$$

and H is a solution of the equation

$$(-\Delta + 1/\epsilon^2\nu^2)\,H(\epsilon\nu, \mathbf{r}) = 0. \tag{2.10}$$

Thus ρ is given by (2.6) once Q_0, Q_1 and H are found.

We have now converted the original boundary value problem for the integro-differential equation (2.1) into the problem of solving the differential equations (2.9) and (2.10), but the boundary values to go with these equations are not yet known. These boundary values must be correctly determined to ensure that (2.6) actually solves (2.4). Thus, we introduce (2.6) into (2.4) to obtain an equation for the determination of the boundary values. In this equation, any particular solutions of (2.9) can be chosen for Q_0 and Q_1, as is shown in Appendix A, and then only the boundary values of H need be found. This equation for the boundary values of H has been solved exactly only for the half-space. To solve it for other domains, we shall consider ϵ to be small and solve the equations asymptotically.

As a first step we shall assume that the boundary value of H are of the form

$$H(\epsilon\nu, \mathbf{r}_0) \sim \sum_{n=0}^{\infty} \epsilon^n B_n(\nu, \mathbf{r}_0), \qquad \mathbf{r}_0 \in \partial D. \tag{2.11}$$

Then formal asymptotic solutions of (2.10) and (2.11) can be found by seeking H of the form[3]

$$H(\epsilon\nu, \mathbf{r}) \sim e^{-S(\mathbf{r})/\epsilon\nu} \sum_{n=0}^{\infty} \epsilon^n H_n(\nu, \mathbf{r}). \tag{2.12}$$

We introduce this ansatz into (2.10) and equate the coefficients of different powers of ϵ to obtain

$$(\nabla S)^2 = 1, \tag{2.13}$$

$$2\nabla H_n \cdot \nabla S + H_n \Delta S \doteq \nu H_{n-1}, \qquad n \geq 0, \tag{2.14}$$

where $H_{-1} = 0$. To make H bounded in D and to satisfy the boundary conditions, we require

$$\begin{aligned} &S(\mathbf{r}_0) = 0, \qquad \mathbf{r}_0 \in \partial D, \\ &S(\mathbf{r}) > 0, \qquad \mathbf{r} \in D, \end{aligned} \tag{2.15}$$

and

$$H_n(\nu, \mathbf{r}_0) = B_n(\nu, \mathbf{r}_0), \qquad \mathbf{r}_0 \in \partial D. \tag{2.16}$$

The solution $S(\mathbf{r})$ of the boundary value problem (2.13), (2.15) is easily constructed. Let $\mathbf{r}$ be near ∂D. Then there is a unique point $\mathbf{r}_0 = \mathbf{r}_0(\mathbf{r}) \in \partial D$ which is nearest $\mathbf{r}$, and

$$\begin{aligned} &\nabla S(\mathbf{r}) = -\,\mathbf{n}(\mathbf{r}_0), \\ &S(\mathbf{r}) \;\; = |\mathbf{r} - \mathbf{r}_0|\,. \end{aligned} \tag{2.17}$$

$S(\mathbf{r})$ is thus the distance from $\mathbf{r}$ to ∂D.

Now we can solve the system (2.14), (2.16) recursively. To describe the solutions H_n, we let $\mathbf{r}$ and $\mathbf{r}_0$ be as above. We consider a Cartesian xyz-coordinate system in which $\mathbf{r}_0 = (0,0,0)$ and $\mathbf{r} = (0,0,t)$. We choose the x and y directions so that the equation for ∂D near $\mathbf{r}_0$ can be expressed in the form

$$z = \tfrac{1}{2}\alpha x^2 + \tfrac{1}{2}\beta y^2 + \cdots. \tag{2.18}$$

Then the functions H_n are defined recursively by

$$H_n(\nu, \mathbf{r}) = \frac{B_n(\nu, \mathbf{r}_0)}{[(1-\alpha t)(1-\beta t)]^{1/2}} + \tfrac{1}{2}\int_0^t \left[\frac{(1-\alpha s)(1-\beta s)}{(1-\alpha t)(1-\beta t)}\right]^{1/2} \cdot \Delta H_{n-1}[\nu, (0,0,s)]\,ds. \tag{2.19}$$

In addition, we set $n = 0$ in (2.19) and (2.14) to obtain the useful formula

$$\Delta S(\mathbf{r}) = [-(\alpha+\beta) + 2\alpha\beta t^2]/(1-\alpha t)(1-\beta t). \tag{2.20}$$

For t near $t^* = \min[\alpha^{-1}, \beta^{-1}]$, the solutions H_n become unbounded and the expansion (2.14) must be modified.[4] However, if ϵ is small enough, H will be exponentially small for t near $t^* - \delta$ and so, with exponentially small error, we can set $H = 0$ for $t > t^* - \delta$. In this paper, we shall assume that this is the case. Additional difficulties occur near corners or vertices of ∂D, if there are any. We shall not consider them here.

In the preceding analysis of Eqs. (2.10), we have assumed that $\epsilon\nu$ is small. This assumption is not valid for $\nu = \nu_0$ if c depends on ϵ and is close to 1. We find from an analysis of the equation $\lambda(\nu_0) = 0$ that if c depends upon ϵ in the manner

$$c = c(\epsilon) = 1 + \sum_{n=2} c_n \epsilon^n,$$

then $\epsilon\nu_0$ is of the form

$$\epsilon\nu_0 = \frac{1}{\sqrt{-3c_2}} + \sum_{n=1}^{\infty} b_n \epsilon^n.$$

For such a case, $\epsilon\nu_0$ is not small and we must modify statements made above about (2.10). The first step in this modification is to choose $c(\epsilon)$. We do this by requiring $\epsilon\nu_0$ to be a constant for all ϵ. This choice of $c(\epsilon)$ makes equation (2.10) independent of ϵ, which eliminates the necessity to expand $H(\epsilon\nu_0, \mathbf{r})$ in a series in ϵ.

Therefore, we define the constant γ by

$$\epsilon\nu_0 = 1/\gamma. \tag{2.21}$$

By assumption $\gamma = O(1)$, and from (2.8) we obtain

$$c = 2\epsilon\gamma\left[\ln\left(\frac{1+\epsilon\gamma}{1-\epsilon\gamma}\right)\right]^{-1} = 1 - \frac{\epsilon^2\gamma^2}{3} + O(\epsilon^4). \tag{2.22}$$

Equation (2.10) now has the form

$$(-\Delta + \gamma^2)\,H(1/\gamma, \mathbf{r}) = 0. \tag{2.23}$$

We shall express many of our results in terms of γ. In real problems, c and ϵ are known and then γ can be determined by inverting (2.22).

We have shown that the form of the solution of (2.10) depends upon whether ν_0 is real and $O(1)$, or $\epsilon\nu_0 = O(1)$. We shall call the first case "subcritical." Here $0 < c < 1$ and c is not near 1. We shall call the second case "near critical". Here c and ν_0 are defined in terms of ϵ and γ by (2.21) and (2.22), and for this case we shall only consider small interior sources and show this by replacing q_0 by ϵq_0 in all equations.

This concludes our discussion of the solutions of (2.10) in the subcritical and near critical cases. We can now give the asymptotic expansions of the right side of (2.6), and we shall prove the result later in Sec. 5.

Theorem 1: The neutron density $\rho(\mathbf{r})$ possesses an asymptotic expansion of the form

$$\rho(\mathbf{r}) \sim \sum_{n=0}^{\infty} \epsilon^n \rho_n(\mathbf{r}, \epsilon). \tag{2.24}$$

In the subcritical case, the leading term ρ_0 is given by

$$\rho_0(\mathbf{r}, \epsilon) = \tfrac{1}{2}\int_\nu e^{-S(\mathbf{r})/\epsilon\nu} A_0(\nu, \mathbf{r})\,d\nu, \tag{2.25}$$

where A_0 satisfies (2. 14) and (2. 19) with $n = 0$. In the near critical case,

$$\rho_0(\mathbf{r},\epsilon) = \tfrac{1}{2}\int_{\nu=0}^{1} e^{-S(\mathbf{r})/\epsilon\nu} A_0(\nu,\mathbf{r})\,d\nu + \tfrac{1}{2}A_0(\nu_0,\mathbf{r}) + 3Q_0(1/\gamma,\mathbf{r}). \quad (2.26)$$

For $0 < \nu < 1$, $A_0(\nu,\mathbf{r})$ satisfies (2. 14) and (2. 19) with $n = 0$. $A_0(\nu_0,\mathbf{r})$ satisfies (2. 23), and $Q_0(1/\gamma,\mathbf{r})$ is the solution of (2. 9) with $\nu = \nu_0 = 1/\epsilon\gamma$, q_0 replaced by ϵq_0, and satisfying $Q(1/\gamma,\mathbf{r}_0) = 0$ for $\mathbf{r}_0 \in \partial D$.

To complete the representation of ρ_0, we must determine boundary values for the functions A_0. We shall do this in Sec. 6 by introducing (2. 25) and (2. 26) into (2. 4) and then performing an asymptotic analysis of the resulting equations. However, we shall state the results here.

Theorem 2: Let the representations for ρ_0 in Theorem 1 hold. Then the boundary values of $A_0(\nu,\mathbf{r})$ for the subcritical and near critical cases are given by

$$A_0(\nu,\mathbf{r}_0) = \frac{\nu}{W(\nu)N(\nu)}\int_{\mu=0}^{1} W(\mu)\phi_\nu(\mu)J_0(\mu,\mathbf{r}_0)\,d\mu, \quad \mathbf{r}_0 \in \partial D. \quad (2.27)$$

In this equation, $\phi_\nu(\mu)$ are the one-speed, isotropic Case eigenfunctions, and $W(\mu)$ and $N(\nu)$ are the half-range weight factor and normalization function respectively.[5] The function J_0 is defined as follows. We let $\mathbf{r}_0 \in \partial D$ and we consider the Cartesian system with origin at $\mathbf{r}_0$ described above equation (2. 18). We define angular coordinates θ,η so that

$$\Omega = (\Omega_x,\Omega_y,\Omega_z) = (\cos\eta\,\sin\theta, \sin\eta\,\sin\theta, \cos\theta).$$

Then with ψ_0 the incoming flux as in (2. 2) and $\mu = \cos\theta$, $J_0(\mu,\mathbf{r}_0)$ is defined by

$$J_0(\cos\theta,\mathbf{r}_0) = (1/2\pi)\int_{\eta=0}^{2\pi}\psi_0(\mathbf{r}_0,\Omega)\,d\eta, \quad 0 \le \theta < \pi/2. \quad (2.28)$$

Theorems 1 and 2 then completely define ρ_0. We have determined ρ_1, and we describe it in Appendix B. The calculations to obtain ρ_1 are very lengthy but are based on the same procedure as that to obtain ρ_0. Therefore, we shall only show the derivation of ρ_0, and we do this in Sec. 5 and 6.

3. THE ASYMPTOTIC REPRESENTATION OF ψ

Here we shall insert the representations (2. 25)–(2. 27) into (2. 3) to obtain expansions for the angular density ψ. To do this, we shall refer to results proved in Secs. 5 and 6.

We first consider the subcritical case. We insert (2. 25) into (2. 3) and use Eq. (6. 5) to obtain

$$\psi(\mathbf{r},\Omega) = \psi_0(\mathbf{r}_0,\Omega)e_0^{-S(\mathbf{r})/\epsilon\mu} + \frac{c}{2}\int_\nu A_0(\nu,\mathbf{r})\frac{\nu}{\nu-\mu} \times (e^{-S(\mathbf{r})/\epsilon\nu} - e_0^{-S(\mathbf{r})/\epsilon\mu})\,d\nu + O(\epsilon),$$

where $\mathbf{r}_0$ is the point on ∂D closest to $\mathbf{r}$ and $\mu = \Omega\cdot\nabla S(\mathbf{r}) = -\Omega\cdot\mathbf{n}(\mathbf{r}_0)$. The function $e_0^{(\,)}$ is defined by (5. 8). We rearrange this equation and use (6. 10) and (2. 28) to obtain, for $\mathbf{r}$ near ∂D,

$$\psi(\mathbf{r},\Omega) = \left(\psi_0(\mathbf{r}_0,\Omega) - \frac{1}{2\pi}\int_{\eta'=0}^{2\pi}\psi_0(\mathbf{r}_0,\Omega)\,d\eta'\right)e_0^{-S(\mathbf{r})/\epsilon\mu} + \int_\nu A_0(\nu,\mathbf{r})\,\phi_\nu(\mu)\,e^{-S(\mathbf{r})/\epsilon\nu}\,d\nu + O(\epsilon). \quad (3.1)$$

In the near-critical case, we obtain near ∂D

$$\psi(\mathbf{r},\Omega) = \left(\psi_0(\mathbf{r}_0,\Omega) - \frac{1}{2\pi}\int_{\eta'=0}^{2\pi}\psi_0(\mathbf{r}_0,\Omega)\,d\eta'\right)e_0^{-S(\mathbf{r})/\epsilon\mu} + 3Q_0(1/\gamma,\mathbf{r}) + \int_{\nu=0}^{1} A_0(\nu,\mathbf{r})\,\phi_\nu(\mu)\,e^{-S(\mathbf{r})/\epsilon\nu} \times d\nu + A_0(\nu_0,\mathbf{r})\,\phi_{\nu_0}(\mu) + O(\epsilon). \quad (3.2)$$

Therefore, to lowest order, the angular density ψ near the boundary is the solution of a local half-space problem with half-range coefficients A_0 which depend upon $\mathbf{r}$. The dependence upon $\mathbf{r}$ is due to focusing caused by the curvature of the boundary.

In the deep interior of D we obtain from Appendix B, for the subcritical case,

$$\psi(\mathbf{r},\Omega) = \epsilon q_0(\mathbf{r})/(1-c) + O(\epsilon^2). \quad (3.3)$$

For the near critical case, we get

$$\psi(\mathbf{r},\Omega) = \tfrac{1}{2}A_0(\nu_0,\mathbf{r}) + 3Q_0(1/\gamma,\mathbf{r}) + O(\epsilon). \quad (3.4)$$

In the subcritical and near critical cases, the "continuous" modes $A_0(\nu,\mathbf{r})$ for $0 < \nu < 1$ form a boundary layer which decays exponentially with distance from ∂D. For the subcritical case, the mode $A_0(\nu_0,r)$ forms the remaining part of the boundary layer. As c increases, the decay rate of this mode becomes less and in the limit as $c \to 1 + O(\epsilon^2)$ in the near critical case, $A_0(\nu_0,\mathbf{r})$ satisfies the diffusion equation (2. 23). The boundary condition appropriate to this equation is given by (2. 27). One can see that this condition is not the usual "Marshak" boundary condition (Ref. 5, p. 155) which in our notation has the form

$$A_0(\nu_0,\mathbf{r}_0) = \int_{\mu=0}^{1}\mu J_0(\mu,\mathbf{r}_0)\,d\mu, \quad (3.5)$$

where J_0 is defined by (2. 28).

An approximate form of Eqs. (3. 1)–(3. 4) was derived by Larsen and Keller[1] using a different method. To obtain their equations, replace $A_0(\nu,\mathbf{r})$ by $A_0(\nu,\mathbf{r}_0)$ for $0 < \nu < 1$ and $\nu = \nu_0$ in Eq. (3. 1), and for $0 < \nu < 1$ in (3. 2). Also, in (3. 2), replace $c = c(\epsilon)$ defined in (2. 22) by $c = 1 - \tfrac{1}{3}\epsilon^2\gamma^2$. These changes are at worst $O(\epsilon)$, and so equations (3. 1)–(3. 4) are, to lowest order, equivalent to those given in Ref. 1.

The solutions given here are more accurate, however, because they show the effects of the boundary curvature and in the near critical case they use the exact value of γ, defined by (2. 27), rather than by the truncated $c = 1 - \tfrac{1}{3}\epsilon^2\gamma^2$. Also, the calculation of the $O(\epsilon)$ correction term in Appendix B leads to a more accurate criticality condition than that given in Ref. 1. We shall formulate this criticality condition next.

4. CRITICAL PROBLEMS

If a positive solution of (2. 4) exists with $q_1 = q_0 = 0$, then we say that D is "critical." In this case we expect D to be near critical. Thus by (2. 6), setting $q_1 = Q_0 = Q_1 = 0$ and replacing the symbol H by A, we obtain

$$\rho(\mathbf{r}) = \int_{\nu=0}^{1} A(\epsilon\nu,\mathbf{r})\,d\nu + A(1/\gamma,\mathbf{r}), \quad (4.1)$$

where

$$A(\epsilon\nu,\mathbf{r}) = e^{-S(\mathbf{r})/\epsilon\nu}[A_0(\nu,\mathbf{r}) + \epsilon A_1(\nu,\mathbf{r})] + O(\epsilon^2) \quad (4.2)$$

and

$$A(1/\gamma,\mathbf{r}) = A_0(\nu_0,\mathbf{r}) + \epsilon A_1(\nu_0,\mathbf{r}) + O(\epsilon^2). \quad (4.3)$$

Here $A(1/\gamma,\mathbf{r})$ and $A_n(\nu_0,\mathbf{r})$ satisfy (2. 23), $A(\epsilon\nu,\mathbf{r})$ for $0 < \nu < 1$ satisfies (2. 10), and $A_n(\nu,\mathbf{r})$ for $0 < \nu < 1$ satisfy (2. 14). By equation (2. 27) of Theorem 2,

$$A_0(\nu, \mathbf{r}_0) = 0, \quad \mathbf{r}_0 \in \partial D, \tag{4.4}$$

and, by Eq. (B4) of Appendix B,

$$A_1(\nu, \mathbf{r}_0) = \frac{\nu}{W(\nu)N(\nu)} \int_{\mu=0}^{1} W(\mu)\, \phi_\nu(\mu)\, \mu T(\mathbf{r}_0)\, d\mu$$
$$= -\,\mathbf{n}(\mathbf{r}_0)\cdot\nabla A_0(\nu_0, \mathbf{r}_0)\, l(\nu), \tag{4.5}$$

where

$$l(\nu) \equiv \tfrac{1}{2}\, \frac{\nu}{W(\nu)N(\nu)} \int_0^1 W(\mu)\, \phi_\nu(\mu)\, \mu\, d\mu. \tag{4.6}$$

Introducing these into (4. 3), we obtain

$$A(1/\gamma, \mathbf{r}_0) = \epsilon[-\,\mathbf{n}(\mathbf{r}_0)\cdot\nabla A_0(\nu_0, \mathbf{r}_0)\, l(\nu_0)] + O(\epsilon^2)$$
$$= -\,\epsilon l(\nu_0)\mathbf{n}(\mathbf{r}_0)\cdot\nabla A(1/\gamma, \mathbf{r}_0) + O(\epsilon^2).$$

Therefore, the asymptotic boundary condition appropriate to $A(1/\gamma, \mathbf{r})$ is

$$A(1/\gamma, \mathbf{r}_0) + \epsilon l(\nu_0)\mathbf{n}(\mathbf{r}_0)\cdot\nabla A(1/\gamma, \mathbf{r}_0) = 0, \quad \mathbf{r}_0 \in \partial D. \tag{4.7}$$

Thus $A(1/\gamma, \mathbf{r})$ is a solution of Eq. (2. 23) with the homogeneous boundary condition (4. 7). To describe such a solution, we consider the eigenvalue problem

$$\begin{aligned} &(-\Delta - \lambda I)\,\phi(\mathbf{r}) = 0, && \mathbf{r} \in D \\ &\phi(\mathbf{r}_0) + \epsilon l(\nu_0)\mathbf{n}(\mathbf{r}_0)\cdot\nabla\phi(\mathbf{r}_0) = 0, && \mathbf{r}_0 \in \partial D. \end{aligned} \tag{4.8}$$

The eigenvalues of (4. 8) are positive and denumerable and can be arranged in an increasing sequence: $0 < \lambda_0 < \lambda_1 \cdots$. Then a nontrivial solution of (2. 23), (4. 7) exists iff $\gamma^2 = -\lambda_n$. By (2. 22), the corresponding critical values of c are

$$c_n = \epsilon\sqrt{\lambda_n}/\arctan\epsilon\sqrt{\lambda_n} = 1 + \tfrac{1}{3}\epsilon^2\lambda_n + O(\epsilon^4). \tag{4.9}$$

The minimum such value is c_0, and the corresponding eigenfunction of (4. 8) is everywhere positive in D. Therefore, we expect the critical value of c to be given asymptotically by

$$c_{crit} = \epsilon\sqrt{\lambda_0}/\arctan\epsilon\sqrt{\lambda_0}, \tag{4.10}$$

where λ_0 is the minimum eigenvalue of (4. 8).

An equivalent criticality condition which employs the so-called "extrapolated endpoint" can be derived. Since

$$l(\nu_0) = \frac{1}{2}\, \frac{\int_0^1 W(\mu)\, \phi_{\nu_0}(\mu)\, \mu\, d\mu}{\int_0^1 W(\mu)\, \phi_{\nu_0}^2(\mu)\, d\mu} = \frac{\int_0^1 W(\mu)\, \mu\, d\mu}{\int_0^1 W(\mu)\, d\mu} + O(\epsilon),$$

and since $W(\mu)$ is of one sign, then $l(\nu_0) > 0$. Let D' be the domain whose boundary points $\mathbf{r}_0'$ are "extrapolated" from the boundary points $\mathbf{r}_0$ of D by $\mathbf{r}_0' = \mathbf{r}_0 + \epsilon l(\nu_0)\mathbf{n}(\mathbf{r}_0)$. Then D' properly contains D, and we consider the eigenvalue problem

$$\begin{aligned} &(-\Delta - \lambda I)\,\phi(\mathbf{r}') = 0, && \mathbf{r}' \in D' \\ &\phi(\mathbf{r}_0') = 0, && \mathbf{r}_0' \in \partial D'. \end{aligned} \tag{4.11}$$

If ϕ is a solution of (4. 11), then

$$0 = \phi[\mathbf{r}_0 + \epsilon l(\nu_0)\,\mathbf{n}(\mathbf{r}_0)]$$
$$= \phi(\mathbf{r}_0) + \epsilon l(\nu_0)\,\mathbf{n}(\mathbf{r}_0)\cdot\nabla\phi(\mathbf{r}_0) + O(\epsilon^2),$$

and so ϕ asymptotically satisfies (4. 8). Therefore, the eigenvalue problems (4. 8), (4. 11) are asymptotically equivalent. The eigenvalues of (4. 11) are positive and can be arranged in an increasing sequence: $0 < \lambda_0' < \lambda_1' \cdots$, and the corresponding critical values of c are given by (4. 9) with λ_n replaced by λ_n'. The minimum such value is c_0', and the corresponding eigenfunction of (4. 11) is positive. Thus we obtain the "extrapolated" criticality condition

$$c_{crit}' = \epsilon\sqrt{\lambda_0'}/\arctan\epsilon\sqrt{\lambda_0'}, \tag{4.12}$$

where λ_0' is the minimum eigenvalue of (4. 11). We expect $\lambda_0 \approx \lambda_0'$, and therefore the conditions (4. 10) and (4. 12) should be asymptotically equivalent.

Finally, if D is critical, then $\rho(\mathbf{r})$ is defined up to $O(\epsilon^2)$ by (4. 1). In this equation, $A(1/\gamma, \mathbf{r})$ is a positive solution of either of the eigenvalue problems (4. 8) or (4. 11) and $A(\epsilon\nu, \mathbf{r})$ is given by (4. 2). The boundary conditions for $A_n(\nu, \mathbf{r})$ are then given by (4. 4) and (4. 5), where we replace $A_0(\nu_0, \mathbf{r})$ by $A(1/\gamma, \mathbf{r})$ in the right side of (4. 5).

5. PROOF OF THEOREM 1

In this section we shall expand each of the terms in (2. 6) asymptotically for $\epsilon \ll 1$ and thereby obtain the asymptotic representations (2. 25)–(2. 26) of ρ. For clarity, we shall perform the work on each term as a separate lemma. We begin with $q_1(\mathbf{r})$, defined in (2. 5).

Lemma 1:

$$q_1(\mathbf{r}) = \tfrac{1}{2} \int_0^1 J(\epsilon\mu, \mathbf{r})\, d\mu + O(\epsilon), \tag{5.1}$$

where, for $0 < \mu < 1$, $J(\epsilon\mu, \mathbf{r})$ satisfies Eq. (2. 10) with ν replaced by μ. The boundary values of J are

$$J(\epsilon\mu, \mathbf{r}_0) = J_0(\mu, \mathbf{r}_0), \tag{5.2}$$

where J_0 is defined by equation (2. 28).

Proof: The proof will consist in asymptotically expanding the left and right sides of Eq. (5. 1) and observing that, to $O(\epsilon)$, the two sides are equal. We first consider $q_1(\mathbf{r})$, defined by (2. 5). We let $\mathbf{r}$ be near ∂D and $\mathbf{r}_0 = \mathbf{r}_0(\mathbf{r}) \in \partial D$ be the point nearest $\mathbf{r}$. Then the main contribution to the integral in (2. 5) comes from the direction $\Omega = -\,n(\mathbf{r}_0)$. Therefore we shall expand $d(\mathbf{r}, \Omega)$ and then the integrand in (2. 5) asymptotically about $\Omega = -\,\mathbf{n}(\mathbf{r}_0)$. To do this, we consider the Cartesian system described above Eq. (2. 18) and the angular coordinates θ, η described above (2. 28). Then the direction $\Omega = -\,\mathbf{n}(\mathbf{r}_0)$ corresponds to $\theta = 0$.

For $\Omega \approx -\,\mathbf{n}(\mathbf{r}_0)$, let (x, y, z) be the point on ∂D intersecting the line $\mathbf{r} - s\Omega$ with $s > 0$. Then

$$(x, y, z) = \mathbf{r} - d(\mathbf{r}, \Omega)\,\Omega$$
$$= (0, 0, S(\mathbf{r})) - d(\mathbf{r}, \Omega)(\cos\eta\, \sin\theta, \sin\eta\, \sin\theta, \cos\theta), \tag{5.3}$$

so

$$x/y = (\cos\eta)/(\sin\eta).$$

Thus if we define $r = (x^2 + y^2)^{1/2}$, then $x = r\cos\eta$ and $y = r\sin\eta$. Using this in (2. 18), we obtain

$$z = \tfrac{1}{2}(\alpha\cos^2\eta + \beta\sin^2\eta)\, r^2 + \cdots \equiv f_2(\mathbf{r}_0, \eta)\, r^2 + O(r^3). \tag{5.4}$$

Also, (5. 3) becomes

$$z = S(\mathbf{r}) - d(\mathbf{r}, \Omega)\cos\theta$$
$$r = \sin\theta\; d(\mathbf{r}, \Omega)$$

Eliminating r and z from these last three equations yields

$$d(\mathbf{r},\Omega)\cos\theta = S(\mathbf{r}) - f_2(\mathbf{r}_0,\eta)\sin^2\theta\, d^2(\mathbf{r},\Omega) + O(r^3).$$

This leads to the expansion

$$d(\mathbf{r},\Omega) = \frac{S(\mathbf{r})}{\cos\theta} - f_2(\mathbf{r}_0,\eta)\frac{\sin^2\theta}{\cos^3\theta}S^2(\mathbf{r}) + \cdots, \quad (5.5)$$

which is valid for $\theta \approx 0$, or for $\Omega \approx -\mathbf{n}(\mathbf{r}_0)$.

Now we introduce (5. 5) into (5. 3) to obtain

$$\mathbf{r} - d(\mathbf{r},\Omega)\Omega = -S(\mathbf{r})\tan\theta\,\Omega_2 + \cdots,$$

where $\Omega_2 = (\cos\eta, \sin\eta, 0)$.

Then, for $\theta \approx 0$,

$$\psi_0[\mathbf{r} - d(\mathbf{r},\Omega)\Omega,\Omega] = \psi_0(\mathbf{r}_0,\Omega) - S(\mathbf{r})(\tan\theta)\Omega_2 \cdot\nabla\psi_0(\mathbf{r}_0,\Omega) + \cdots. \quad (5.6)$$

We also obtain from (5. 5)

$$e^{-d(\mathbf{r},\Omega)/\epsilon} = e_0^{-S(\mathbf{r})/(\epsilon\cos\theta)}\left(1 + \frac{S^2(\mathbf{r})}{\epsilon}f_2(\mathbf{r}_0,\eta)\frac{\sin^2\theta}{\cos^3\theta}\right) + O(\epsilon^2). \quad (5.7)$$

In this equation, we have introduced the function

$$\epsilon_0^t = \begin{cases} e^t, & t \le 0 \\ 0, & t > 0 \end{cases}. \quad (5.8)$$

This makes (5. 7) valid for all angles.

Now we introduce (5. 6) and (5. 7) into (2. 5) and integrate to obtain

$$q_1(\mathbf{r}) = \tfrac{1}{2}\int_0^1 J_0(\mu,\mathbf{r})\, e^{-S(\mathbf{r})/\epsilon\mu}\, d\mu + O(\epsilon), \quad (5.9)$$

where J_0 is defined by (2. 28) and (2. 14) with $n = 0$.

Next we consider the right side of (5. 1). We showed in Sec. 2 that the solutions $J(\epsilon\mu,\mathbf{r})$ of (2. 10) with boundary conditions (5. 2) can be expanded in the form

$$J(\epsilon\mu,\mathbf{r}) = e^{-S(\mathbf{r})/\epsilon\mu}J_0(\mu,\mathbf{r}) + O(\epsilon),$$

where J_0 satisfies (2. 14) with $n = 0$, with boundary values given by (2. 28). Integrating this equation over μ and comparing the result to (5. 9) verifies equation (5. 1). This proves the lemma. QED

Lemma 2: There exist particular solutions Q_1 of (2. 9) such that

$$\int_\nu g(c,\nu)\,Q_1(\epsilon\nu,\mathbf{r})\,d\nu = \frac{1}{2}\int_{\mu=0}^1\left(\frac{\lambda(\mu)}{\lambda^2(\mu) + (\pi c\mu/2)^2} - 1\right)J(\epsilon\mu,\mathbf{r})\,d\mu + O(\epsilon), \quad (5.10)$$

where $J(\epsilon\mu,\mathbf{r})$ is defined in Lemma 1. This equation is valid for all values of c.

Proof: From the representation (5. 1), we obtain the following solution of (2. 9):

$$Q_1(\epsilon\nu,\mathbf{r}) = \frac{c}{2}\int_{\mu=0}^1 \frac{\nu^2\mu^2}{\mu^2-\nu^2}J(\epsilon\mu,\mathbf{r})\,d\mu + O(\epsilon). \quad (5.11)$$

For $0 < \nu < 1$, the integral is treated as a principal value. We multiply by $g(c,\nu)$ and "integrate" over ν to obtain

$$\int_\nu g(c,\nu)\,Q_1(\epsilon\nu,\mathbf{r})\,d\nu = \frac{1}{2}\int_{\mu=0}^1\left(c\int_\nu \frac{\nu^2\mu^2}{\nu^2-\mu^2}g(c,\nu)\,d\nu\right)J(\epsilon\mu,\mathbf{r})\,d\mu + O(\epsilon). \quad (5.12)$$

The inner integral is evaluated in (A12) in Appendix A, and its value is such that (5. 10) and (5. 12) agree. This completes the proof of the lemma. QED

Lemma 3: In the subcritical case, a solution Q_0 of (2. 9) exists such that

$$\int_\nu g(c,\nu)\,Q_0(\epsilon\nu,\mathbf{r})\,d\nu = \epsilon q_0(\mathbf{r})/(1-c) + O(\epsilon^3). \quad (5.13)$$

In the near critical case, every $O(1)$ solution Q_0 of (2. 9) (with q_0 replaced by ϵq_0) satisfies

$$\int_\nu g(c,\nu)\,Q_0(\epsilon\nu,\mathbf{r})\,d\nu = 3Q_0(1/\gamma,\mathbf{r}) + O(\epsilon^2). \quad (5.14)$$

Proof: We first consider the subcritical case. An asymptotic solution Q_0 of (2. 9), valid throughout D, is

$$Q_0(\epsilon\nu,\mathbf{r}) = \epsilon\nu^2 q_0(\mathbf{r}) + \epsilon^3\nu^4\Delta q_0(\mathbf{r}) + \cdots. \quad (5.15)$$

Therefore,

$$\int_\nu g(c,\nu)\,Q_0(\epsilon\nu,\mathbf{r})\,d\nu = \epsilon q_0(\mathbf{r})\int_\nu \nu^2 g(c,\nu)\,d\nu + O(\epsilon^3).$$

The integral on the right side is evaluated in Eq. (A13) of Appendix A, and we obtain (5. 13).

In the near critical case, we replace q_0 by ϵq_0 in all relevant equations. For $0 < \nu < 1$, we take $Q_0(\epsilon\nu,\mathbf{r})$ to be defined by (5. 15). Then these functions are $O(\epsilon^2)$. Let $Q_0(1/\gamma,\mathbf{r})$ be any solution of (2. 9) with $\nu = \nu_0 = 1/\epsilon\gamma$. Then

$$\int_\nu g(c,\nu)\,Q_0(\epsilon\nu,\mathbf{r})\,d\nu = g(c,\nu_0)\,Q_0(1/\gamma,\mathbf{r}) + O(\epsilon^2).$$

By (A14) of Appendix A, $g(c,\nu_0) = 3 + O(\epsilon^2)$. This verifies (5. 14) and completes the proof of the lemma. QED

The above lemmas can now be combined to prove Theorem 1.

Proof of Theorem 1: For the subcritical case, we combine equations (2. 6), (2. 12), and the above lemmas to obtain

$$\rho(\mathbf{r}) = \tfrac{1}{2}\int_\nu A(\epsilon\nu,\mathbf{r})\,d\nu + O(\epsilon), \quad (5.16)$$

where $A(\epsilon\nu,\mathbf{r})$ is a linear combination of $H(\epsilon\nu,\mathbf{r})$ and $J(\epsilon\nu,\mathbf{r})$, and so A satisfies Eq. (2. 10). We substitute the expansions (2. 12)–(2. 14) for A into (5. 16) to obtain (2. 25). Equation (2. 26) is obtained in the same way, where for notational convenience we have replaced $A(1/\gamma,\mathbf{r})$ in this equation by $A(\nu_0,\mathbf{r})$. This completes the proof of the theorem. QED

6. PROOF OF THEOREM 2

In this section we shall insert the representations for ρ in Theorem 1 into Eq. (2. 4) to obtain equations for the boundary values of the function $A_0(\nu,\mathbf{r})$. Equation (2. 4) can be written in the form

$$(I - cL)\rho(\mathbf{r}) = q_1(\mathbf{r}) + \epsilon Lq_0(\mathbf{r}), \quad (6.1)$$

where L is the operator

$$(L\omega)(\mathbf{r}) \equiv \frac{1}{4\pi}\int_{|\Omega|=1}\int_{t=0}^{d(\mathbf{r},\Omega)}\frac{e^{-t/\epsilon}}{\epsilon}\,\omega(\mathbf{r}-t\Omega)\,dt\,d\Omega. \quad (6.2)$$

From Theorem 1 we see that ρ and q_n can be represented in terms of two classes of functions ω for which

$L\omega$ has different asymptotic expansions. These classes ω can be characterized by:

(i) $\omega(\mathbf{r})$ is $O(1)$ and has $O(1)$ derivatives.

(ii) $\omega(\mathbf{r}) = e^{-S(\mathbf{r})/\epsilon\nu}\, h(\mathbf{r})$, where $h(\mathbf{r})$ satisfies (i).

We shall obtain the asymptotic expansion of $L\omega$ for these classes of functions. The analysis will be separated into two lemmas.

Lemma 4: Let $\omega(\mathbf{r})$ be $O(1)$ and have $O(1)$ derivatives. Then

$$(L\omega)(\mathbf{r}) = \omega(\mathbf{r}) - \frac{\omega(\mathbf{r})}{2} \int_{\mu=0}^{1} e^{-S(\mathbf{r})/\epsilon\mu}\, d\mu + O(\epsilon). \quad (6.3)$$

Proof: Using (5.5), (5.7), $\omega(\mathbf{r} - t\boldsymbol{\Omega}) = \omega(\mathbf{r}) - t\boldsymbol{\Omega}\cdot\nabla\omega(\mathbf{r}) + \cdots$, and setting $\mu = \cos\theta$, we obtain

$$\int_0^{d(\mathbf{r},\boldsymbol{\Omega})} \frac{e^{-t/\epsilon}}{\epsilon}\, \omega(\mathbf{r} - t\boldsymbol{\Omega})\, dt = \omega(\mathbf{r})\,[1 - e_0^{-S(\mathbf{r})/\epsilon\mu}] + O(\epsilon).$$

We integrate this equation over $\boldsymbol{\Omega}$ to obtain (6.3). This completes the proof of the lemma. QED

Lemma 5: Let $h(\mathbf{r})$ be $O(1)$ and have $O(1)$ derivatives. Also, let $\nu > 0$ be an $O(1)$ constant. Then

$$L(e^{-S(\mathbf{r})/\epsilon\nu}\, h(\mathbf{r})) = h(\mathbf{r}) \left(e^{-S(\mathbf{r})/\epsilon\nu}\, \frac{\nu}{2} \int_{-1}^{1} \frac{d\mu}{\nu - \mu} - \frac{\nu}{2} \int_0^1 e^{-S(\mathbf{r})/\epsilon\mu}\, \frac{d\mu}{\nu - \mu}\right) + O(\epsilon). \quad (6.4)$$

Proof: We proceed as in the previous lemma. We set $\mu = \boldsymbol{\Omega}\cdot\nabla S(\mathbf{r})$ to obtain $S(\mathbf{r} - t\boldsymbol{\Omega}) = S(\mathbf{r}) - t\mu + O(t^2)$ and then, using (5.7), we get

$$\begin{aligned}\int_0^{d(\mathbf{r},\boldsymbol{\Omega})} (e^{-t/\epsilon}/\epsilon)\, e^{-S(\mathbf{r}-t\boldsymbol{\Omega})/\epsilon\nu}\, h(\mathbf{r} - t\boldsymbol{\Omega})\, dt \\ = h(\mathbf{r})\, e^{-S(\mathbf{r})/\epsilon\nu} \int_0^{d(\mathbf{r},\boldsymbol{\Omega})} e^{-(t/\epsilon)(1-\mu/\nu)}\, dt + O(\epsilon) \\ = h(\mathbf{r})\,[\nu/(\nu - \mu)]\,[e^{-S(\mathbf{r})/\epsilon\nu} - e_0^{-S(\mathbf{r})/\epsilon\mu}] + O(\epsilon).\end{aligned} \quad (6.5)$$

We integrate this equation over $\boldsymbol{\Omega}$ to obtain (6.4). This completes the proof of the lemma. QED

Now we shall use these lemmas to obtain expansions for Lq_0 and $(I - cL)\rho$.

First, we use Eq. (5.9) as the expansion for q_1 and we use Lemma 4 to obtain an expansion for ϵLq_0. Adding these, we obtain

$$q_1(\mathbf{r}) + \epsilon Lq_0(\mathbf{r}) = \tfrac{1}{2} \int_0^1 e^{-S(\mathbf{r})/\epsilon\mu}\, J_0(\mu, \mathbf{r})\, d\mu + O(\epsilon). \quad (6.6)$$

This is the expansion we shall use for the right side of (6.1).

Next we shall expand $(I - cL)\rho$ for the subcritical case. From Lemma 5 we obtain, after "integrating" over ν and using (2.25),

$$(I - cL)\rho(\mathbf{r}) = \tfrac{1}{2} \int_{\mu=0}^{1} \left(\int_\nu A_0(\nu, \mathbf{r})\, \phi_\nu(\mu)\, d\nu\right) e^{-S(\mathbf{r})/\epsilon\mu} \times d\mu + O(\epsilon), \quad (6.7)$$

where $\phi_\nu(\mu)$ are the Case eigenfunctions,

$$\phi_\nu(\mu) = \lambda(\nu)\, \delta(\nu - \mu) + \tfrac{1}{2}\, c\nu(\nu - \mu)^{-1}. \quad (6.8)$$

It remains to compute the expansion of $(I - cL)\rho$ for the near critical case. Since in this case $c = 1 + O(\epsilon^2)$ and $\frac{1}{2} = \phi_{\nu_0}(\mu) + O(\epsilon)$, then we obtain, from Lemma 4 and $Q_0(1/\gamma, \mathbf{r}_0) = 0$,

$$\begin{aligned}(I - cL)\,[\tfrac{1}{2} A_0(\nu_0, \mathbf{r}) + 3Q_0(1/\gamma, \mathbf{r}] \\ = \tfrac{1}{2} A_0(\nu_0, \mathbf{r}) \int_0^1 \phi_{\nu_0}(\mu)\, e^{-S(\mathbf{r})/\epsilon\mu}\, d\mu + O(\epsilon).\end{aligned} \quad (6.9)$$

Now we combine Eqs. (2.26), (6.7) for $0 < \nu < 1$ only, and (6.9) to obtain exactly Eq. (6.7). Therefore, to leading order, the boundary conditions for the subcritical and near critical cases will have the same form. For both cases, then, we combine (6.1), (6.6), and (6.7) to obtain

$$\begin{aligned}&\int_\nu A_0(\nu, \mathbf{r}_0)\, \phi_\nu(\mu)\, d\nu = J_0(\mu, \mathbf{r}_0), \\ &0 < \mu < 1, \quad \mathbf{r}_0 \in \partial D.\end{aligned} \quad (6.10)$$

The solution $A_0(\nu, \mathbf{r}_0)$ of this equation is expressed in terms of the half-range formulas in Eq. (2.27). The boundary values of A_0 have thus been determined and Theorem 2 has been proved.

ACKNOWLEDGMENTS

The author thanks Professor Joseph B. Keller for suggesting the problem to him, and for some discussions and help in the preparation of the manuscript. He also acknowledges the work of Professor George J. Habetler and Professor Bernard J. Matkowsky,[6] who obtained many of the results in this paper for the one-dimensional slab.

APPENDIX A

In this appendix we shall outline the method used by Gibbs[2] to obtain the representation (2.6). We also shall derive certain formulas needed in Secs. 5 and 6.

First we rewrite Eq. (2.4) by making a change of variables to Cartesian coordinates. By letting $\mathbf{r}' = \mathbf{r} - t\boldsymbol{\Omega}$ and $dt\, d\boldsymbol{\Omega} = |\mathbf{r} - \mathbf{r}'|^{-2} d\mathbf{r}'$, (2.4) becomes

$$\rho(\mathbf{r}) = q_1(\mathbf{r}) + \frac{1}{\epsilon} \int_D \frac{e^{-|\mathbf{r}-\mathbf{r}'|/\epsilon}}{4\pi|\mathbf{r} - \mathbf{r}'|^2}\, [c\rho(\mathbf{r}') + \epsilon q_0(\mathbf{r}')]\, d\mathbf{r}' \quad (A1)$$

Now we use the fact that

$$\frac{e^{-|\mathbf{r}-\mathbf{r}'|/\epsilon}}{4\pi|\mathbf{r} - \mathbf{r}'|^2} = \frac{1}{\epsilon} \int_{\mu=0}^{1} G(\mathbf{r} - \mathbf{r}', \epsilon\mu)\, \frac{d\mu}{\mu^2}, \quad (A2)$$

where G satisfies

$$G(\mathbf{r} - \mathbf{r}', \epsilon\mu) = e^{-|\mathbf{r}-\mathbf{r}'|/\epsilon\mu}/4\pi\,|\mathbf{r} - \mathbf{r}'|$$

$$\times (-\Delta + 1/\epsilon^2\mu^2)\, G(\mathbf{r} - \mathbf{r}', \epsilon\mu) = \delta(\mathbf{r} - \mathbf{r}').$$

We introduce (A2) into (A1) and invert the order of integrations to obtain

$$\rho(\mathbf{r}) = q_1(\mathbf{r}) + \int_{\mu=0}^{1} F(\mathbf{r}, \mu)\, \frac{d\mu}{\mu^2}, \quad (A3)$$

where F is defined by

$$F(\mathbf{r}, \mu) = \epsilon^{-2} \int_D G(\mathbf{r} - \mathbf{r}', \epsilon\mu)\, [c\rho(\mathbf{r}') + \epsilon q_0(\mathbf{r}')]\, d\mathbf{r}'.$$

To obtain an equation for F, we operate on the above equation by $-\Delta + (\epsilon\mu)^{-2}$ and use (A3). The result is

$$-\Delta F(\mathbf{r}, \mu) + \epsilon^{-2} K^{-2} F(\mathbf{r}, \mu) = \epsilon^{-1} q_0(\mathbf{r}) + c\epsilon^{-2} q_1(\mathbf{r}), \quad (A4)$$

where the operator K^{-2} acts only on the μ variable, and for $c \neq 1$ the operator K^{+2} is found to be

$$(K^2\phi)(\mu) \equiv \mu^2 \left(\phi(\mu) + \frac{c}{1 - c} \int_0^1 \phi(s)\, ds\right). \quad (A5)$$

To obtain the general solution of (A4), we must derive the resolution of the identity of K^2. We find that for any function $h(\mu)$ which is Hölder continuous on $(0,1)$, one can write

$$h(\mu) = \int_\nu A(\nu) f(\nu,\mu)\, d\nu, \quad 0 < \mu < 1, \tag{A6}$$

where

$$f(\nu,\mu) = \mu^2 \lambda(\mu)\, \delta(\nu-\mu) + c\nu^2\mu^2/(\nu^2-\mu^2),$$

$$A(c,\nu) = g(c,\nu) \int_0^1 [f(\nu,\mu)/\mu^2] h(\mu)\, d\mu,$$

and

$$g(c,\nu) = \begin{cases} \nu^{-2}[\lambda^2(\nu) + (\frac{1}{2}\pi c\nu)^2]^{-1}, & 0 < \nu < 1, \\ (2/c\nu_0^4)\,[c/(\nu_0^2-1) - 1/\nu_0^2]^{-1}, & \nu = \nu_0. \end{cases} \tag{A7}$$

The eigenfunctions $f(\nu,\mu)$ are normalized to satisfy

$$1 = \int_0^1 \frac{f(\nu,\mu)}{\mu^2}\, d\mu. \tag{A8}$$

We remark that the operator in (A5) has been called K^2 because for h given by (A6), one has

$$(K^2 h)(\mu) = \int_\nu \nu^2 A(\nu) f(\nu,\mu)\, d\nu.$$

Now we obtain the solution of (A4) by writing F in the form

$$F(\mathbf{r},\mu) = \int_\nu L(\nu,\mathbf{r}) f(\nu,\mu)\, d\nu. \tag{A9}$$

Introducing this into (A4) and using $1 = \int_\nu g(c,\nu) f(\nu,\mu)\, d\nu$, we obtain

$$L(\nu,\mathbf{r}) = [Q_0(\epsilon\nu,\mathbf{r}) + Q_1(\epsilon\nu,\mathbf{r})]\, g(c,\nu) + H(\epsilon\nu,\mathbf{r}), \tag{A10}$$

where Q_0, Q_1, and H satisfy (2. 9) and (2. 10). Now we combine (A10), (A9), (A3), and (A8) to obtain the representation (2. 6).

Using the resolution of the identity of K^2, we can derive certain formulas used in Secs. 5 and 6.

For the function $\phi(\mu) = 1 = \int_\nu g(c,\nu) f(\nu,\mu)\, d\nu$, we compute, using (A5),

$$\frac{\mu^2}{z^2-\mu^2}\,\frac{1}{\lambda(z)} = (z^2 I - K^2)^{-1} K^2 \int_\nu g(c,\nu) f(\nu,\mu)\, d\nu = \int_\nu \frac{\nu^2}{z^2-\nu^2}\, g(c,\nu) f(\nu,\mu)\, d\nu.$$

We multiply this equation by $cz^2\mu^{-2}$, integrate over μ from 0 to 1, and use (A8) to obtain

$$c\int_\nu \frac{z^2\nu^2}{z^2-\nu^2}\, g(c,\nu)\, d\nu = \frac{1}{\lambda(z)} - 1. \tag{A11}$$

Now for $0 < \mu < 1$, we let $z \to \mu \pm i0$ and use the Plemelj formulas to obtain

$$c\int_\nu \frac{\mu^2\nu^2}{\mu^2-\nu^2}\, g(c,\nu)\, d\nu = \frac{\lambda(\mu)}{\lambda^2(\mu) + (\pi c\mu/2)^2} - 1. \tag{A12}$$

Also, we let $z \to \infty$ in (A11) and use $\lambda(\infty) = 1 - c$ to obtain

$$\int_\nu \nu^2 g(c,\nu)\, d\nu = 1/(1-c). \tag{A13}$$

Finally, we use (2. 21) and (2. 22) in (A7) to obtain

$$g(c,\nu_0) = 3 + O(\epsilon^2). \tag{A14}$$

APPENDIX B

In this Appendix we give the term $\rho_1(\mathbf{r},\epsilon)$ in the series (2. 24). The lengthy calculations to obtain ρ_1 involve carrying out the procedure given in Secs. 5 and 6 to $O(\epsilon)$ accuracy. The only difficulties occur with the term $q_1(\mathbf{r})$. Here one can show that Eq. (5. 1) holds to $O(\epsilon^2)$ if one adds the correction term $\epsilon J_1(\mu,\mathbf{r}_0)$ to the boundary condition (5. 2). J_1 is defined below. Also, one must use $\Delta S(\mathbf{r}_0) = -(\alpha+\beta)$, which is obtained from (2. 20). The calculation of Q_1 and the other terms proceed just as before. We now state the results.

In the subcritical case, we have

$$\rho_1(\mathbf{r},\epsilon) = \tfrac{1}{2}\int_\nu e^{-S(\mathbf{r})/\epsilon\nu} A_1(\nu,\mathbf{r})\, d\nu + q_0(\mathbf{r})/(1-c). \tag{B1}$$

Let $A_0(\nu,\mathbf{r})$ be the function defined in Theorems 1 and 2 for the subcritical case. Then A_1 in (B1) with A_0 satisfy (2. 14) and (2. 19) with $n = 1$. The boundary condition for A_1 is

$$A_1(\nu,\mathbf{r}_0) = \frac{\nu}{W(\nu)N(\nu)} \int_{\mu=0}^1 W(\mu)\,\phi_\nu(\mu) \times [J_1(\mu,\mathbf{r}_0) - q_0(\mathbf{r}_0)/(1-c)]\, d\mu. \tag{B2}$$

In the near critical case, we have

$$\rho_1(\mathbf{r},\epsilon) = \tfrac{1}{2}\int_{\nu=0}^1 e^{-S(\mathbf{r})/\epsilon\nu} A_1(\nu,\mathbf{r})\, d\nu + \tfrac{1}{2} A_1(\nu_0,\mathbf{r}), \tag{B3}$$

where for $0 < \nu < 1$, $A_1(\nu,\mathbf{r})$ is defined in terms of the near-critical $A_0(\nu,\mathbf{r})$ by (2. 14) and (2. 19) with $n = 1$. $A_1(\nu_0,\mathbf{r})$ satisfies Eq. (2. 23). The boundary condition for A_1 is

$$A_1(\nu,\mathbf{r}_0) = \frac{\nu}{W(\nu)N(\nu)} \int_{\mu=0}^1 W(\mu)\,\phi_\nu(\mu)[J_1(\mu,\mathbf{r}_0) + \mu T(\mathbf{r}_0)]\, d\mu, \tag{B4}$$

where

$$T(\mathbf{r}_0) = \tfrac{1}{2}\gamma A_0(\nu_0,\mathbf{r}_0) - \mathbf{n}(\mathbf{r}_0) \cdot \nabla[\tfrac{1}{2} A_0(\nu_0,\mathbf{r}_0) + 3Q_0(1/\gamma,\mathbf{r}_0)].$$

The function $J_1(\mu,\mathbf{r}_0)$ will now be described. Let $\mathbf{r}$, $\mathbf{r}_0$, θ, and η be as in Theorem 2. Let $\boldsymbol{\Omega}_2 = (\cos\eta, \sin\eta, 0)$ and let $f_2(\mathbf{r}_0,\eta) = \frac{1}{2}(\alpha\cos^2\eta + \beta\sin^2\eta)$. We define the functions

$$\Psi_1(\cos\theta,\mathbf{r}_0) = \frac{1}{2\pi}\int_{\eta=0}^{2\pi} \boldsymbol{\Omega}_2 \cdot \nabla\psi_0(\mathbf{r}_0,\boldsymbol{\Omega})\, d\eta,$$

$$\Psi_2(\cos\theta,\mathbf{r}_0) = \frac{1}{2\pi}\int_{\eta=0}^{2\pi} f_2(\mathbf{r}_0,\eta)\,\psi_0(\mathbf{r}_0,\boldsymbol{\Omega})\, d\eta.$$

Then J_1 is given by

$$J_1(\mu,\mathbf{r}_0) = \frac{d}{d\mu}\left(\mu(1-\mu^2)^{1/2}\,\Psi_1(\mu,\mathbf{r}_0) + \mu^2\frac{d}{d\mu}\left(\frac{1-\mu^2}{\mu} \times \Psi_2(\mu,\mathbf{r}_0)\right) + \frac{\alpha+\beta}{4}\,\mu^2 J_0(\mu,\mathbf{r}_0)\right), \tag{B5}$$

where J_0 is defined by (2. 28).

*The research was partially supported by the Air Force Office of Scientific Research under Grant No. AFOSR-71-2013 and by a National Science Foundation Institutional Grant to New York University.

[1] E. W. Larsen and J. B. Keller, J. Math. Phys. 15, 75 (1974).
[2] A. G. Gibbs, J. Math. Phys. 10, 875 (1969).
[3] R. M. Lewis and J. B. Keller, "Asymptotic Methods for Partial Differential Equations: The Reduced Wave Equation and Maxwell's Equation," New York Univ. Inst. Math. Research Rep. EM-194 (Jan. 1964).
[4] D. Ludwig, Comm. Pure Appl. Math. 19, 215 (1966).
[5] K. M. Case and P. F. Zweifel, *Linear Transport Theory* (Addison-Wesley, Reading, Mass., 1967), p. 86.
[6] G. J. Habetler and B. J. Matkowsky, "On Boundary Value Problems in Diffusion Theory as Limits of Boundary Value Problems in Transport Theory," 1971, unpublished.

Spinor calculus in five-dimensional relativity

Elhanan Leibowitz

Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada
(Received 22 June 1973)

A consistent spinor calculus is developed within the framework of five-dimensional relativity. The formalism is manifestly five-covariant, and in special coordinate systems and special spin frames it reduces to a familiar spinor formalism in curved space–time. The five-dimensional formulation is free of the difficulties involving the coupling of the electromagnetic field, which characterize the four-dimensional approach to spinor calculus. Some theorems and useful identities in the five-dimensional spinor calculus are proved.

1. INTRODUCTION

If the idea, first proposed by Kaluza[1] and subsequently developed by Klein,[2] of formulating the laws of nature in the framework of a five-dimensional Riemannian manifold is taken as a genuine physical approach rather than merely a mathematical curiosity, it becomes necessary to incorporate all physical fields into the five-dimensional picture. As spinor fields have proved to play a major role in the description of physical reality, the question arises whether it is possible to incorporate these fields into the five-dimensional formalism. It is the purpose of the present work to answer this question in the affirmative, and to develop a consistent five-dimensional two-component spinor calculus.

The theory of spinors in four-dimensional space is based upon the homomorphism between the group of Lorentz transformations $L_+^\uparrow$ and the group of unimodular linear transformations $SL(2,C)$. The significance of Lorentz transformations persists in the five-dimensional theory, because one direction in 5-space is singled out by the isometry of the space. Hence it is to be expected that a covariant spinor formalism can in fact be introduced. The problem of the construction of such a formalism has been tackled by Schmutzer[3] in the framework of projective relativity. Since his discussion is confined to spinor algebra, and does not touch upon spinor analysis, the full four-dimensional meaning of his approach has not been evaluated. It seems, however, that his spinor algebra is consistent with the one introduced here. From a different point of view initial steps in establishing a five-dimensional 4-component spinor calculus have also been investigated by Littlewood.[4]

The spinor calculus developed in the present work is free of the complications which are introduced into the standard four-dimensional spinor formalism in order to account for the coupling of the electromagnetic field. In the present approach the coupling is automatically accomplished by requiring periodicity of five-dimensional spinor fields. This point will be demonstrated in a subsequent paper,[5] where the spinor calculus developed here will be applied to simple spinor field equations.

2. FIVE-DIMENSIONAL GEOMETRY

The fundamentals of five-dimensional relativity theory were given in a recent paper.[6] The results obtained there which are needed for the present discussion will be summarized briefly.

The underlying manifold V_5 is a five-dimensional hyperbolic Riemannian space with signature -3, admitting an isometry defined by a spacelike vector field ξ^μ satisfying

$$\mathcal{L}_\xi \gamma_{\mu\nu} \equiv \xi_{\mu\|\nu} + \xi_{\nu\|\mu} = 0, \tag{2.1}$$

$$\xi_\mu \xi^\mu = -1. \tag{2.2}$$

Here $\gamma_{\mu\nu}$ is the five-dimensional metric tensor,[7] $\mathcal{L}_\xi$ denotes the Lie derivative in the direction of the vector ξ^μ, and a double stroke denotes five-dimensional covariant differentiation.

A family of five-dimensional coordinate systems, called "ξ systems," is defined by the conditions

$$\xi^5 = 1, \quad \xi^k = 0. \tag{2.3}$$

The most general transformation between ξ systems is given by

$$\bar{x}^5 = x^5 + \Lambda(x^l), \tag{2.4a}$$

$$\bar{x}^k = \bar{x}^k(x^l), \tag{2.4b}$$

where Λ and $\bar{x}^k$ are arbitrary functions of $x^1, \ldots, x^4$. In ξ systems the components of the 5-metric tensor and the ξ field are related to the components of the four-dimensional (space-time) metric tensor g_{kl} and the electromagnetic potential vector ϕ_k according to

$$\left.\begin{aligned} &\gamma_{55} = -1, \quad \gamma^{55} = -1 + \phi_k\phi^k, \\ &\gamma_{5k} = \phi_k, \quad \gamma^{5k} = \phi^k, \\ &\gamma_{kl} = g_{kl} - \phi_k\phi_l, \quad \gamma^{kl} = g^{kl}, \\ &\xi_5 = -1, \quad \xi_k = \phi_k. \end{aligned}\right\} \tag{2.5}$$

Under the transformation (2.4b) ϕ_k transforms like a 4-vector, while under the transformation (2.4a) it undergoes a gauge transformation

$$\bar{\phi}_k = \phi_k + \Lambda_{,k}, \tag{2.6}$$

where a comma denotes ordinary partial differentiation.

Using (2.5), one can calculate the five-dimensional Christoffel symbols $\Gamma^\lambda_{\mu\nu}$ in a ξ system. One obtains

$$\begin{aligned} &\Gamma^\mu_{55} = 0, \\ &\Gamma^5_{5k} = \tfrac{1}{2}\phi^r F_{rk}, \\ &\Gamma^5_{kl} = -\tfrac{1}{2}(\phi_{k;l} + \phi_{l;k}) - \tfrac{1}{2}\phi^r(\phi_k F_{rl} + \phi_l F_{rk}), \\ &\Gamma^k_{5l} = \tfrac{1}{2}F^k{}_l, \\ &\Gamma^k_{lm} = \left\{{k \atop lm}\right\} - \tfrac{1}{2}(\phi_l F^k{}_m + \phi_m F^k{}_l), \end{aligned} \tag{2.7}$$

where $\left\{{k \atop lm}\right\}$ is the four-dimensional Christoffel symbol,

$$F_{kl} = \phi_{k,l} - \phi_{l,k}$$

is the electromagnetic field tensor, and a semicolon denotes a 4-covariant derivative.

3. SURVEY OF SPINOR CALCULUS IN FOUR-DIMENSIONAL CURVED SPACE-TIME

The theory of spinors in curved space-time has been discussed by many authors.[8-10] In the present work we will mainly follow the notation of Ref. 8. The fundamentals of this theory which are needed for the present discussion will be outlined briefly.

Two-component contravariant spinors of the first rank

$$\psi^A(x^k), \quad A = 1, 2,$$

are elements of a complex two-dimensional linear vector space defined at every point of space-time. Under change of the spin-frame they transform according to

$$\psi'^A = S^A{}_B \psi^B,$$

where $S^A{}_B(x)$ is a unimodular complex 2×2 matrix. The conjugate spinor $\psi^{\dot{A}}$ transforms according to

$$\psi'^{\dot{A}} = S^{\dot{A}}{}_{\dot{B}} \psi^{\dot{B}}, \quad S^{\dot{A}}{}_{\dot{B}} = (S^A{}_B)^*.$$

First-rank covariant spinors χ_A transform according to

$$\chi'_A = \chi_B S^{-1B}{}_A, \quad \chi'_{\dot{A}} = \chi_{\dot{B}} S^{-1\dot{B}}{}_{\dot{A}}.$$

Higher rank spinors transform like products of first-rank spinors.

Spinor indices are raised and lowered by the fundamental antisymmetric spinor ϵ_{AB}, e.g.,

$$\psi^A = \epsilon^{AB}\psi_B, \quad \psi_A = \psi^B \epsilon_{BA}$$
$$\psi^{\dot{A}} = \epsilon^{\dot{A}\dot{B}}\psi_{\dot{B}} \quad \psi_{\dot{A}} = \psi^{\dot{B}}\epsilon_{\dot{B}\dot{A}},$$

with

$$\epsilon_{AB} = -\epsilon_{AB}, \quad \epsilon^{AB} = -\epsilon^{BA},$$
$$\epsilon_{12} = \epsilon_{\dot{1}\dot{2}} = \epsilon^{12} = \epsilon^{\dot{1}\dot{2}} = 1, \tag{3.1}$$
$$\epsilon_{AC}\epsilon^{BC} = \delta^B_A.$$

There exists a one-to-one correspondence between vectors T_k and Hermitian second-rank spinors $T_{\dot{A}B}$, viz.,

$$\left.\begin{aligned} T_{\dot{A}B} &= \sigma^k{}_{\dot{A}B} T_k \\ T_k &= \sigma_k{}^{\dot{A}B} T_{\dot{A}B} \end{aligned}\right\}, \quad T_{\dot{A}B} = T_{B\dot{A}},$$

where the "mixed quantities" $\sigma^k{}_{\dot{A}B}$ are four Hermitian matrices satisfying the algebraic relations

$$\sigma^k{}_{\dot{A}B}\sigma^{l\dot{A}B} = g^{kl},$$
$$\sigma^k{}_{\dot{A}B}\sigma_k{}^{\dot{C}D} = \delta^{\dot{C}}_{\dot{A}}\delta^D_B,$$
$$\sigma^{k\dot{C}A}\sigma^l{}_{\dot{C}B} + \sigma^{l\dot{C}A}\sigma^k{}_{\dot{C}B} = g^{kl}\delta^A_B.$$

The mixed quantities $\sigma^k{}_{\dot{A}B}$ are determined by the algebraic relations up to 6 real parameters (corresponding to Lorentz transformations).[8]

At an arbitrary point of space-time a geodesic coordinate system and a spin frame can be chosen in such a way that the first derivatives of the metric tensor and the first derivatives of the mixed quantities vanish at the point, and furthermore at the same point the metric tensor takes the form $\eta_{kl} = \mathrm{diag}[-1, -1, -1, 1]$, while the values of the mixed quantities reduce to the Pauli matrices[8]

$$\sigma^{(1)\dot{A}B} = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \sigma^{(2)\dot{A}B} = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix},$$
$$\sigma^{(3)\dot{A}B} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \sigma^{(4)\dot{A}B} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$\sigma_{(1)}{}^{\dot{A}B} = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}, \quad \sigma_{(2)}{}^{\dot{A}B} = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix}, \tag{3.2}$$
$$\sigma_{(3)}{}^{\dot{A}B} = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \sigma_{(4)}{}^{\dot{A}B} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The covariant derivative of a spinor is defined by

$$\psi_{A;k} = \psi_{A,k} - \left\{ \begin{matrix} B \\ A \; k \end{matrix} \right\} \psi_B,$$
$$\psi^A{}_{;k} = \psi^A{}_{,k} + \left\{ \begin{matrix} A \\ B \; k \end{matrix} \right\} \psi^B.$$

together with the usual properties of differentiation (linearity, reality, and Leibnitz rule), where the matrices $\left\{ \begin{matrix} A \\ B \; k \end{matrix} \right\}$ ("spin connections") are to be determined by additional conditions imposed on the covariant differentiation. It is natural and common to require

$$\sigma^k{}_{\dot{A}B;l} = 0, \tag{3.3}$$

which implies

$$\left\{ \begin{matrix} A \\ B \; k \end{matrix} \right\} = \tfrac{1}{2}\sigma_l{}^{\dot{C}A}\left(\sigma^l{}_{\dot{C}B,k} + \left\{ \begin{matrix} l \\ r \; k \end{matrix} \right\}\sigma^r{}_{\dot{C}B}\right) - \tfrac{1}{2}\left\{ \begin{matrix} C \\ C \; k \end{matrix} \right\}\delta^A_B$$
$$= \tfrac{1}{2}\sigma_l{}^{\dot{C}A}\left(\sigma^l{}_{\dot{C}B,k} + \left\{ \begin{matrix} l \\ r \; k \end{matrix} \right\}\sigma^r{}_{\dot{C}B}\right) + \tfrac{1}{2}\epsilon^{AC}\epsilon_{BC;k}. \tag{3.4}$$

The assumption (3.3) is made by most authors. From here two distinct approaches are possible. Firstly, the spinor formalism attains its simplest form with the additional requirement

$$\epsilon_{AB;k} = 0, \quad \epsilon^{AB}{}_{;k} = 0. \tag{3.5}$$

It turns out, however, that with this assumption it is difficult to incorporate the electromagnetic field into spinor field equations. This led Infeld and van der Waerden,[11] followed by most authors, to develop the second spinor formalism. According to their formalism (3.5) is not assumed, but the traces $\left\{ \begin{matrix} A \\ A \; k \end{matrix} \right\}$, which are left undetermined by (3.4), are interpreted as being related to the electromagnetic potentials. However, this approach has the disadvantage, in addition to the obvious inconveniences associated with raising and lowering spinor indices, of requiring different forms for particles differing in their electric charge (unless a more complicated formalism is utilized, using the concepts of "spinor densities"). It will be shown in a subsequent paper[5] that one of the advantages of the five-dimensional approach is that the proper coupling of the electromagnetic field is automatically achieved by the periodicity of the spinor fields, so that there is no need for the complicated version of the spinor formalism. We therefore retain the simpler spinor analysis emerging from the assumptions (3.3) and (3.5). We then have

$$\left\{ \begin{matrix} A \\ B \; k \end{matrix} \right\} = \tfrac{1}{2}\sigma_l{}^{\dot{C}A}\left(\sigma^l{}_{\dot{C}B,k} + \left\{ \begin{matrix} l \\ r \; k \end{matrix} \right\}\sigma^r{}_{\dot{C}B}\right),$$

from which it follows

$$\left\{ \begin{matrix} A \\ A\ k \end{matrix} \right\} = 0.$$

The commutation rule for the covariant derivatives is

$$\psi^A{}_{;kl} - \psi^A{}_{;lk} = P^A{}_{Bkl}\psi^B,$$

where the curvature spinor $P^A{}_{Bkl}$ is given by

$$P^A{}_{Bkl} = \left\{ \begin{matrix} A \\ B\ k \end{matrix} \right\}_{,l} - \left\{ \begin{matrix} A \\ B\ l \end{matrix} \right\}_{,k} + \left\{ \begin{matrix} A \\ C\ l \end{matrix} \right\}\left\{ \begin{matrix} C \\ B\ k \end{matrix} \right\} - \left\{ \begin{matrix} A \\ C\ k \end{matrix} \right\}\left\{ \begin{matrix} C \\ B\ l \end{matrix} \right\},$$

and satisfies the relations

$$P_{ABkl} = -P_{ABlk} = P_{BAkl},$$

$$P^A{}_{Akl} = 0,$$

$$P^A{}_{Bkl} = \tfrac{1}{2}\sigma_r{}^{\dot{C}A}\sigma^s{}_{\dot{C}B}R^r{}_{skl}$$

($R^r{}_{skl}$ being the curvature tensor[12]),

$$\sigma^{l\dot{A}C}P^B{}_{C kl} - \sigma^{l\dot{C}B}P^{\dot{A}}{}_{\dot{C}kl} = 0,$$

and

$$\sigma^{k\dot{C}A}\sigma^{l}{}_{\dot{C}D}P^D{}_{Bkl} = -\tfrac{1}{4}R\delta^A_B,$$

$$\sigma^k{}_{\dot{C}A}\sigma^{l\dot{C}D}P^B{}_{Dkl} = +\tfrac{1}{4}R\delta^B_A.$$

4. FIVE-DIMENSIONAL SPINOR ALGEBRA

At every point of V_5, a two-dimensional linear vector space ("spin space") is defined. This space is spanned by complex two-component quantities

$$\Psi^A(x^\mu), \quad A = 1, 2$$

which are functions of the five coordinates. These are called first-rank contravariant spinors. They transform in a similar way to the four-dimensional spinors

$$\Psi'^A = S^A{}_B(x^\mu)\Psi^B, \quad \det(S^A{}_B) = 1, \tag{4.1}$$

where the unimodular matrix $S^A{}_B$ which determines the spin frame is now a function of the five coordinates. Covariant spinors X_A transform according to

$$X'_A = X_B S^{-1}{}^B{}_A.$$

Similar rules apply for dotted spinor indices. The fundamental antisymmetric spinor is defined by (3.1).

A vector Υ_μ orthogonal to ξ_μ, defined at a point in V_5 is, like a Hermitian spinor, determined by four real numbers. We can map this vector onto a Hermitian spinor whose components are linear functions of the vector components, viz.,

$$\Upsilon_{\dot{A}B} = \sigma^\mu{}_{\dot{A}B}\Upsilon_\mu, \quad \Upsilon_\mu\xi^\mu = 0, \quad \Upsilon_{\dot{A}B} = \Upsilon_{B\dot{A}}. \tag{4.2}$$

The "mixed quantities"[13] $\sigma^\mu{}_{\dot{A}B}$ are five Hermitian matrices, determined by 20 real parameters. [There also exists a correspondence between the bivectors in V_5 orthogonal to ξ_μ and the symmetric spinors. To the bivector

$$\Upsilon_{\mu\nu} = -\Upsilon_{\nu\mu}, \quad \Upsilon_{\mu\nu}\xi^\nu = 0$$

there corresponds the symmetric spinor

$$\Upsilon^{(s)}{}_{A}{}_{B} = \sigma^{\mu\dot{C}A}\sigma^\nu{}_{\dot{C}B}\Upsilon_{\mu\nu},$$

$$\Upsilon^{(s)}{}_{AB} = \Upsilon^{(s)}{}_{BA}.$$

In Appendix B a somewhat different method to relate tensors to spinors is introduced.]

There is a great deal of freedom in the choice of the mixed qualities $\sigma^\mu{}_{\dot{A}B}$, and simple algebraic constraints will now be imposed in order to restrict this freedom. With the aid of the ξ field four independent real conditions can be imposed, namely

$$\sigma^\mu{}_{\dot{A}B}\xi_\mu = 0. \tag{4.3}$$

The other set of constraints is similar to the one imposed in the four-dimensional case: the correspondence (4.2) is required to preserve the "scalar products"

$$\Upsilon_{\dot{A}B}\Upsilon^{\dot{A}B} = \Upsilon_\mu\Upsilon^\mu. \tag{4.4}$$

This yields

$$(\sigma^\mu{}_{\dot{A}B}\sigma^{\nu\dot{A}B} - \gamma^{\mu\nu})\Upsilon_\mu\Upsilon_\nu = 0 \tag{4.5}$$

identically for any vector Υ_μ orthogonal to ξ_μ. In particular, (4.5) must hold for

$$\Upsilon_\mu = g_\mu{}^\nu\eta_\nu,$$

where[14]

$$g_{\mu\nu} = \gamma_{\mu\nu} + \xi_\mu\xi_\nu \tag{4.6}$$

and η_ν is an arbitrary vector in V_5. Hence

$$(\sigma^\mu{}_{\dot{A}B}\sigma^{\nu\dot{A}B} - g^{\mu\nu})\eta_\mu\eta_\nu = 0$$

for arbitrary η_μ.

The condition (4.4) is therefore equivalent to the relation

$$\sigma^\mu{}_{\dot{A}B}\sigma^{\nu\dot{A}B} = g^{\mu\nu} \tag{4.7}$$

Equations (4.3) and (4.7) together constitute a system of 14 independent relations among the 20 real parameters of $\sigma^\mu{}_{\dot{A}B}$, and we are left with six degrees of freedom. By using (4.7) the mapping (4.2) can be inverted

$$\Upsilon_\mu = \sigma_\mu{}^{\dot{A}B}\Upsilon_{\dot{A}B},$$

and when this is substituted into (4.2) we find

$$(\sigma^\mu{}_{\dot{A}B}\sigma_\mu{}^{\dot{C}D} - \delta^{\dot{C}}_{\dot{A}}\delta^D_B)\Upsilon_{\dot{C}D} = 0$$

for an arbitrary Hermitian spinor $\Upsilon_{\dot{C}D}$. Hence we have the relation

$$\sigma^\mu{}_{\dot{A}B}\sigma_\mu{}^{\dot{C}D} = \delta^{\dot{C}}_{\dot{A}}\delta^D_B. \tag{4.8}$$

A summary of the algebraic relations obtained so far is given in Appendix A. These relations imply a generalized "anticommutation rule," as can be shown in the following way. If

$$\delta^{A_1\ldots A_m}_{B_1\ldots B_m} \equiv \det(\delta^{A_i}_{B_j})$$

is the generalized Kronecker delta, then as the spinor indices attain two values only, we have identically

$$\delta^{ABC}_{DEF} \equiv 0.$$

From this we get

$$0 = \delta^{ABC}_{DEF}\sigma^{\mu\dot{H}D}\sigma^\nu{}_{\dot{H}C}\epsilon^{EF}$$

$$= 2[(\sigma^{\mu\dot{H}A}\epsilon^{BF} - \sigma^{\mu\dot{H}B}\epsilon^{AF})\sigma^\nu{}_{\dot{H}F} + g^{\mu\nu}\epsilon^{AB}],$$

where the right-hand side is obtained by expanding the Kronecker delta and using (4.8).

Hence

$$\sigma^{\mu}{}_{\dot{C}A}\sigma^{\nu}{}_{\dot{C}B} + \sigma^{\nu}{}_{\dot{C}A}\sigma^{\mu}{}_{\dot{C}B} = g^{\mu\nu}\delta^{A}_{B}. \tag{4.9}$$

So far no four-dimensional structure has been extracted from the five-dimensional spinor formalism. In order to do that, we first note that in ξ systems the relation (4.3) takes the form

$$\sigma_5{}^{\dot{A}B} = 0,$$

and it appears that the four matrices $\sigma_k{}^{\dot{A}B}$ can be related to the four-dimensional mixed quantities. However, in general $\sigma_k{}^{\dot{A}B}$ will depend on all the five coordinates. A concept in the spin space, similar to the concept of ξ system in the Riemannian space, proves to be useful here.

Definition: A spin-frame is called a "ξ frame" if the following relation holds in that frame:

$$\sigma_{\mu}{}^{\dot{A}B}{}_{,\nu}\xi^{\nu} + \sigma_{\nu}{}^{\dot{A}B}\xi^{\nu}{}_{,\mu} = 0. \tag{4.10}$$

Note that the condition (4.10) is invariant under a general coordinate transformation in V_5, as can easily be seen by writing it in an equivalent form

$$(\sigma_{\mu}{}^{\dot{A}B}{}_{,\nu} - \Gamma^{\lambda}_{\mu\nu}\sigma_{\lambda}{}^{\dot{A}B})\xi^{\nu} + \sigma_{\nu}{}^{\dot{A}B}\xi^{\nu}{}_{\|\mu} = 0.$$

In ξ systems the condition for ξ frame reads

$$\sigma_{\mu}{}^{\dot{A}B}{}_{,5} = 0.$$

Let us now illustrate one way to construct a ξ frame. We choose a ξ system. Since the metric is independent of x^5, it is always possible to select an orthonormal tetrad $\zeta^{(a)}{}_{\mu}$, $a = 1, 2, 3, 4$, of vector fields orthogonal to the ξ field and independent of x^5, namely

$$\left.\begin{aligned} &\zeta^{(a)}{}_{\mu}\xi^{\mu} = 0 \\ &\gamma^{\mu\nu}\zeta^{(a)}{}_{\mu}\zeta^{(b)}{}_{\nu} = \eta^{ab} \\ &\zeta^{(a)}{}_{\mu,5} = 0 \end{aligned}\right\}, \quad a, b = 1, 2, 3, 4, \tag{4.11}$$

where $\eta^{ab} = \mathrm{diag}[-1, -1, -1, 1]$.

Now we define the following five Hermitian matrices:

$$\sigma_{\mu}{}^{\dot{A}B} = \zeta^{(a)}{}_{\mu}\sigma_{(a)}{}^{\dot{A}B},$$

where $\sigma_{(a)}{}^{\dot{A}B}$ are the Pauli matrices (3.2).

It is easily verified with the aid of (4.11) that these matrices satisfy all the rules of the five-dimensional algebra for the mixed quantities, and furthermore they are independent of x^5. Hence they determine a ξ frame.

The general transformation between ξ frames is given by (4.1), where $S^{A}{}_{B}(x^{\mu})$ is a unimodular matrix satisfying

$$S^{A}{}_{B,\mu}\xi^{\mu} = 0 \tag{4.12}$$

(which, in ξ systems, states $S^{A}{}_{B,5} = 0$).

Now we are in the position to identify the four-dimensional formalism implied by the foregoing five-dimensional theory. The spinor algebra in V_5, expressed in terms of $\sigma_{\mu}{}^{\dot{A}B}$ and ξ_{μ}, reduces in a ξ system and ξ frame to the familiar four-dimensional spinor algebra for the quantities $\sigma_k{}^{\dot{A}B}$. Hence we identify $\sigma_k{}^{\dot{A}B}$ in a ξ system and ξ frame, with the mixed quantities of general relativity.

Another identity involving the mixed quantities, which will be required in the subsequent analysis, will be established now. Let $E_{\mu\nu\lambda\tau\rho}$ denote the five-dimensional alternating tensor, i.e.,

$$\left.\begin{aligned} E_{\mu\nu\lambda\tau\rho} &= \gamma^{1/2}\epsilon_{\mu\nu\lambda\tau\rho} \\ E^{\mu\nu\lambda\tau\rho} &= \gamma^{-1/2}\epsilon^{\mu\nu\lambda\tau\rho} \end{aligned}\right\}, \quad (\gamma > 0), \tag{4.13}$$

where $\epsilon_{\mu\nu\lambda\tau\rho}$ and $\epsilon^{\mu\nu\lambda\tau\rho}$ are the Levi-Civita symbols in V_5 ($\epsilon_{12345} = \epsilon^{12345} = 1$). We define

$$e_{\mu\nu\lambda\tau} = -E_{\mu\nu\lambda\tau\rho}\xi^{\rho}. \tag{4.14}$$

This tensor is completely antisymmetric in all its four indices, and satisfies the identity:

$$e_{\mu\nu\lambda\tau}\xi^{\tau} = 0. \tag{4.15}$$

Equation (4.15) shows that in ξ systems all the components of $e_{\mu\nu\lambda\tau}$ with any one of the indices having the value 5 vanish, whereas according to (4.14) and (4.13) one finds in ξ systems that

$$\begin{aligned} e_{klrs} &= -(-g)^{1/2}\epsilon_{klrs}, \\ e^{klrs} &= (-g)^{-1/2}\epsilon^{klrs} \end{aligned}$$

(ϵ_{klrs} and ϵ^{klrs} are the Levi-Civita symbols in V_4, $\epsilon_{1234} = \epsilon^{1234} = 1$). Thus e_{klrs} in ξ systems is the four-dimensional alternating tensor.

The relation (4.14) between $E_{\mu\nu\lambda\tau\rho}$ and $e_{\mu\nu\lambda\tau}$ can be inverted in the following way. The tensor $e_{[\mu\nu\lambda\tau}\xi_{\rho]}$ (square brackets denote total antisymmetrization) must be proportional to $E_{\mu\nu\lambda\tau\rho}$:

$$e_{[\mu\nu\lambda\tau}\xi_{\rho]} = \Psi E_{\mu\nu\lambda\tau\rho}$$

for some scalar Ψ. Contracting this equation with ξ^{ρ} and using (4.14) and (4.15) one finds $\Psi = 1/5$. Hence

$$E_{\mu\nu\lambda\tau\rho} = 5e_{[\mu\nu\lambda\tau}\xi_{\rho]}.$$

It easily verified (by working in a ξ systems and ξ frame) that the tensor $e_{\mu\nu\lambda\tau}$ has the following representation in terms of the mixed quantities:

$$e^{\mu\nu\lambda\tau} = i(\sigma^{\mu\dot{A}B}\sigma^{\nu\dot{C}D} - \sigma^{\nu\dot{A}B}\sigma^{\mu\dot{C}D})\sigma^{\lambda}{}_{\dot{A}D}\sigma^{\tau}_{\dot{C}B}. \tag{4.16}$$

This implies also

$$(\sigma^{\mu\dot{A}C}\sigma^{\nu}{}_{\dot{D}B} - \sigma^{\nu\dot{A}C}\sigma^{\mu}{}_{\dot{D}B})\sigma^{\lambda}{}_{\dot{D}C} = ie^{\mu\nu\lambda\tau}\sigma_{\tau}{}^{\dot{A}B}. \tag{4.17}$$

5. FIVE-DIMENSIONAL SPINOR ANALYSIS

The covariant derivative of a spinor in V_5 is defined by

$$\begin{aligned} \Psi_{A\|\mu} &\equiv \Psi_{A,\mu} - \Gamma_{A}{}^{B}{}_{\mu}\Psi_{B}, \\ \Psi^{A}{}_{\|\mu} &\equiv \Psi^{A}{}_{,\mu} + \Gamma_{B}{}^{A}{}_{\mu}\Psi^{B}, \end{aligned} \tag{5.1}$$

together with the usual properties of differentiation. The five-dimensional spin connections $\Gamma^{A}_{B\mu}$ are determined by the additional conditions imposed on the covariant differentiation. As explained in Sec. 3, a natural and convenient assumption is

$$\epsilon_{AB\|\mu} = 0, \quad \epsilon^{AB}{}_{\|\mu} = 0. \tag{5.2}$$

In contrast to the formalism in V_4, however, here it is impossible to require the vanishing of the covariant derivative of the mixed quantities, as this supposition is inconsistent with (4.3) (for this will imply another algebraic relation among the mixed quantities, the metric, and the ξ field with its derivatives). Let us, therefore, consider what conditions can be imposed on the covariant derivatives of the mixed quantities. If we restrict ourselves to equations not higher than first order in the derivatives of the metric, and not containing products of more than two derivatives or two σ's, then we are led to the following expression:

$$\sigma_\mu{}^{\dot{A}B}{}_{||\nu} = (a\xi_\mu\xi^\lambda{}_{||\nu} + b\xi_\nu\xi^\lambda{}_{||\mu})\sigma_\lambda{}^{\dot{A}B} + c\xi_\mu\sigma_\nu{}^{\dot{A}B} + d\xi_\nu\sigma_\mu{}^{\dot{A}B}, \tag{5.3}$$

with a, b, c, and d being constants, to be determined in part by the requirement that (5.3) should not constitute algebraic relations on the σ's. Contracting (5.3) with ξ^μ yields

$$[(a-1)\xi_{\mu||\nu} + c\gamma_{\mu\nu}]\sigma^{\mu\dot{A}B} = 0.$$

Hence we must take

$$a = 1, \quad c = 0.$$

Substituting this in (5.3) and contracting with $\sigma^\mu{}_{\dot{A}B}$, one finds

$$4d\xi_\nu = 0.$$

Hence

$$d = 0.$$

We are, therefore, led to the following condition:

$$\sigma_\mu{}^{\dot{A}B}{}_{||\nu} = (\xi_\mu\xi^\lambda{}_{||\nu} + b\xi_\nu\xi^\lambda{}_{||\mu})\sigma_\lambda{}^{\dot{A}B}, \tag{5.4}$$

with the only freedom being left in the choice of the arbitrary constant b. One can choose $b = 0$, in which case (5.4) takes the form

$$g_\mu{}^\lambda\sigma_\lambda{}^{\dot{A}B}{}_{||\nu} = 0.$$

Subsequent analysis shows, however, that a more coherent formalism emerges with the choice $b = 1$ [which makes (5.4) symmetrical in the two tensorial indices], and we will adhere to this choice. Thus (5.4) becomes

$$\sigma_\mu{}^{\dot{A}B}{}_{||\nu} = -(\xi_\mu\xi_\nu)_{||\lambda}\sigma^{\lambda\dot{A}B}. \tag{5.5}$$

This condition can be written in a different way, in terms of projections and the Lie derivative. Equation (5.5) is equivalent to

$$\mathcal{L}_\xi\sigma_\mu{}^{\dot{A}B} \equiv \sigma_\mu{}^{\dot{A}B}{}_{||\nu}\xi^\nu + \sigma_\nu{}^{\dot{A}B}\xi^\nu{}_{||\mu} = 0,$$
$$g_\mu{}^\lambda g_\nu{}^\tau\sigma_\lambda{}^{\dot{A}B}{}_{||\tau} = 0.$$

We have, then, identically

$$\sigma^\mu{}_{\dot{A}B||\mu} \equiv 0, \quad \sigma^\mu{}_{\dot{A}B||\nu}\sigma^{\nu\dot{A}B} \equiv 0, \quad \sigma^\mu{}_{\dot{A}B||\nu}\sigma_\mu{}^{\dot{A}B} \equiv 0.$$

We summarize this discussion by stating that the five-dimensional spinor analysis is defined by (5.1), (5.2), and (5.5). From (5.2) one finds

$$0 = \epsilon_{AB}\epsilon^{\dot{A}B}{}_{||\mu} = 2\Gamma^A_{\dot{A}\mu}. \tag{5.6}$$

In order to express the spin connections explicitly in terms of the mixed quantities, we write (5.5) in full, viz.,

$$\sigma^\mu{}_{\dot{A}B,\nu} + \Gamma^\mu_{\lambda\nu}\sigma^\lambda{}_{\dot{A}B} - \Gamma^{\dot{C}}_{\dot{A}\nu}\sigma^\mu{}_{\dot{C}B} - \Gamma^C_{B\nu}\sigma^\mu{}_{\dot{A}C} = -(\xi^\mu\xi_\nu)_{||\lambda}\sigma^\lambda{}_{\dot{A}B}.$$

Multiplying by $\sigma_\mu{}^{\dot{A}E}$ and using (4.3), (4.8), and (5.6), we find

$$\Gamma_B{}^A{}_\mu = \tfrac{1}{2}\sigma_\nu{}^{\dot{C}A}(\sigma^\nu{}_{\dot{C}B,\mu} + \Gamma^\nu_{\lambda\mu}\sigma^\lambda{}_{\dot{C}B}) + \tfrac{1}{2}\xi_\mu\xi_{\nu||\lambda}\sigma^{\nu\dot{C}A}\sigma^\lambda{}_{\dot{C}B}. \tag{5.7}$$

The last equation expressed in a ξ system and ξ frame implies

$$\Gamma^A_{B5} = 0, \quad \Gamma_B{}^A{}_k = \left\{ {A \atop B\ k} \right\}. \tag{5.8}$$

Thus we see that the five-dimensional spinor analysis, as well as the spinor algebra, reduces in ξ systems and ξ frames to the familiar four-dimensional spinor calculus.

6. FIVE-DIMENSIONAL SPINOR CURVATURE

As in four dimensions, the five-dimensional curvature spinor $\Pi^A{}_{B\mu\nu}$ is defined by the commutation rule of the covariant derivatives:

$$\Psi^A{}_{||\mu\nu} - \Psi^A{}_{||\nu\mu} = \Pi^A{}_{B\mu\nu}\Psi^B,$$

for an arbitrary five-dimensional spinor Ψ^A. It is given by

$$\Pi^A{}_{B\mu\nu} = \Gamma^A_{B\mu,\nu} - \Gamma^A_{B\nu,\mu} + \Gamma^A_{C\nu}\Gamma^C_{B\mu} - \Gamma^A_{C\mu}\Gamma^C_{B\nu}. \tag{6.1}$$

The curvature spinor possesses the following symmetries:

$$\Pi_{AB\mu\nu} = -\Pi_{AB\nu\mu} = \Pi_{BA\mu\nu},$$

where

$$\Pi_{AB\mu\nu} = \Pi^C{}_{B\mu\nu}\epsilon_{CA}$$

(the second identity follows from $\epsilon_{AB||\mu\nu} - \epsilon_{AB||\nu\mu} = 0$). This implies

$$\Pi^A{}_{A\mu\nu} \equiv 0.$$

In a ξ system and ξ frame, (6.1) reduces to

$$\Pi^A{}_{B5\mu} = 0, \quad \Pi^A{}_{Bkl} = P^A{}_{Bkl}.$$

Considering the commutator $\sigma_\lambda{}^{\dot{A}B}{}_{||\mu\nu} - \sigma_\lambda{}^{\dot{A}B}{}_{||\nu\mu}$ and using (5.5) one arrives at the expression of the curvature spinor in terms of the curvature tensor $\Theta^\lambda{}_{\tau\mu\nu}$ and the ξ field:

$$\begin{aligned}\Pi^A{}_{B\mu\nu} = \tfrac{1}{2}\sigma_\lambda{}^{\dot{C}A}\sigma^\tau{}_{\dot{C}B}[&\Theta^\lambda{}_{\tau\mu\nu} + 2\xi^\lambda{}_{||\tau}\xi_{\mu||\nu} \\ &+ (\xi^\lambda{}_{||\mu}\xi_{\tau||\nu} - \xi^\lambda{}_{||\nu}\xi_{\tau||\mu}) \\ &+ (\xi_\mu\xi^\lambda{}_{||\tau\nu} - \xi_\nu\xi^\lambda{}_{||\tau\mu})].\end{aligned} \tag{6.2}$$

The following identities, which are useful for manipulation in five-dimensional spinor curvature theory (and proved in Appendix C), are satisfied by the curvature spinor

$$\sigma^{\nu\dot{A}C}\Pi^B{}_{C\mu\nu} - \sigma^{\nu\dot{C}B}\Pi^{\dot{A}}{}_{\dot{C}\mu\nu} = 0, \tag{6.3}$$

$$\begin{aligned}\sigma^{\mu\dot{C}A}\sigma^\nu{}_{\dot{C}D}\Pi^D{}_{B\mu\nu} &= \Phi\delta^A_B, \\ \sigma^\mu{}_{\dot{C}A}\sigma^{\nu\dot{C}D}\Pi^B{}_{D\mu\nu} &= -\Phi\delta^B_A\end{aligned} \tag{6.4}$$

with

$$\Phi = \tfrac{1}{4}g^{\lambda\nu}g^{\tau\mu}(\Theta_{\lambda\tau\mu\nu} + 3\xi_{\lambda||\tau}\xi_{\mu||\nu}).$$

7. CONCLUSION

It has been shown in the present paper that the five-dimensional theory of relativity and electromagnetism

admits a consistent spinor formalism. The spinor algebra and the spinor analysis have been constructed in a covariant way, and their four-dimensional contents have been investigated. The advantage of the five-dimensional approach over the usual four-dimensional one has been discussed. In a subsequent paper[5] the spinor calculus developed here is applied to a few physical cases, in which the benefits of the five-dimensional starting point will be illustrated.

ACKNOWLEDGMENTS

Part of this work was carried out in the Department of Physics, Technion-Israel Institute of Technology, Haifa, Israel. I would like to express my gratitude to Professor Nathan Rosen for his assistance and encouragement. I also wish to thank Professor David Lovelock for helpful advice. This work was supported in part by the National Research Council of Canada under grant No. A7342.

APPENDIX A: SYMMARY OF THE FUNDAMENTAL RELATIONS OF FIVE-DIMENSIONAL SPINOR CALCULUS

I. Spinor algebra

$$\sigma^{\mu}{}_{\dot{A}B} = \sigma^{\mu}{}_{B\dot{A}},$$

$$\sigma^{\mu}{}_{\dot{A}B}\xi_{\mu} = 0,$$

$$\sigma^{\mu}{}_{\dot{A}B}\sigma^{\nu\dot{A}B} = g^{\mu\nu} \equiv \gamma^{\mu\nu} + \xi^{\mu}\xi^{\nu},$$

$$\sigma^{\mu}{}_{\dot{A}B}\sigma_{\mu}{}^{\dot{C}D} = \delta_{\dot{A}}{}^{\dot{C}}\delta_{B}^{D},$$

$$\sigma^{\mu\dot{C}A}\sigma^{\nu}{}_{\dot{C}B} + \sigma^{\nu\dot{C}A}\sigma^{\mu}{}_{\dot{C}B} = g^{\mu\nu}\delta_{B}^{A},$$

$$i(\sigma^{\mu\dot{A}B}\sigma^{\nu\dot{C}D} - \sigma^{\nu\dot{A}B}\sigma^{\mu\dot{C}D})\sigma^{\lambda}{}_{\dot{A}D}\sigma^{\tau}{}_{\dot{C}B} = e^{\mu\nu\lambda\tau} \equiv -E^{\mu\nu\lambda\tau\rho}\xi_{\rho}.$$

II. Spinor analysis

$$\epsilon_{AB||\mu} = 0, \qquad \epsilon^{AB}{}_{||\mu} = 0,$$

$$\sigma_{\mu}{}^{\dot{A}B}{}_{||\nu} = -(\xi_{\mu}\xi_{\nu})_{||\lambda}\sigma^{\lambda\dot{A}B}.$$

Spin connections:

$$\Gamma^{A}_{B\mu} = \tfrac{1}{2}\sigma_{\nu}{}^{\dot{C}A}(\sigma^{\nu}{}_{\dot{C}B,\mu} + \Gamma^{\nu}_{\lambda\mu}\sigma^{\lambda}{}_{\dot{C}B}) + \tfrac{1}{2}\xi_{\mu}\xi_{\nu||\lambda}\sigma^{\nu\dot{C}A}\sigma^{\lambda}{}_{\dot{C}B}$$

Curvature spinor:

$$\Pi^{A}{}_{B\mu\nu} = \Gamma^{A}_{B\mu,\nu} - \Gamma^{A}_{B\nu,\mu} + \Gamma^{A}_{C\nu}\Gamma^{C}_{B\mu} - \Gamma^{A}_{C\mu}\Gamma^{C}_{B\nu},$$

$$\begin{aligned}\Pi^{A}{}_{B\mu\nu} = {} & \tfrac{1}{2}\sigma_{\lambda}{}^{\dot{C}A}\sigma^{\tau}{}_{\dot{C}B}[\Theta^{\lambda}{}_{\tau\mu\nu} + 2\xi^{\lambda}{}_{||\tau}\xi_{\mu||\nu} \\ & + (\xi^{\lambda}{}_{||\mu}\xi_{\tau||\nu} - \xi^{\lambda}{}_{||\nu}\xi_{\tau||\mu}) \\ & + (\xi_{\mu}\xi^{\lambda}{}_{||\tau\nu} - \xi_{\nu}\xi^{\lambda}{}_{||\tau\mu})].\end{aligned}$$

III. ξ Frame

It is defined by

$$\sigma_{\mu}{}^{\dot{A}B}{}_{,\nu}\xi^{\nu} + \sigma_{\nu}{}^{\dot{A}B}\xi^{\nu}{}_{,\mu} = 0.$$

The general transformation between ξ frames is

$$S^{A}{}_{B}(x^{\mu}); \det(S^{A}{}_{B}) = 1, \qquad S^{A}{}_{B,\mu}\xi^{\mu} = 0.$$

The following holds in ξ systems and ξ frames:

$$\sigma_{\mu}{}^{\dot{A}B}{}_{,5} = 0,$$

$$\sigma_{5}{}^{\dot{A}B} = 0, \qquad \sigma_{k}{}^{\dot{A}B} = \text{the four-dimensional mixed quantities},$$

$$\Gamma^{A}_{B5} = 0, \qquad \Gamma_{B}{}^{A}{}_{k} = \begin{Bmatrix} A \\ B\ k \end{Bmatrix},$$

$$\Pi^{A}{}_{B5\mu} = 0, \qquad \Pi^{A}{}_{Bkl} = P^{A}{}_{Bkl}.$$

APPENDIX B: SPINORS AND BIVECTORS

In Sec. 4 a correspondence between 5-vectors orthogonal to ξ_{μ} and Hermitian spinors has been analyzed. In this appendix it will be demonstrated that as a basis for the spinor calculus of this paper we can take a correspondence between spinors and a set of tensors which are not necessarily restricted to being orthogonal to ξ_{μ}.

First we observe that with the aid of the mixed quantities defined in Sec. 4, a one-to-one mapping can be established between the 5-bivectors orthogonal to ξ_{μ} and the symmetric second rank spinors. Namely, if

$$\hat{\Phi}_{\mu\nu} = -\hat{\Phi}_{\nu\mu}$$

is a 5-bivector orthogonal to ξ_{μ}

$$\hat{\Phi}_{\mu\nu}\xi^{\nu} = 0,$$

then we define the spinor

$$\Phi^{(s)}{}_{AB} = \sigma^{\mu\dot{C}A}\sigma^{\nu}{}_{\dot{C}B}\hat{\Phi}_{\mu\nu},$$

which is symmetric

$$\Phi^{(s)}{}_{AB} = \Phi^{(s)}{}_{BA}.$$

This relation can be inverted

$$\begin{aligned}\hat{\Phi}_{\mu\nu} &= \tfrac{1}{2}(\sigma_{\mu\dot{C}A}\sigma_{\nu}{}^{\dot{C}B}\Phi^{(s)}{}_{AB} + \text{c.c.}) \\ &= \tfrac{1}{2}\sigma_{\mu\dot{C}A}(\sigma_{\nu}{}^{\dot{C}B}\Phi^{(s)}{}_{AB} + \sigma_{\nu}{}^{\dot{B}A}\Phi^{(s)}{}_{\dot{B}}{}^{\dot{C}}).\end{aligned}$$

The "scalar products" are related according to

$$\hat{\Phi}_{\mu\nu}\hat{\Phi}^{\mu\nu} = \tfrac{1}{2}(\Phi^{(s)}{}_{AB}\Phi^{(s)AB} + \text{c.c.}),$$

$$\Phi^{(s)}{}_{AB}\Phi^{(s)AB} = \hat{\Phi}_{\mu\nu}\hat{\Phi}^{\mu\nu} - \tfrac{1}{2}ie^{\mu\nu\lambda\tau}\hat{\Phi}_{\mu\nu}\hat{\Phi}_{\lambda\tau}.$$

Now let $\Phi_{\mu\nu} = -\Phi_{\nu\mu}$ be an arbitrary 5-bivector. Then[15] it defines in a unique way a pair $(\hat{\Phi}_{\mu}, \hat{\Phi}_{\mu\nu})$ consisting of a vector and a bivector orthogonal to ξ_{μ}, such that

$$\begin{aligned}&\Phi_{\mu\nu} = \hat{\Phi}_{\mu\nu} + \xi_{\mu}\hat{\Phi}_{\nu} - \xi_{\nu}\hat{\Phi}_{\mu}, \\ &\hat{\Phi}_{\nu}\xi^{\nu} = \hat{\Phi}_{\mu\nu}\xi^{\nu} = 0.\end{aligned} \tag{B1}$$

Solving for $\hat{\Phi}_{\mu}$ and $\hat{\Phi}_{\mu\nu}$, we find

$$\hat{\Phi}_{\mu} = \Phi_{\mu\nu}\xi^{\nu},$$

$$\hat{\Phi}_{\mu\nu} = g_{\mu}{}^{\lambda}g_{\nu}{}^{\tau}\Phi_{\lambda\tau}.$$

Conversely, every pair $(\hat{\Phi}_{\mu}, \hat{\Phi}_{\mu\nu})$ consisting of a vector and a bivector orthogonal to ξ_{μ} defines a bivector $\Phi_{\mu\nu}$ according to (B1).

Combining the results obtained in this appendix and in Sec. 4, we see that there is a one-to-one mapping between the set of all 5-bivectors $\Phi_{\mu\nu}$ and the set of

pairs $(\Phi^{(h)}{}_{\dot{A}B}, \Phi^{(s)}{}_{AB})$ of a Hermitian and a symmetric second rank spinors. For a given bivector $\Phi_{\mu\nu}$, the corresponding pair of spinors is defined by

$$\Phi^{(h)}{}_{\dot{A}B} = \sigma^{\mu}{}_{\dot{A}B}\Phi_{\mu\nu}\xi^{\nu},$$

$$\Phi^{(s)}{}_{AB} = \sigma^{\mu}{}_{\dot{C}A}\sigma^{\nu}{}_{\dot{C}B}g^{\lambda}_{\mu}g_{\nu}{}^{\tau}\Phi_{\lambda\tau} = \sigma^{\mu}{}_{\dot{C}A}\sigma^{\nu}{}_{\dot{C}B}\Phi_{\mu\nu}.$$

Conversely, given a pair $(\Phi^{(h)}{}_{\dot{A}B}, \Phi^{(s)}{}_{AB})$, the corresponding bivector is defined by

$$\Phi_{\mu\nu} = \tfrac{1}{2}(\sigma_{\mu\dot{C}A}\sigma_{\nu}{}^{\dot{C}B}\Phi^{(s)}{}_{AB} + \text{c.c.}) + (\xi_{\mu}\sigma_{\nu}{}^{\dot{A}B} - \xi_{\nu}\sigma_{\mu}{}^{\dot{A}B})\Phi^{(h)}{}_{\dot{A}B}.$$

The "scalar products" are related according to

$$\Phi_{\mu\nu}\Phi^{\mu\nu} = \tfrac{1}{2}(\Phi^{(s)}{}_{AB}\Phi^{(s)AB} + \text{c.c.}) - 2\Phi^{(h)}{}_{\dot{A}B}\Phi^{(h)\dot{A}B},$$

$$\Phi^{(h)}{}_{\dot{A}B}\Phi^{(h)\dot{A}B} = \Phi_{\mu\nu}\Phi^{\mu\lambda}\xi^{\nu}\xi_{\lambda},$$

$$\Phi^{(s)}{}_{AB}\Phi^{(s)AB} = (g^{\mu\lambda}g^{\nu\tau} - \tfrac{1}{2}ie^{\mu\nu\lambda\tau})\Phi_{\mu\nu}\Phi_{\lambda\tau}.$$

APPENDIX C: IDENTITIES IN FIVE-DIMENSIONAL SPINOR CURVATURE THEORY

In this appendix identities (6.3) and (6.4) are proved. By substituting the expression (6.2) and its complex conjugate, and using (4.3), one finds

$$\begin{aligned}\Psi_{\mu}^{\dot{A}B} &\equiv \sigma^{\nu\dot{A}C}\Pi^{B}{}_{C\mu\nu} - \sigma^{\nu\dot{C}B}\Pi^{\dot{A}}{}_{\dot{C}\mu\nu}\\ &= \tfrac{1}{2}(\sigma^{\nu\dot{A}C}\sigma^{\lambda}{}_{\dot{D}B} - \sigma^{\lambda\dot{A}C}\sigma^{\nu\dot{D}B})\sigma^{\tau}{}_{\dot{D}C}\\ &\quad\times[\Theta_{\lambda\tau\mu\nu} + 2\xi_{\lambda\|\tau}\xi_{\mu\|\nu}\\ &\quad + (\xi_{\lambda\|\mu}\xi_{\tau\|\nu} - \xi_{\lambda\|\nu}\xi_{\tau\|\mu}) + \xi_{\mu}\xi_{\lambda\|\tau\nu}].\end{aligned}$$

This, in view of (4.17), becomes

$$\Psi_{\mu}^{\dot{A}B} = \tfrac{1}{2}ie^{\nu\lambda\tau\rho}\sigma_{\rho}^{\dot{A}B}[\Theta_{\lambda\tau\mu\nu} + 2\xi_{\lambda\|\tau}\xi_{\mu\|\nu} + (\xi_{\lambda\|\mu}\xi_{\tau\|\nu} - \xi_{\lambda\|\nu}\xi_{\tau\|\mu}) + \xi_{\mu}\xi_{\lambda\|\tau\nu}].$$

Due to the cyclic identity of the curvature tensor, we have

$$e^{\nu\lambda\tau\rho}\Theta_{\lambda\tau\mu\nu} \equiv 0.$$

Also, evidently we have

$$e^{\nu\lambda\tau\rho}[2\xi_{\lambda\|\tau}\xi_{\mu\|\nu} + (\xi_{\lambda\|\mu}\xi_{\tau\|\nu} - \xi_{\lambda\|\nu}\xi_{\tau\|\mu})] \equiv 0$$

and

$$e^{\nu\lambda\tau\rho}\xi_{\lambda\|\tau\nu} = -e^{\nu\lambda\tau\rho}\Theta_{\lambda\tau\nu\sigma}\xi^{\sigma} \equiv 0.$$

Hence

$$\Psi_{\mu}^{\dot{A}B} = 0.$$

For the proof of (6.4) the expression (6.2) is used again:

$$\begin{aligned}\Psi^{A}{}_{B} &\equiv \sigma^{\mu\dot{C}A}\sigma^{\nu}{}_{\dot{C}D}\Pi^{D}{}_{B\mu\nu}\\ &= \tfrac{1}{2}\sigma^{\mu\dot{C}A}\sigma^{\nu}{}_{\dot{C}D}\sigma^{\lambda\dot{E}D}\sigma^{\tau}{}_{\dot{E}B}[\Theta_{\lambda\tau\mu\nu} + 2\xi_{\lambda\|\tau}\xi_{\mu\|\nu}\\ &\quad + (\xi_{\lambda\|\mu}\xi_{\tau\|\nu} - \xi_{\lambda\|\nu}\xi_{\tau\|\mu})].\end{aligned}$$

In the product of the four σ's one can commute the indices ν and λ using (4.9) to get

$$\begin{aligned}\Psi^{A}{}_{B} &= [\tfrac{1}{2}g^{\nu\lambda}\sigma^{\mu\dot{C}A}\sigma^{\tau}{}_{\dot{C}B} - \tfrac{1}{4}(\sigma^{\mu\dot{C}A}\sigma^{\nu\dot{E}D} - \sigma^{\nu\dot{C}A}\sigma^{\mu\dot{E}D})\sigma^{\lambda}{}_{\dot{C}D}\sigma^{\tau}{}_{\dot{E}B}]\\ &\quad\times[\Theta_{\lambda\tau\mu\nu} + 2\xi_{\lambda\|\tau}\xi_{\mu\|\nu} + (\xi_{\lambda\|\mu}\xi_{\tau\|\nu} - \xi_{\lambda\|\nu}\xi_{\tau\|\mu})]\\ &= \tfrac{1}{4}[g^{\nu\lambda}(\sigma^{\mu\dot{C}A}\sigma^{\tau}{}_{\dot{C}B} + \sigma^{\tau\dot{C}A}\sigma^{\mu}{}_{\dot{C}B}) + ie^{\mu\nu\lambda\rho}\sigma_{\rho}{}^{\dot{C}A}\sigma^{\tau}{}_{\dot{C}B}]\\ &\quad\times[\Theta_{\lambda\tau\mu\nu} + 2\xi_{\lambda\|\tau}\xi_{\mu\|\nu} + (\xi_{\lambda\|\mu}\xi_{\tau\|\nu} - \xi_{\lambda\|\nu}\xi_{\tau\|\mu})].\end{aligned}$$

The term containing $e^{\mu\nu\lambda\rho}$ vanishes, and for the remaining term we use again (4.9) to get finally

$$\Psi^{A}{}_{B} = \tfrac{1}{4}g^{\nu\lambda}g^{\mu\tau}(\Theta_{\lambda\tau\mu\nu} + 3\xi_{\lambda\|\tau}\xi_{\mu\|\nu})\delta^{A}_{B},$$

which completes the proof of the first equation of (6.4). The second equation follows from the first, since

$$\begin{aligned}\sigma^{\mu}{}_{\dot{C}A}\sigma^{\nu\dot{C}D}\Pi^{B}{}_{D\mu\nu} &= -\sigma^{\mu\dot{C}E}\sigma^{\nu}{}_{\dot{C}F}\Pi^{B}{}_{D\mu\nu}\epsilon_{EA}\epsilon^{DF}\\ &= \sigma^{\mu\dot{C}E}\sigma^{\nu}{}_{\dot{C}F}\Pi^{F}{}_{D\mu\nu}\epsilon_{EA}\epsilon^{BD} = \Phi\delta^{E}_{D}\epsilon_{EA}\epsilon^{BD}\\ &= -\Phi\delta^{A}_{B}.\end{aligned}$$

[1] Th. Kaluza, Sitzungsber. Preuss. Akad. Wiss. Berlin, Math.-Phys. Kl., 966 (1921).

[2] O. Klein, Z. Phys. **37**, 895 (1926); **46**, 188 (1927).

[3] E. Schmutzer, Z. Naturforschg. **15a**, 831 (1960).

[4] D. E. Littlewood, Int. J. Theor. Phys. **4**, 403 (1971).

[5] E. Leibowitz, "Five-Dimensional approach to the neutrino and electron Theories", J. Math. Phys., **15**, 313 (1974).

[6] E. Leibowitz and N. Rosen, "Five-Dimensional Relativity Theory," GRG Journal **4**, 449 (1973).

[7] Greek indices run from 1 to 5, italic indices from 1 to 4. The summations convention for repeated indices holds.

[8] W. L. Bade and H. Jehle, Rev. Mod. Phys. **25**, 714 (1953).

[9] F. A. E. Pirani in *Lectures on General Relativity*, 1964 Brandeis Summer Institute (Prentice-Hall, Engelwood Cliffs, N. J., 1965).

[10] F. Cap, W. Majerotto, W. Raab, and P. Unteregger, Fortschr. Phys. **14**, 205 (1966).

[11] L. Infeld and B. L. van der Waerden, Sitzungsber. Preuss. Akad. Wiss. Phys.-Math. Kl., 380 (1933).

[12] The convention used here is $T^{r}{}_{;kl} - T^{r}{}_{;lk} = R^{r}{}_{skl}T^{s}$, $R = R^{kl}{}_{kl}$.

[13] Strictly speaking a different notation should be used for these quantities, because $\sigma^{k}{}_{\dot{A}B}$ has already been used to denote the four-dimensional mixed quantities.

[14] For a discussion of this tensor see Ref. 6, Sec. I.

[15] E. Leibowitz and N. Rossen, *"Periodic Fields in Five-Dimensional Relativity,"* GRG Journal (to be published).

Five-dimensional approach to the neutrino and electron theories

Elhanan Leibowitz

Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada
(Received 22 June 1973)

The theories of the neutrino and the electron are studies from a five-dimensional point of view. The concept of periodic 5-spinors is examined, and the formalism of five-dimensional spinor calculus is applied to the manipulations of spinor equations. It is shown that the five-dimensional approach reproduces the results of the familiar four-dimensional approach in a unified and coherent manner. Furthermore, the present approach predicts in a natural way the correct coupling to the electromagnetic field, without ad hoc assumptions.

1. INTRODUCTION

Following the idea of taking the five-dimensional relativity theory[1] as a genuine consistent approach to the description of physical reality, it is appropriate to examine the theory of the neutrino and the electron from a five-dimensional point of view. The tools needed for this purpose are the geometry of periodic five-dimensional objects, which has already been studied[2] regarding tensorial objects, and a five-dimensional spinor calculus, which has been developed elsewhere[3] (this work will be referred to as SC).

The present work starts with an analysis of the concept of periodic 5-spinors, expressed in a 5-covariant manner, which is motivated by the usefulness of the periodicity concept in the theory of 5-tensors and by its simple geometrical characterization. Then the simplest spinor equations for periodic 5-spinors are investigated, with the application of the general five-dimensional spinor calculus. The merits of the five-dimensional approach as compared to the common four-dimensional approach are discussed.

2. THE CONCEPT OF PERIODIC SPINORS

In a ξ-system and ξ-frame[4], let us consider a 5-spinor $\Psi_A(x)$ which is periodic in x^5,

$$\Psi_A(x^k, x^5 + 2\pi/\alpha) = \Psi_A(x^k, x^5),$$

where α is a real constant. This can be expanded in a Fourier series,

$$\Psi_A(x) = \sum_{n=-\infty}^{\infty} \psi_A{}^{(n)}(x^k) e^{in\alpha x^5},$$

where $\Psi_A{}^{(n)}$ are arbitrary functions of $x^1, \ldots, x^4$. Writing

$$\Psi_A{}^{(n)} = \psi_A{}^{(n)} e^{in\alpha x^5},$$

so that

$$\Psi_A = \sum_{n=-\infty}^{\infty} \Psi_A{}^{(n)},$$

we have

$$\Psi_A{}^{(n)}{}_{,5} = in\alpha \Psi_A{}^{(n)}.$$

In view of (SC-5.8)[5] this first order differential equation can be written in a 5-covariant manner (covariant under both coordinate and spin transformations),

$$\mathcal{L}_\xi \Psi_A{}^{(n)} \equiv \Psi_A{}^{(n)}{}_{\|\mu} \xi^\mu = in\alpha \Psi_A{}^{(n)}$$

(note that the Lie-derivative operator acts on tensorial indices only, and not on spinor indices).

This form will be the basis for the definition of the term periodicity. A five-dimensional spinor $\Psi_{AB\ldots}$ will be called periodic with periodicity constant α if

$$\mathcal{L}_\xi \Psi_{AB\ldots} = i\alpha \Psi_{AB\ldots}, \tag{2.1}$$

where α is a real constant.

It follows that such a periodic 5-spinor can be written in a ξ-system and ξ-frame

$$\Psi_{AB\ldots}(x^\mu) = e^{i\alpha x^5} \psi_{AB\ldots}(x^k), \tag{2.2}$$

and the four-dimensional quantities $\psi_{AB\ldots}$ induced by the 5-spinor will be interpreted as physical spinor fields in space–time. In the case $\alpha = 0$ the spinor Ψ_{AB} is called cylindrical.[6]

Example: The fundamental antisymmetric spinor ϵ_{AB} and the five-dimensional mixed quantities $\sigma_\mu{}^{\dot{A}B}$ are cylindrical.

This is obviously true for ϵ_{AB}, and for $\sigma_\mu{}^{\dot{A}B}$ it is proved by using (SC-5.5):

$$\begin{aligned}\mathcal{L}_\xi \sigma_\mu{}^{\dot{A}B} &\equiv \sigma_\mu{}^{\dot{A}B}{}_{\|\nu} \xi^\nu + \sigma_\nu{}^{\dot{A}B} \xi^\mu{}_{\|\mu} \\ &= -(\xi_\mu \xi_\lambda)_{\|\nu} \xi^\lambda \sigma^\nu{}^{\dot{A}B} + \xi_{\nu\|\mu} \sigma^\nu{}^{\dot{A}B} \\ &= (\xi_{\mu\|\nu} + \xi_{\nu\|\mu}) \sigma^\nu{}^{\dot{A}B} = 0.\end{aligned}$$

We shall now investigate the properties of a periodic spinor Ψ_A in more detail. In a ξ system and ξ frame it has the representation

$$\Psi_A(x) = e^{i\alpha x^5} \psi_A(x^k). \tag{2.3}$$

Under the general transformation between ξ frames (SC-4.12) ψ_A transforms like a four-dimensional spinor. However, with regard to the general coordinate transformation between ξ systems (SC-2.4), we find that under (SC-2.4b) ψ_A behaves as a 4-scalar, whereas under (SC-2.4a) ψ_A transforms according to

$$\overline{\psi}_A = \psi_A e^{-i\alpha\Lambda},$$

which can be interpreted as a gauge transformation of the first kind. Thus *the periodicity condition for the 5-spinor ψ_A automatically ensures that ψ_A is endowed with all the transformation properties of a charged spinor field in space-time in the presence of an electromagnetic field.*

In a ξ system and ξ frame the following expressions are easily obtained:

$$\begin{aligned}\Psi_{A\|5} &= e^{i\alpha x^5}(i\alpha\psi_A), \\ \Psi_A{}^{\|k} &\equiv \gamma^{k\mu}\Psi_{A\|\mu} = e^{i\alpha x^5}(\partial^k \psi_A), \\ \Psi_A{}^{\|\mu}{}_{\|\mu} &\equiv \gamma^{\mu\nu}\Psi_{A\|\mu\nu} = e^{i\alpha x^5}(\partial_k \partial^k \psi_A + \alpha^2 \psi_A),\end{aligned} \tag{2.4}$$

$$\sigma^{\mu\dot{A}B}\Psi_{B\|\mu} = e^{i\alpha x^5}\sigma^{k\dot{A}B}\partial_k\psi_B, \tag{2.5}$$

where ∂_k is the four-dimensional "gauge invariant" covariant derivative:

$$\partial_k\psi_A \equiv \psi_{A;k} + i\alpha\phi_k\psi_A. \tag{2.6}$$

3. THE NEUTRINO EQUATION

In this section we shall investigate the consequences of assuming the existence of a 5-spinor, which in addition to being periodic satisfies also a simple spinor equation. Let then $\Psi_A(x^\mu)$ be a five-dimensional periodic spinor, viz.,

$$\begin{aligned}\Psi_{A\|\mu}\xi^\mu &= i\alpha\Psi_A,\\ \Psi_{\dot{A}\|\mu}\xi^\mu &= -i\alpha\Psi_{\dot{A}}.\end{aligned} \tag{3.1}$$

From now on the complex conjugate counterparts of the equations obtained will be understood and will not be given explicitly. The simplest five-dimensional spinor equations which can be set for this spinor is

$$\sigma^{\mu\dot{A}B}\Psi_{B\|\mu} = 0. \tag{3.2}$$

In a ξ frame (3.2) reduces to (according to the results of Sec. 2)

$$\Psi_A(x^\mu) = e^{i\alpha x^5}\psi_A(x^k), \tag{3.3}$$

$$\sigma^{k\dot{A}B}\partial_k\psi_B = 0. \tag{3.4}$$

Equation (3.4) is precisely the Weyl equation[7] for a charged "neutrino" with electric charge $e = -\hbar\alpha$. The case of physical electrically neutral neutrino is obtained by setting $\alpha = 0$ [i.e., $\Psi_A(x^\mu)$ is a cylindrical spinor].

In order to incorporate the neutrino field into a gravitational–electromagnetic theory, a real, symmetric, cylindrical, divergence-free tensor $\Upsilon_{\mu\nu}$ is required[8], to construct the energy–momentum 5-tensor. It is easily proved that the 5-tensor

$$\begin{aligned}\Upsilon_{\mu\nu} = i[&\sigma_\mu{}^{\dot{A}B}(\delta_\nu{}^\lambda - \xi_\nu\xi^\lambda) + \sigma_\nu{}^{\dot{A}B}(\delta_\mu{}^\lambda - \xi_\mu\xi^\lambda)]\\ &\times(\Psi_{\dot{A}\|\lambda}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\lambda})\end{aligned} \tag{3.5}$$

is real, symmetric, and cylindrical, and in Appendix A we show that

$$\Upsilon_\mu{}^\nu{}_{\|\nu} = 0. \tag{3.6}$$

This tensor can also be written in the equivalent forms

$$\begin{aligned}\Upsilon_{\mu\nu} = i(&\sigma_\mu{}^{\dot{A}B}\delta_\nu{}^\lambda + \sigma_\nu{}^{\dot{A}B}\delta_\mu{}^\lambda)(\Psi_{\dot{A}\|\lambda}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\lambda})\\ &+ \tfrac{1}{2}(J_\mu\xi_\nu + J_\nu\xi_\mu),\end{aligned} \tag{3.7}$$

$$\begin{aligned}\Upsilon_{\mu\nu} = i(&\sigma_\mu{}^{\dot{A}B}g_\nu{}^\lambda + \sigma_\nu{}^{\dot{A}B}g_\mu{}^\lambda)(\Psi_{\dot{A}\|\lambda}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\lambda})\\ &+ (J_\mu\xi_\nu + J_\nu\xi_\mu),\end{aligned} \tag{3.8}$$

where

$$J_\mu = -4\alpha\sigma_\mu{}^{\dot{A}B}\Psi_{\dot{A}}\Psi_B.$$

The four-dimensional quantities induced by this tensor in a ξ system[8] are

$$-\Upsilon_5{}^k = -4\alpha\sigma^{k\dot{A}B}\psi_{\dot{A}}\psi_B = J^k,$$

$$\Upsilon^{kl} = i(\sigma^{k\dot{A}B}g^{lr} - \sigma^{l\dot{A}B}g^{kr})[(\partial^*{}_r\psi_{\dot{A}})\psi_B - \psi_{\dot{A}}(\partial_r\psi_B)] \equiv T^{kl}.$$

[In the second equation the expressions (2.4) have been used.] The four-dimensional vector J^k is the "charged neutrino" charge–current density, while T^{kl} is the four-dimensional neutrino energy-momentum density. Following the general procedure outlined elsewhere[8], the five-dimensional gravitational–electromagnetic field equations become

$$E_{\mu\nu} \equiv K_{\mu\nu} + \tfrac{1}{2}\Upsilon_{\mu\nu} = 0,$$

since

$$\Upsilon \equiv \gamma^{\mu\nu}\Upsilon_{\mu\nu} = 0.$$

These equations are precisely the Einstein–Maxwell equations for curved space–time containing a "charged neutrino" field. The 5-conservation law

$$E_\mu{}^\nu{}_{\|\nu} = 0$$

is equivalent to charge conservation together with the conservation of four-dimensional energy–momentum.

4. THE DIRAC EQUATION

Motivated by the four-dimensional spinor theory, we shall now proceed to consider periodic 5-spinors subject to the five-dimensional Dirac equation. Let then $(\Psi_A, X_{\dot{A}})$ be a pair of five-dimensional periodic spinors with the same periodicity constant α:

$$\begin{aligned}\Psi_{A\|\mu}\xi^\mu &= i\alpha\Psi_A,\\ X_{\dot{A}\|\mu}\xi^\mu &= i\alpha X_{\dot{A}}.\end{aligned} \tag{4.1}$$

As the field equations to be satisfied by these spinors we will take the five-dimensional Dirac equations, viz.,

$$\begin{aligned}\sigma^{\mu\dot{A}B}\Psi_{B\|\mu} + MX^{\dot{A}} &= 0,\\ \sigma^{\mu\dot{B}A}X_{\dot{B}\|\mu} + M\Psi^A &= 0,\end{aligned} \tag{4.2}$$

With the same real constant M. These equations are consistent with the periodicity conditions (4.1). In a ξ system and ξ frame Eqs. (4.1) and (4.2) reduce to

$$\Psi_A(x^\mu) = e^{i\alpha x^5}\psi_A(x^k), \qquad X_{\dot{A}}(x^\mu) = e^{i\alpha x^5}\chi_{\dot{A}}(x^k), \tag{4.3}$$

$$\begin{aligned}\sigma^{k\dot{A}B}\partial_k\psi_B + M\chi^{\dot{A}} &= 0,\\ \sigma^{k\dot{B}A}\partial_k\chi_{\dot{B}} + M\psi^A &= 0\end{aligned} \tag{4.4}$$

[∂_k is defined in (2.6)]. $(\psi_A, \chi_{\dot{A}})$ is a pair of four-dimensional two-component spinors, endowed with a gauge transformation of the first kind. Equation (4.4) is precisely the Dirac equation[9] in curved space–time in the presence of an electromagnetic field for a particle with an electric charge[10]

$$e = -\hbar\alpha$$

and mass

$$m = \sqrt{2}\hbar M.$$

Note that the coupling to the electromagnetic field through the "gauge invariant" covariant derivative is a result of the 5-periodicity of the spinor field, and not an additional assumption.

In Appendix B it is proved that the real, symmetric, and cylindrical 5-tensor

$$\begin{aligned}\Upsilon_{\mu\nu} = i[&\sigma_\mu{}^{\dot{A}B}(\delta_\nu{}^\lambda - \xi_\nu\xi^\lambda) + \sigma_\nu{}^{\dot{A}B}(\delta_\mu{}^\lambda - \xi_\mu\xi^\lambda)]\\ &\times(\Psi_{\dot{A}\|\lambda}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\lambda} - X_{\dot{A}\|\lambda}X_B + X_{\dot{A}}X_{B\|\lambda})\end{aligned} \tag{4.5}$$

is divergence-free:

$$\Upsilon_{\mu}{}^{\nu}{}_{\|\nu} = 0, \tag{4.6}$$

and hence can be used for the construction of the five-dimensional energy–momentum tensor. Equivalent forms for the $\Upsilon_{\mu\nu}$ tensor are

$$\Upsilon_{\mu\nu} = i(\sigma_{\mu}{}^{\dot{A}B}\delta_{\nu}{}^{\lambda} + \sigma_{\nu}{}^{\dot{A}B}\delta_{\mu}{}^{\lambda})(\Psi_{\dot{A}\|\lambda}\Psi_{B} - \Psi_{\dot{A}}\Psi_{B\|\lambda} - X_{\dot{A}\|\lambda}X_{B} + X_{\dot{A}}X_{B\|\lambda}) + \tfrac{1}{2}(I_{\mu}\xi_{\nu} + I_{\nu}\xi_{\mu}) \tag{4.7}$$

and

$$\Upsilon_{\mu\nu} = i(\sigma_{\mu}{}^{\dot{A}B}g_{\nu}{}^{\lambda} + \sigma_{\nu}{}^{\dot{A}B}g_{\mu}{}^{\lambda})(\Psi_{\dot{A}\|\lambda}\Psi_{B} - \Psi_{\dot{A}}\Psi_{B\|\lambda} - X_{\dot{A}\|\lambda}X_{B} + X_{\dot{A}}X_{B\|\lambda}) + (I_{\mu}\xi_{\nu} + I_{\nu}\xi_{\mu}), \tag{4.8}$$

where

$$I_{\mu} = -4\alpha\sigma_{\mu}{}^{\dot{A}B}(\Psi_{\dot{A}}\Psi_{B} + X_{\dot{A}}X_{B}). \tag{4.9}$$

The four-dimensional quantities induced by this tensor in a ξ system are

$$\begin{aligned} -\Upsilon_{5}{}^{k} &= -4\alpha\sigma^{k\dot{A}B}(\psi_{\dot{A}}\psi_{B} + \chi_{\dot{A}}\chi_{B}) = J^{k}, \\ \Upsilon^{kl} &= i(\sigma^{k\dot{A}B}g^{lr} + \sigma^{l\dot{A}B}g^{kr}) \\ &\quad \times [(\partial_{r}^{*}\psi_{\dot{A}})\psi_{B} - \psi_{\dot{A}}(\partial_{r}\psi_{B}) - (\partial_{r}\chi_{\dot{A}})\chi_{B} + \chi_{\dot{A}}(\partial_{r}^{*}\chi_{B})] \\ &\equiv T^{kl}. \end{aligned} \tag{4.10}$$

The four-dimensional tensors J^{k} and T^{kl} are the electron charge-current and energy–momentum densities. The five-dimensional gravitational–electromagentic field equations are

$$E_{\mu\nu} \equiv K_{\mu\nu} + \tfrac{1}{2}(\Upsilon_{\mu\nu} + \xi_{\mu}\xi_{\nu}\Upsilon),$$

with

$$\Upsilon \equiv \gamma^{\mu\nu}\Upsilon_{\mu\nu} = 2M(\Psi_{\dot{A}}X^{\dot{A}} - \Psi_{A}X^{A}),$$

and they are equivalent to the combined set of the four-dimensional Einstein–Maxwell equations with the Dirac field as a source. The 5-conservation law is equivalent to charge conservation together with the conservation of four-dimensional energy–momentum.

The five dimensional field equations of this section are derived from a 5-variational principle, where the Lagrangian is integrated over a domain consisting of entire periods, and the variations of the metric and the electron fields are restricted to those preserving cylindricity and periodicity conditions, respectively (in the case of the neutrino theory, of course, the variations of both fields must preserve the cylindricity conditions).

It is interesting to note that in contrast to the case of periodic tensor fields satisfying second order field equations (which have been discussed in Ref. 2), no special role is played here by a spinor field with a charge-to-mass ratio $|e/m| = 1$.

5. CONCLUSION

In the present paper the theories of neutrino and electron have been worked out in the framework of the five-dimensional unified theory of relativity and electromagnetism. The spinor calculus developed in a previous paper has been fully employed. It has been shown that the results of the familiar four-dimensional approach to the theory of neutrino and electron are reproduced by the five-dimensional approach in a coherent way and on the basis of simple assumptions. It is striking that the proper coupling to the electromagnetic field turns out to be a consequence of the periodicity of the 5-spinor fields, with the implication that no special provisions in the spinor formalism are required for that purpose (as is the case in the usual four-dimensional approach).

This paper demonstrates also that the concept of periodic 5-fields, which has already been utilized for the analysis of various physical tensor fields, can without difficulty be extended to include spinor fields. There are reasons to believe that the benefits of working with such fields have not yet been fully unearthed, and the concept of periodicity may appear to be fruitful in other contexts also, and therefore deserves further study.

ACKNOWLEDGMENTS

Part of this work was carried out in the Department of Physics, Technion-Israel Institute of Technology, Haifa, Israel. I am indebted to Professor Nathan Rosen for his help and suggestions. I am grateful to Professor David Lovelock for interesting discussions. This work was supported in part by the National Research Council of Canada under grant No. A7342.

APPENDIX A: NEUTRINO ENERGY-MOMENTUM TENSOR

In this appendix the vanishing of the divergence of the neutrino energy–momentum 5-tensor (3.5) will be proved. In order to do so, we shall first establish a few identities.

Throughout, let Ψ_{A} represent a periodic 5-spinor satisfying the neutrino equation (3.2). Then we prove the following five identities:

(i) $$g^{\mu\nu}\Psi_{A\|\mu\nu} = \Phi\Psi_{A} + 2i\alpha\xi_{\mu\|\nu}\sigma^{\mu}{}_{\dot{C}A}\sigma^{\nu\dot{C}B}\Psi_{B},$$

with

$$\Phi = \tfrac{1}{4}g^{\mu\tau}g^{\nu\lambda}(\Theta_{\mu\nu\lambda\tau} + 3\xi_{\mu\|\nu}\xi_{\lambda\|\tau}). \tag{A.1}$$

Proof: We differentiate the neutrino equation

$$\sigma^{\mu\,\dot{C}B}\Psi_{B\|\mu} = 0$$

covariantly with respect to x^{ν} and multiply by $\sigma^{\nu}{}_{\dot{C}A}$, using (SC-5.5), (SC-4.3), and (3.1), to get

$$-i\alpha\xi_{\mu\|\nu}\sigma^{\mu}{}_{\dot{C}A}\sigma^{\nu\,\dot{C}B}\Psi_{B} + \sigma^{\nu}{}_{\dot{C}A}\sigma^{\mu\,\dot{C}B}\Psi_{B\|\mu\nu} = 0. \tag{A.2}$$

Writing

$$\Psi_{B\|\mu\nu} = \tfrac{1}{2}(\Psi_{B\|\mu\nu} + \Psi_{B\|\nu\mu}) - \tfrac{1}{2}\Pi^{C}{}_{B\mu\nu}\Psi_{C},$$

and using (SC-4.9) and (SC-6.4), we have

$$\begin{aligned} \sigma^{\nu}{}_{\dot{C}A}\sigma^{\mu\,\dot{C}B}\Psi_{B\|\mu\nu} &= \tfrac{1}{2}(\sigma^{\nu}{}_{\dot{C}A}\sigma^{\mu\,\dot{C}B} + \sigma^{\mu}{}_{\dot{C}A}\sigma^{\nu\,\dot{C}B})\Psi_{B\|\mu\nu} \\ &\quad + \tfrac{1}{2}\sigma^{\mu}{}_{\dot{C}A}\sigma^{\nu\dot{C}B}\Pi^{D}{}_{B\mu\nu}\Psi_{D} \\ &= \tfrac{1}{2}g^{\mu\nu}\Psi_{A\|\mu\nu} - \tfrac{1}{2}\Phi\Psi_{A}. \end{aligned}$$

This is substituted in (A.2) to give (i).

(ii) $$\Psi_{A}{}^{\|\mu}{}_{\|\mu} = (\Phi + \alpha^{2})\Psi_{A} + 2i\alpha\xi_{\mu\|\nu}\sigma^{\mu}{}_{\dot{C}A}\sigma^{\nu\,\dot{C}B}\Psi_{B}.$$

Proof: Using the periodicity condition we get

$$\Psi_{A}{}^{\|\mu}{}_{\|\mu} = (g^{\mu\nu} - \xi^{\mu}\xi^{\nu})\Psi_{A\|\mu\nu} = g^{\mu\nu}\Psi_{A\|\mu\nu} + \alpha^{2}\Psi_{A}.$$

From this and (i) follows (ii).

(iii) $$i\sigma_{\mu}{}^{\dot{A}B}(\Psi_{\dot{A}}{}^{\|\nu}{}_{\|\nu}\Psi_{B} - \Psi_{\dot{A}}\Psi_{B}{}^{\|\nu}{}_{\|\nu}) = -\xi_{\mu\|\nu}J^{\nu}.$$

Proof: Substituting (ii) we find

$$\begin{aligned} &i\sigma_{\mu}{}^{\dot{A}B}(\Psi_{\dot{A}}{}^{\|\nu}{}_{\|\nu}\Psi_{B} - \Psi_{\dot{A}}\Psi_{B}{}^{\|\nu}{}_{\|\nu}) \\ &\quad = -2\alpha(\sigma_{\mu}{}^{\dot{A}C}\sigma^{\nu}{}_{\dot{D}B} + \sigma^{\nu}{}_{\dot{A}C}\sigma_{\mu}{}^{\dot{D}B})\sigma^{\lambda}{}_{\dot{D}C}\xi_{\nu\|\lambda}\Psi_{\dot{A}}\Psi_{B}. \end{aligned}$$

In the second term the formula (SC-4.9) is applied twice to commute the σ's, and we get (iii).

$$\text{(iv)}\quad i\sigma^{\nu\dot{A}B}(\Psi_{\dot{A}\|\mu\nu}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\mu\nu}) = -i\sigma^{\nu\dot{A}B}{}_{\|\mu}(\Psi_{\dot{A}\|\nu}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\nu}).$$

Proof: The left-hand side is equal to

$$i\sigma^{\nu\dot{A}B}(\Psi_{\dot{A}\|\nu\mu}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\nu\mu}) + i(\sigma^{\nu\dot{A}C}\Pi^B{}_{C\mu\nu} - \sigma^{\nu\dot{C}B}\Pi^{\dot{A}}{}_{\dot{C}\mu\nu})\Psi_{\dot{A}}\Psi_B.$$

The second term vanishes due to the identity (SC-6.3), while the first term, with the aid of the neutrino equation, becomes the right-hand side of (iv).

$$\text{(v)}\quad J^{\mu}{}_{\|\mu} = 0, \qquad J_{\mu\|\nu}\xi^{\nu} = \xi_{\mu\|\nu}J^{\nu}.$$

The proof is straightforward.

Now we are in the position to prove (3.6). We calculate the divergence of $\Upsilon_{\mu\nu}$ in its form (3.7), taking into account the neutrino equation and (v):

$$\Upsilon_{\mu}{}^{\nu}{}_{\|\nu} = i\sigma_{\mu}{}^{\dot{A}B\|\nu}(\Psi_{\dot{A}\|\nu}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\nu}) + i\sigma_{\mu}{}^{\dot{A}B}(\Psi_{\dot{A}}{}^{\|\nu}{}_{\|\nu}\Psi_B - \Psi_{\dot{A}}\Psi_B{}^{\|\nu}{}_{\|\nu}) + i\sigma^{\nu\dot{A}B}(\Psi_{\dot{A}\|\mu\nu}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\mu\nu}) + \xi_{\mu\|\nu}J^{\nu}.$$

For the second term (iii) is used, and for the third term (iv) is invoked:

$$\Upsilon_{\mu}{}^{\nu}{}_{\|\nu} = i(\sigma_{\mu}{}^{\dot{A}B\|\nu} - \sigma^{\nu\dot{A}B}{}_{\|\mu})(\Psi_{\dot{A}\|\nu}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\nu}).$$

This expression vanishes, since according to (SC-5.5) the first bracket is identically zero.

APPENDIX B: ELECTRON ENERGY-MOMENTUM TENSOR

In this appendix the vanishing of the divergence of (4.5), which is used for the construction of the electron energy–momentum 5-tensor, will be proved. The steps of the proof run parallel to the steps in Appendix A, and the details of the calculations here can be omitted whenever they are very similar to those of Appendix A. Again a few lemmas are needed.

Throughout, let (Ψ_A, X_A) represent a pair of periodic 5-spinors [i.e., subject to conditions (4.1)], satisfying Dirac equations (4.2). Then we have the following five identities:

$$\text{(i)}\quad g^{\mu\nu}\Psi_{A\|\mu\nu} = \Omega\Psi_A + 2i\alpha\xi_{\mu\|\nu}\sigma^{\mu}{}_{\dot{C}A}\sigma^{\nu\dot{C}B}\Psi_B,$$

with

$$\Omega = \Phi - 2M^2$$

and Φ being given by (A.1).

Likewise,

$$g^{\mu\nu}X_{\dot{A}\|\mu\nu} = \Omega X_{\dot{A}} + 2i\alpha\xi_{\mu\|\nu}\sigma^{\mu}{}_{\dot{A}C}\sigma^{\nu\dot{B}C}X_{\dot{B}}.$$

[There is a complete symmetry between the roles of ψ_A and $X_{\dot{A}}$ in the theory. Hence all the statements hereafter, which will be expressed in terms of $\psi_A(X_{\dot{A}})$, can also be expressed in terms of $X_{\dot{A}}(\psi_A)$, and we will not bother to write this down explicitly.]

Proof: We differentiate the Dirac equation

$$\sigma^{\mu\dot{C}B}\Psi_{B\|\mu} = -MX^{\dot{C}}$$

covariantly with respect to x^{ν} and multiply by $\sigma^{\nu}{}_{\dot{C}A}$:

$$-i\alpha\xi_{\mu\|\nu}\sigma^{\mu}{}_{\dot{C}A}\sigma^{\nu\dot{C}B}\Psi_B + \sigma^{\nu}{}_{\dot{C}A}\sigma^{\mu\dot{C}B}\Psi_{B\|\mu\nu} = -M\sigma^{\nu}{}_{\dot{C}A}X^{\dot{C}}{}_{\|\nu} = -M^2\Psi_A.$$

The second term on the left-hand side, as has been shown in (i) of Appendix A, is equal to $\frac{1}{2}g^{\mu\nu}\Psi_{A\|\mu\nu} - \frac{1}{2}\Phi\Psi_A$, and hence (i) here follows.

$$\text{(ii)}\quad \Psi_A{}^{\|\mu}{}_{\|\mu} = (\Omega + \alpha^2)\Psi_A + 2i\alpha\xi_{\mu\|\nu}\sigma^{\mu}{}_{\dot{C}A}\sigma^{\nu\dot{C}B}\Psi_B.$$

$$\text{(iii)}\quad i\sigma_{\mu}{}^{\dot{A}B}(\Psi_{\dot{A}}{}^{\|\nu}{}_{\|\nu}\Psi_B - \Psi_{\dot{A}}\Psi_B{}^{\|\nu}{}_{\|\nu} - X_{\dot{A}}{}^{\|\nu}{}_{\|\nu}X_B + X_{\dot{A}}X_B{}^{\|\nu}{}_{\|\nu}) = -\xi_{\mu\|\nu}I^{\nu}.$$

$$\text{(iv)}\quad i\sigma^{\nu\dot{A}B}(\Psi_{\dot{A}\|\mu\nu}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\mu\nu} - X_{\dot{A}\|\mu\nu}X_B + X_{\dot{A}}X_{B\|\mu\nu}) = -i\sigma^{\nu\dot{A}B}{}_{\|\mu}(\Psi_{\dot{A}\|\nu}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\nu} - X_{\dot{A}\|\nu}X_B + X_{\dot{A}}X_{B\|\nu}) + iM(\Psi_{\dot{A}}X^{\dot{A}} - \Psi_A X^A)_{\|\mu}.$$

Proof: The left-hand side is equal to

$$i\sigma^{\nu\dot{A}B}(\Psi_{\dot{A}\|\nu\mu}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\nu\mu} - X_{\dot{A}\|\nu\mu}X_B + X_{\dot{A}}X_{B\|\nu\mu}) + i(\sigma^{\nu\dot{A}C}\Pi^B{}_{C\mu\nu} - \sigma^{\nu\dot{C}B}\Pi^{\dot{A}}{}_{\dot{C}\mu\nu})(\Psi_{\dot{A}}\Psi_B - X_{\dot{A}}X_B).$$

The second term vanishes, while the first term, with the aid of Dirac equation, becomes the right-hand side of (iv).

$$\text{(v)}\quad I^{\mu}{}_{\|\mu} = 0, \qquad I_{\mu\|\nu}\xi^{\nu} = \xi_{\mu\|\nu}I^{\nu}.$$

Now we are in the position to prove (4.6). We calculate the divergence of $\Upsilon_{\mu\nu}$ in the form (4.7), taking into account Diract equation and (v):

$$\Upsilon_{\mu}{}^{\nu}{}_{\|\nu} = i\sigma_{\mu}{}^{\dot{A}B\|\nu}(\Psi_{\dot{A}\|\nu}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\nu} - X_{\dot{A}\|\nu}X_B + X_{\dot{A}}X_{B\|\nu}) - iM(\Psi_{\dot{A}}X^{\dot{A}} - \Psi_A X^A)_{\|\mu} + i\sigma_{\mu}{}^{\dot{A}B}(\Psi_A{}^{\|\nu}{}_{\|\nu}\Psi_B - \Psi_{\dot{A}}\Psi_B{}^{\|\nu}{}_{\|\nu} - X_{\dot{A}}{}^{\|\nu}{}_{\|\nu}X_B + X_{\dot{A}}X_B{}^{\|\nu}{}_{\|\nu}) + i\sigma^{\nu\dot{A}B}(\Psi_{\dot{A}\|\mu\nu}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\mu\nu} - X_{\dot{A}\|\mu\nu}X_B + X_{\dot{A}}X_{B\|\mu\nu}) + \xi_{\mu\|\nu}I^{\nu}.$$

Substituting (iii) for the third term and (iv) for the fourth term, we find

$$\Upsilon_{\mu}{}^{\nu}{}_{\|\nu} = i(\sigma_{\mu}{}^{\dot{A}B\|\nu} - \sigma^{\nu\dot{A}B}{}_{\|\mu})(\Psi_{\dot{A}\|\nu}\Psi_B - \Psi_{\dot{A}}\Psi_{B\|\nu} - X_{\dot{A}\|\nu}X_B + X_{\dot{A}}X_{B\|\nu}) = 0.$$

[1]E. Leibowitz and N. Rosen, "Five-Dimensional Relativity Theory," GRG 4, 449 (1973).
[2]E. Leibowitz and N. Rosen, "Periodic Fields in Five-Dimensional Relativity", GRG (to be published)
[3]E. Leibowitz, "Spinor Calculus in Five-Dimensional Relativity" J. Math. Phys. 15, 306 (1974).
[4]We follow throughout this paper the notation of Ref. 3.
[5]This notation means "Eq. (5.8) in Ref. 3."
[6]See Ref. 1, Sec. IV.
[7]O. Bergmann, J. Math. Phys. 1, 172 (1960).
[8]Ref. 2, Sec. 2.
[9]W. L. Bade and H. Jehle, Rev. Mod. Phys. 25, 714 (1953).
[10]For a discussion of the system of units employed here, see Appendix A of Ref. 1.

Smoothed boundary conditions for randomly rough surfaces

Alan R. Wenzel*

Rosenstiel School of Marine and Atmospheric Science, University of Miami, Coral Gables, Florida 33124
(Received 8 June 1973)

The problem of scalar wave propagation in the half-space bounded by a rigid, randomly rough surface which is a small perturbation of an infinite plane is considered. It is shown that, if the boundary roughness is statistically homogeneous, the coherent (or average) wave satisfies a generalized impedance boundary condition on the average boundary. This is referred to as a "smoothed" boundary condition. Applying it to plane waves yields an expression for the effective plane-wave reflection coefficient C_e of the boundary. For the case of isotropic roughness, approximate expressions for C_e are obtained for both long (relative to the correlation length of the boundary roughness) and short waves. These expressions show that generally $|C_e| < 1$, and therefore that the amplitude of the coherent wave is diminished upon reflection from the boundary. This is the result of scattering of energy out of the coherent wave by the boundary roughness. It is also shown that this type of boundary can support a surface wave. This wave propagates at near-grazing incidence with a speed slightly less than the free-space propagation speed. Its amplitude decreases with propagation distance, also as a result of scattering by the boundary roughness.

INTRODUCTION

We consider here the problem of scalar wave propagation in the half-space bounded by a rigid, randomly rough surface which is a small perturbation of an infinite plane. The procedure is to express the solution for any realization of the boundary in terms of a perturbation expansion in a small parameter measuring the deviation of the boundary from the plane, which is also the average boundary. This expansion is carried out to terms of second order, which is equivalent to a double scattering approximation. By taking ensemble averages, we obtain an expression for the coherent (or average) wave. We use this expression to show that, if the boundary roughness is statistically homogeneous, the coherent wave satisfies a generalized impedance boundary condition on the average boundary. We call this a "smoothed" boundary condition. Together with the original equation of motion, and a radiation condition, it constitutes a boundary value problem for the coherent wave.

By considering plane-wave solutions of this boundary value problem we get an expression for the effective plane-wave reflection coefficient C_e of the boundary. This expression shows that generally C_e depends on the direction of propagation as well as on the frequency and angle of incidence. However, if the boundary roughness is statistically isotropic, C_e is independent of the propagation direction.

For the case of isotropic roughness we obtain approximate expressions for C_e for both long (relative to the correlation length of the boundary roughness) and short waves. We show also, by a modification of this analysis, that a rough boundary can support a surface wave. These results are described in more detail in the abstract and in Secs. IIB and IIC.

Equivalent impedance boundary conditions for rough surfaces have been obtained previously by Twersky[1,2] and Biot.[3] Twersky's model of rough surface reflection and scattering consists generally of a plane wave incident on a random distribution of arbitrary, but identical, bosses on a plane surface. His results for arbitrary bosses have been specialized to hemispheres and semicylinders,[1,2] and to semielliptic cylinders.[4] They have also been used to plot graphs of both magnitude and phase of reflection amplitudes for the case of hemispherical bosses.[5] Biot used a uniform distribution of sources and dipoles on a plane as a model of a rough surface. His boundary condition is general in the sense that it is not limited to plane waves; however, it is valid only for the case in which the wavelength is much greater than the scale of the surface roughness. In contrast, the present results are not restricted as to wavelength.

Derivations of impedance boundary conditions for rough surfaces have also been given by Senior,[6] who treated the electromagnetic case, and by Lysanov,[7] who considered the acoustic problem with a pressure-release surface. Both of these investigators used perturbation methods. More recently, Zipfel and De Santo[8] have studied the scattering of scalar waves by a rigid, randomly rough surface with the aid of diagram methods. The problem of differential equations with random boundary conditions has also been examined from a general viewpoint by Keller,[9] who adapted a method which has previously been applied to the study of differential equations with random coefficients.[10]

More complete reviews of the extensive literature on scattering and reflection of waves by randomly rough surfaces have been given by Beckmann and Spizzichino,[11] Fortuin,[12] and Horton.[13]

I. DERIVATION OF THE SMOOTHED BOUNDARY CONDITION

The type of problem we wish to consider involves, generally, finding certain statistical properties of the function $\phi(x, y, z)$ which is a solution of the reduced wave equation

$$(\nabla^2 + k_0^2)\phi = f \tag{1}$$

in the region $z > \epsilon\mu(x, y)$, subject to the boundary condition

$$\frac{\partial\phi}{\partial\nu} = 0 \tag{2}$$

on $z = \epsilon\mu(x, y)$, and a radiation condition. The (nonrandom) source function $f(x, y, z)$ is assumed to be given, and the random function μ, which defines the boundary, is assumed to have zero mean. Also, ν is a unit vector normal to the boundary, ϵ is a small parameter measuring the deviation of the boundary from the x–y plane, k_0 is a positive constant, and a common time factor $\exp\{-i\beta t\}$ $(\beta > 0)$ has been dropped. Here we

shall be concerned with the ensemble average of ϕ, which we call the coherent wave. It should be noted, however, that the method we employ can also be used to obtain other statistical properties of ϕ, such as the mean square, covariance, etc. It can also be used to treat cases in which the average boundary is other than a flat plane.

We begin by assuming a solution of Eqs. (1) and (2) of the form

$$\phi(x, y, z; \epsilon) = \phi^{(0)}(x, y, z) + \epsilon\phi^{(1)}(x, y, z) + \epsilon^2\phi^{(2)}(x, y, z) + \cdots. \quad (3)$$

By substituting Eq. (3) into Eqs. (1) and (2), expanding in powers of ϵ, and collecting terms in like powers of ϵ we find that the functions $\phi^{(n)}$ must satisfy the equation

$$(\nabla^2 + k_0^2)\phi^{(n)} = \begin{cases} f, & n = 0, \\ 0, & n > 0, \end{cases} \quad (4)$$

in the region $z > 0$, subject to the boundary condition

$$\phi_z^{(n)} = \begin{cases} 0, & n = 0, \\ \mu_x\phi_x^{(n-1)} + \mu_y\phi_y^{(n-1)} - \mu\phi_{zz}^{(n-1)}, & n > 0, \end{cases} \quad (5)$$

on the plane $z = 0$. In addition, we require that each of the functions $\phi^{(n)}$ satisfy a radiation condition at ∞.

Equations (4) and (5) can be solved with the aid of the appropriate Green's function. This yields the following recursion formulas for the functions $\phi^{(n)}$:

$$\phi^{(0)}(x, y, z) = \iiint_{\zeta>0} G(x, y, z; \xi, \eta, \zeta) f(\xi, \eta, \zeta) d\xi d\eta d\zeta, \quad (6)$$

$$\phi^{(n)}(x, y, z) = (2\pi)^{-1} \int_{-\infty}^{\infty} \int F(x - \xi, y - \eta, z)[\mu_x(\xi, \eta) \times \phi_x^{(n-1)}(\xi, \eta, 0) + \mu_y(\xi, \eta)\phi_y^{(n-1)}(\xi, \eta, 0) - \mu(\xi, \eta) \times \phi_{zz}^{(n-1)}(\xi, \eta, 0)] d\xi d\eta, \quad n > 0. \quad (7)$$

Here the Green's function G is given by

$$G(x, y, z; \xi, \eta, \zeta) = (4\pi)^{-1}[F(x - \xi, y - \eta, z - \zeta) + F(x - \xi, y - \eta, z + \zeta)], \quad (8)$$

where

$$F(\alpha_1, \alpha_2, \alpha_3) = -s^{-1}e^{ik_0 s} \quad (9)$$

and

$$s = (\alpha_1^2 + \alpha_2^2 + \alpha_3^2)^{1/2}. \quad (10)$$

We now wish to use Eqs. (6) and (7) to write the solution for ϕ, correct through terms of order ϵ^2, in terms of $\phi^{(0)}$, G, and μ. To do this, we must first substitute into Eq. (7) (with $n = 2$) the formula for $\phi^{(1)}$ obtained from Eq. (7) with $n = 1$. However, upon differentiating Eq. (7) (with $n = 1$) we find that we cannot set $z = 0$ in the resulting expressions because the integrals appearing in them become divergent. This difficulty can be overcome if we first integrate by parts to eliminate derivatives of F before setting $z = 0$. In doing this we use the fact that the function F satisfies the equation

$$F_{\alpha_1\alpha_1} + F_{\alpha_2\alpha_2} + F_{\alpha_3\alpha_3} + k_0^2 F = 0$$

when $s \neq 0$. Then by substituting the resulting expressions for $\phi_x^{(1)}$, $\phi_y^{(1)}$, and $\phi_{zz}^{(1)}$ into Eq. (7) (with $n = 2$), we obtain a rather unwieldy expression for $\phi^{(2)}$ which we shall not write out. Instead, we give here the result of averaging that expression. It is

$$\overline{\phi^{(2)}}(x, y, z) = (4\pi^2)^{-1} \iint_{-\infty}^{\infty} F(x - \xi, y - \eta, z) \iint_{-\infty}^{\infty} F(\xi', \eta', 0) \times [R_{\xi\xi}(\xi', \eta')\phi_{xx}^{(0)}(\xi + \xi', \eta + \eta', 0) + 2R_{\xi\eta}(\xi', \eta')\phi_{xy}^{(0)}(\xi + \xi', \eta + \eta', 0) + R_{\eta\eta}(\xi', \eta')\phi_{yy}^{(0)}(\xi + \xi', \eta + \eta', 0) - 2R_{\xi}(\xi', \eta')\phi_{xzz}^{(0)}(\xi + \xi', \eta + \eta', 0) - 2R_{\eta}(\xi', \eta')\phi_{yzz}^{(0)}(\xi + \xi', \eta + \eta', 0) + R(\xi', \eta')\phi_{zzzz}^{(0)}(\xi + \xi', \eta + \eta', 0)] \times d\xi' d\eta' d\xi d\eta. \quad (11)$$

The overbar denotes an ensemble average. In deriving Eq. (11) we have assumed that the random function μ is statistically homogeneous, and we have introduced the correlation function

$$R(\xi, \eta) = \overline{\mu(x, y)\mu(x + \xi, y + \eta)}. \quad (12)$$

The solution for $\overline{\phi}$, correct through terms of order ϵ^2, can now be obtained by substituting Eq. (11) into the expression obtained by averaging Eq. (3). This expression is

$$\overline{\phi} = \phi^{(0)} + \epsilon^2\overline{\phi^{(2)}} + O(\epsilon^3). \quad (13)$$

In deriving Eq. (13) we have used the fact that $\overline{\phi^{(1)}} = 0$ and that $\phi^{(0)}$ is nonrandom.

Equations (11) and (13) give the solution for the coherent wave, correct through terms of order ϵ^2, in terms of $\phi^{(0)}$, which we call the incident wave, and the correlation function R. This form of the solution is useful if we wish to find the coherent wave field resulting from the scattering by the boundary of a known incident wave. For certain purposes, however, it may be more convenient to formulate a boundary value problem for the coherent wave which does not involve the incident wave. This can be done by first noting that, as a consequence of Eq. (13),

$$\phi^{(0)} = \overline{\phi} + O(\epsilon^2).$$

Hence, we can replace $\phi^{(0)}$ by $\overline{\phi}$ in Eq. (11) without introducing an error of magnitude greater than $O(\epsilon^3)$ in Eq. (13). Next, we differentiate Eq. (13) with respect to z and set $z = 0$. By using Eq. (9) and the definition of the Green's function, and noting that $\phi_z^{(0)}(x, y, 0) = 0$, we obtain

$$\overline{\phi}_z(x, y, 0) = -\frac{\epsilon^2}{2\pi} \iint_{-\infty}^{+\infty} \rho^{-1} e^{ik_0\rho}[R_{\xi\xi}(\xi, \eta)\overline{\phi}_{xx}(x + \xi, y + \eta, 0) + 2R_{\xi\eta}(\xi, \eta)\overline{\phi}_{xy}(x + \xi, y + \eta, 0) + R_{\eta\eta}(\xi, \eta)\overline{\phi}_{yy}(x + \xi, y + \eta, 0) - 2R_{\xi}(\xi, \eta)\overline{\phi}_{xzz}(x + \xi, y + \eta, 0) - 2R_{\eta}(\xi, \eta)\overline{\phi}_{yzz}(x + \xi, y + \eta, 0) + R(\xi, \eta)\overline{\phi}_{zzzz}(x + \xi, y + \eta, 0)] d\xi d\eta. \quad (14)$$

Here $\rho = (\xi^2 + \eta^2)^{1/2}$, and we have neglected terms of order ϵ^3.

Equation (14) shows that the coherent wave satisfies a generalized impedance boundary condition on the average boundary $z = 0$. We call this a "smoothed" boundary condition because it is applied at the average, or smoothed, boundary, and also because the above analysis is

somewhat analogous to the smoothing method, which has been used to study wave propagation in random media.[14] Equation (14), together with Eq. (1) (with ϕ replaced by $\bar{\phi}$) and the radiation condition, constitutes the desired boundary value problem for the coherent wave.

II. PLANE WAVES

In order to make clear some of the implications of the smoothed boundary condition, we consider plane-wave solutions of the boundary value problem for $\bar{\phi}$. We being by assuming a solution of the form

$$\bar{\phi}(x, y, z) = \psi(z)e^{i(k_1x + k_2y)}. \tag{15}$$

Upon substituting Eq. (15) into Eq. (1) (after setting $f = 0$), we find that ψ must satisfy

$$\psi''(z) + m^2\psi(z) = 0, \tag{16}$$

where $m^2 = k_0^2 - k^2$, $k^2 = k_1^2 + k_2^2$, and the primes denote differentiation. Applying the smoothed boundary condition to Eq. 15 yields

$$\psi'(0) + \kappa\psi(0) = 0, \tag{17}$$

where

$$\kappa = \epsilon^2 K(\mathbf{k}, k_0), \tag{18}$$

$$K(\mathbf{k}, k_0) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int \rho^{-1}e^{ik_0\rho}\{m^4R(\xi, \eta) + 2im^2[k_1R_\xi(\xi, \eta) + k_2R_\eta(\xi, \eta)] - [k_1^2R_{\xi\xi}(\xi, \eta) + 2k_1k_2R_{\xi\eta}(\xi, \eta) + k_2^2R_{\eta\eta}(\xi, \eta)]\}\, e^{i(k_1\xi + k_2\eta)}d\xi\, d\eta, \tag{19}$$

and $\mathbf{k} = (k_1, k_2)$. The radiation condition is relaxed in this case, since the source is, in effect, at ∞. Instead, we require merely that the solution remain bounded as $z \to \infty$.

Note that the quantity κ/ik_0 may be regarded as the effective boundary admittance for plane waves.

Equation (16) has the solution

$$\psi(z) = A_-e^{-imz} + A_+e^{imz}, \tag{20}$$

where A_- and A_+ are undetermined constants. By invoking the boundary condition [Eq. (17)], we find that A_- and A_+ must satisfy

$$(\kappa - im)A_- + (\kappa + im)A_+ = 0. \tag{21}$$

If $\kappa \neq \pm im$, Eq. (21) shows that neither A_- nor A_+ can vanish (except in the trivial case); hence, the boundary condition at $z = \infty$ applied to Eq. (20) shows that m must be real. Consequently, k must be real, with $k^2 < k_0^2$. Thus, Eqs. (15) and (20) represent a system of incident and reflected waves.

The effective plane-wave reflection coefficient C_e of the boundary is defined to be the ratio of the reflected to the incident wave amplitude. Equation (21) shows that

$$C_e = \frac{1 + i\kappa/k_0\cos\chi}{1 - i\kappa/k_0\cos\chi}, \tag{22}$$

where χ, the angle of incidence, is equal to $\sin^{-1}(k/k_0)$. By neglecting terms of order ϵ^4, we can write Eq. (22) in the form

$$C_e = 1 + 2i\kappa/k_0\cos\chi. \tag{23}$$

Note that C_e depends on the direction of propagation as well as on the frequency and angle of incidence.

We can use Eq. (22) to relate the present theory to that of Twersky (Ref. 6) by noting that Eq. (22) is identical to Twersky's corresponding result [his Eq. (37)] provided that we set

$$i\kappa/k_0\cos\chi = Z_+. \tag{24}$$

The quantity Z_+ is related to Twersky's effective boundary impedance by his Eq. (38). Substituting for Z_+ in Eq. (24) from Eq. (76) of Ref. 6 yields

$$\pi\rho f = ik_0\kappa, \tag{25}$$

where, in the notation of Ref. 6, ρ is the average number of protuberances per unit area and f is the scattering amplitude of a single protuberance. Thus, we see that the present theory is formally equivalent to that of Ref. 6, at least insofar as plane-wave reflection is concerned, provided that we relate the quantity κ of the present theory to the product ρf of Ref. 6 through Eq. (25). Note also from Eq. (25) that the quantity $(i/\pi)k_0\kappa$ of the present theory may be regarded as an effective scattering amplitude per unit surface area.

If $\kappa = \pm im$, another type of solution of Eq. (16) may exist. In that case Eq. (20) becomes

$$\psi(z) = Ae^{-\kappa z}, \tag{26}$$

where A is an arbitrary constant. By applying the boundary condition at $z = \infty$ to Eq. (26) we see that a necessary condition for the existence of a nontrivial solution of this type is that Re $\kappa \geq 0$. Moreover, the vector $\mathbf{k}$ for this case must satisfy the equation

$$(k^2 - k_0^2)^{1/2} = \epsilon^2 K(\mathbf{k}, k_0). \tag{27}$$

(Here the square root is defined so that it has nonnegative real part.)

If Re$\kappa > 0$, the solution given by Eqs. (15) and (26) represents a surface wave, i.e., a wave whose amplitude decreases exponentially with distance from the boundary. We shall consider this type of wave in more detail in the isotropic case.

A. The isotropic case

Let us assume now that the function $\mu(x, y)$ is statistically isotropic, so that the correlation function can be written $R(\xi, \eta) = R(\rho)$. Then the angular integration in Eq. (19) can be carried out, after which the expression for K becomes

$$K(k, k_0) = \int_0^\infty e^{ik_0\rho}\{[m^4R(\rho) - k^2\rho^{-1}R'(\rho)]J_0(k\rho) + 2km^2R'(\rho)J_0'(k\rho) + k^2\rho[\rho^{-1}R'(\rho)]'J_0''(k\rho)\}d\rho. \tag{28}$$

Here J_n denotes the Bessel function of order n. By making use of known relationships involving derivatives of Bessel functions, we can write Eq. (28) in the alternate form

$$K(k, k_0) = \int_0^\infty e^{ik_0\rho}\ (\{m^4R(\rho) - k^2\rho^{-1}R'(\rho)]J_0(k\rho) - k\{2m^2R'(\rho) + [\rho^{-1}R'(\rho)]'\}J_1(k\rho) + k^2\rho[\rho^{-1}R'(\rho)]' \times J_2(k\rho))d\rho. \tag{29}$$

Equation (28) shows that K, and therefore κ, depends only on the magnitude of the vector $\mathbf{k}$. In this case, then, C_e is independent of the direction of propagation.

We can now solve Eq. (27) by assuming a solution of the form

$$k = \sum_{n=0}^{\infty} \epsilon^n k^{(n)}, \tag{30}$$

where $k^{(0)} = k_0$. Upon substituting Eq. (30) into Eq. (27), expanding in powers of ϵ, and collecting terms in like power of ϵ, we find that $k^{(1)} = k^{(2)} = k^{(3)} = 0$, while $k^{(4)}$ is given by

$$[2k_0 k^{(4)}]^{1/2} = K(k_0, k_0). \tag{31}$$

Hence, if $\mathrm{Re}K(k_0, k_0) > 0$, a surface-wave solution exists. By inserting the expression for $k^{(4)}$ obtained from Eq. (31) into Eq. (30), we find that

$$k_* = k_0 + (\epsilon^4/2k_0)K^2(k_0, k_0), \tag{32}$$

where k_* is the horizontal propagation constant of the surface wave. The corresponding value of κ is obtained from Eq. (18) and is given by

$$\kappa_* = \epsilon^2 K(k_0, k_0). \tag{33}$$

In deriving Eqs. (32) and (33) we have neglected higher-order terms in ϵ.

B. Long waves

We now assume that $k_0 l \ll 1$, where l is the correlation length of the boundary roughness. We assume also that k is at most of order k_0. Then we can get an approximate expression for $K(k, k_0)$ by expanding the term $e^{ik_0\mu}$, as well as the terms involving the Bessel functions, in Eq. (28) in power series and integrating the first few terms. The result is

$$K(k, k_0) = \tfrac{1}{2}k^2 N_0 + (k_0^4 - \tfrac{3}{4}k_0^2 k^2 + \tfrac{5}{16}k^4)M_0 + ik_0^3(k_0^2 + \tfrac{1}{3}k^2)M_1 + O(k_0^6), \tag{34}$$

where

$$N_0 = -\int_0^\infty \rho^{-1} R'(\rho)\, d\rho \tag{35}$$

and

$$M_n = \int_0^\infty \rho^n R(\rho)\, d\rho, \qquad n = 0, 1, 2, \cdots. \tag{36}$$

The value of C_e for this case is obtained with the aid of Eqs. (18), (23), and (34). Upon dropping terms of order k_0^6, we get

$$C_e = 1 - 2\epsilon^2 k_0^4 M_1(1 + \tfrac{1}{3}\sin^2\chi)\sec\chi + 2i\epsilon^2 k_0 \times [\tfrac{1}{2}N_0 \sin^2\chi + k_0^2 M_0(1 - \tfrac{3}{4}\sin^2\chi + \tfrac{5}{16}\sin^4\chi)]\sec\chi. \tag{37}$$

From Eq. (37) we find that, to lowest order in ϵ,

$$|C_e| = 1 - 2\epsilon^2 k_0^4 M_1(1 + \tfrac{1}{3}\sin^2\chi)\sec\chi. \tag{38}$$

In Appendix A we show that $M_1 \geq 0$. Here we shall assume that $M_1 > 0$, which is the case for most correlation functions of interest. As a consequence, we see from Eq. (38) that $|C_e| < 1$, i.e., the amplitude of the coherent wave is diminished upon reflection from the boundary. This is the result of scattering of energy out of the coherent wave by the boundary roughness. We see also from Eq. (38) that this effect increases with χ, i.e., the loss of coherent energy upon reflection increases with the angle of incidence.

Equation (38) shows that $|C_e|$ increases as k_0 decreases with χ fixed. Thus, at any angle of incidence, longer waves are reflected more coherently.

We show also in Appendix A that both M_0 and N_0 are positive. Hence, we see from Eq. (37) that $\mathrm{Im}C_e > 0$, and therefore that the coherent wave undergoes a phase retardation upon reflection.

We can get a condition for the validity of Eq. (37) as follows. We introduce h, the rms boundary roughness, which is defined by

$$h^2 = \epsilon^2 R(0); \tag{39}$$

also, we note that

$$M_n = O[l^{n+1}R(0)], \qquad n = 0, 1, 2\cdots, \tag{40}$$

and that

$$N_0 = O[l^{-1}R(0)]. \tag{41}$$

We now use Eqs. (39)–(41) to estimate the magnitudes of the terms occurring in Eq. (37), and note that the condition $C_e \simeq 1$ is necessary for the validity of the perturbation analysis leading to this equation. Thus we find that Eq. (37) is valid only if

$$k_0 h^2/l \ll \cos\chi. \tag{42}$$

Hence we must have

$$k_0 h^2/l \ll 1. \tag{43}$$

Moreover, Eq. (42) shows that the formula for C_e given by Eq. (37) is not valid at incidence angles too near grazing, i.e., for $\cos\chi \lesssim k_0 h^2/l$.

We showed previously that a surface-wave solution exists if $\mathrm{Re}K(k_0, k_0) > 0$. By setting $k = k_0$ in Eq. (34) and neglecting terms of order k_0^6 we find that

$$K(k_0, k_0) = \tfrac{1}{2}k_0^2 N_0 + \tfrac{9}{16}k_0^4 M_0 + \tfrac{4}{3}ik_0^5 M_1. \tag{44}$$

Equation (44) shows that $\mathrm{Re}K(k_0, k_0) > 0$; hence, a surface-wave solution exists in this case. The horizontal propagation constant k_* of the surface wave is obtained by substituting Eq. (44) into Eq. (32) and neglecting terms of order k_0^7. The result is

$$k_* = k_0[1 + \tfrac{1}{2}\epsilon^4 N_0 k_0^2(\tfrac{1}{4}N_0 + \tfrac{9}{16}M_0 k_0^2 + \tfrac{4}{3}iM_1 k_0^3)]. \tag{45}$$

With the aid of Eqs. (15), (26), (33), (44), and (45) we see that the surface wave propagates at a slight downward angle toward the boundary, with a speed slightly less than the free-space propagation speed. Since the propagation speed depends on k_0, this type of wave is dispersive. Also, Eq. (45) shows that the amplitude of the surface wave decreases with horizontal propagation distance. Again, this is the result of scattering of energy out of the coherent wave by the boundary roughness.

A necessary condition for the validity of the perturbation analysis leading to Eq. (45) is that $k_* \simeq k_0$. Thus, by applying Eqs. (39)–(41) to Eq. (45), we see that the condition given by Eq. (43) is also necessary for the validity of the surface-wave analysis.

C. Short waves

We consider now the case in which $k_0 l >> 1$. We shall study first the system of incident and reflected waves.

We begin by substituting the integral expression

$$J_0(k\rho) = \frac{1}{2\pi}\int_0^{2\pi} e^{ik\rho\cos\theta}\, d\theta \tag{46}$$

for the Bessel function into Eq. (28) and reversing the order of integration. Next we use integration by parts

to simplify some of the resulting integrals with respect to ρ, after which Eq. (28) can be written

$$K(k, k_0) = \frac{1}{2\pi} \int_0^{2\pi} \{[m^2 + (k_0 + k\cos\theta)\, k\cos\theta]^2 P(k_0 + k\cos\theta) + k^2 \sin^2\theta Q(k_0 + k\cos\theta)\}\, d\theta - \tfrac{1}{2} ik_0 k^2 R(0). \quad (47)$$

Here

$$P(\lambda) = \int_0^\infty e^{i\lambda\rho} R(\rho) d\rho, \quad (48)$$

$$Q(\lambda) = -\int_0^\infty e^{i\lambda\rho} \rho^{-1} R'(\rho) d\rho. \quad (49)$$

We now assume that k, as well as k_0, is large, and that $k_0 \pm k = O(k_0)$. Then the term $k_0 + k\cos\theta$ in Eq. (47) is of order k_0; hence we can substitute for the functions P and Q in this equation their respective asymptotic expansions for large values of the argument. These expansions are obtained from Eqs. (48) and (49) by successive integration by parts, and are given by

$$P(\lambda) = iR(0)\lambda^{-1} - iR''(0)\lambda^{-3} + O(\lambda^{-5}), \quad (50)$$

$$Q(\lambda) = -iR''(0)\lambda^{-1} + O(\lambda^{-3}). \quad (51)$$

Upon substituting Eqs. (50) and (51) into Eq. (47) and integrating over θ we obtain, after some algebra,

$$K(k, k_0) = i\{R(0)(k_0^2 - k^2)^{3/2} - R''(0)(k_0^2 - \tfrac{1}{2}k^2)(k_0^2 - k^2)^{-1/2} + O(k_0^{-1})\}. \quad (52)$$

We can get an expression for C_e by using Eqs. (18), (23), and (52). In terms of the incidence angle it is

$$C_e = 1 - 2\epsilon^2[k_0^2 R(0)\cos^2\chi - R''(0)(1 - \tfrac{1}{2}\sin^2\chi)\sec^2\chi]. \quad (53)$$

Here we have neglected terms of order k_0^{-2}.

We see from Eq. (53) that, to this degree of approximation, C_e is real. Thus, short waves reflect with no phase shift. We see also that $|C_e| < 1$, which shows that, as in the long wave case, coherent energy is lost upon reflection.

The presence of the term $\sec^2\chi$ in Eq. (53) shows that this equation is not valid for incidence angles too near grazing, i.e., for $\cos\chi \lesssim (k_0 l)^{-1/2}$. Away from these angles, however, we can neglect the term involving $\sec^2\chi$, after which Eq. (53) can be written

$$C_e = 1 - 2k_0^2 h^2 \cos^2\chi. \quad (54)$$

This expression for C_e is identical to the one obtained by Rice[15] for the corresponding electromagnetic problem. It also agrees with the first two terms of the expansion in powers of $k_0 h\cos\chi$ of an expression for the reflection coefficient obtained by Tolstoy and Clay[16] using the Kirchhoff approximation.

The condition $C_e \simeq 1$ is necessary for the validity of the analysis leading to Eq. (54). Hence we must have $k_0 h \ll 1$. But since $k_0 l \gg 1$ for short waves, we see that a necessary condition for the validity of the short-wave analysis is that

$$h/l \ll 1, \quad (55)$$

i.e., the characteristic surface slope must be small.

As we have already noted, Eqs. (53) and (54) are not valid at incidence angles too near grazing. In order to consider propagation at such angles we turn now to the surface-wave analysis. The existence of a surface wave depends, as we have seen, on the sign of $\mathrm{Re}K(k_0, k_0)$.

For the case of short waves, $K(k_0, k_0)$ is found by setting $k = k_0$ in Eq. (28) and using the results of Appendix B to evaluate the resulting integrals approximately for large k_0. The result is

$$K(k_0, k_0) = k_0^2[\tfrac{1}{4}(\pi k_0)^{-1/2}(1 + i)N_1 + O(k_0^{-1})], \quad (56)$$

where

$$N_1 = -\int_0^\infty \rho^{-3/2} R'(\rho) d\rho. \quad (57)$$

It is shown in Appendix A that N_1 is positive. Hence, Eq. (56) shows that $\mathrm{Re}K(k_0, k_0) > 0$, and so a surface-wave solution exists in this case.

The horizontal propagation constant k_* for the surface wave is obtained by substituting Eq. (56) into Eq. (32), after dropping the term of order k_0^{-1} in the brackets in Eq. (56). The result is

$$k_* = k_0 + i\epsilon^4 (N_1^2/16\pi) k_0^2. \quad (58)$$

Equations (15), (26), (33), (56), and (58) show that the surface wave has the same properties as in the long-wave case, except that here its horizontal phase speed is equal to the free-space propagation speed.

A necessary condition for the validity of the analysis leading to Eq. (58) is that $k_* \simeq k_0$. By applying Eq. (39) to Eq. (58) and noting that

$$N_1 = O[l^{-3/2} R(0)],$$

we find that the condition given by Eq. (55) is also necessary for the validity of the surface-wave analysis.

ACKNOWLEDGMENTS

Research supported by the Atmospheric Sciences Section, National Science Foundation, NSF Grant GA-31561. The author is grateful to Professor Joseph B. Keller of the Courant Institute of Mathematical Sciences, New York University, for his helpful comments on this paper.

APPENDIX A

Here we show that $M_0 > 0$ and that $M_1 \geq 0$, where, for $n = 0, 1, 2, \cdots$,

$$M_n = \int_0^\infty r^n R(r) dr. \quad (A1)$$

We show also that both N_0 and N_1 are positive, where

$$N_0 = -\int_0^\infty r^{-1} R'(r) dr, \quad (A2)$$

$$N_1 = -\int_0^\infty r^{-3/2} R'(r) dr. \quad (A3)$$

We assume that the correlation function R is such that these integrals, as well as any others appearing in the following analysis, exist.

We begin by introducing S, the Hankel transform of R, where

$$S(s) = \int_0^\infty rR(r) J_0(rs) dr, \quad (A4)$$

$$R(r) = \int_0^\infty sS(s) J_0(rs) ds. \quad (A5)$$

Since $S(s)$ is the Fourier transform of a correlation function, it is nonnegative. Hence, we see immediately from Eq. (A4) that

$$M_1 = S(0) \geq 0. \quad (A6)$$

To show that $M_0 > 0$, we first define the function $p(\alpha)(\alpha > 0)$ by

$$p(\alpha) = \int_0^\infty e^{-\alpha r} R(r) dr. \tag{A7}$$

Next we substitute Eq. (A5) into Eq. (A7) and reverse the order of integration. This yields

$$p(\alpha) = \int_0^\infty sS(s) \int_0^\infty e^{-\alpha r} J_0(rs) dr\, ds. \tag{A8}$$

The integral over r in Eq. (A8) is absolutely convergent, which justifies changing the order of integration. Upon evaluating this integral[17] we get

$$p(\alpha) = \int_0^\infty s(\alpha^2 + s^2)^{-1/2} S(s) ds. \tag{A9}$$

If we now let $\alpha \to 0$ and use the Lebesgue convergence theorem,[18] we obtain from Eqs. (A7) and (A9)

$$M_0 = \int_0^\infty S(s) ds. \tag{A10}$$

Since S can not vanish identically (unless R does), Eq. (A10) shows that $M_0 > 0$.

To show that $N_0 > 0$ we assume that $R(r)$ has a convergent power series expansion in some interval $0 \le r \le c$, $c > 0$. Then since R is an even function we have, for r in this interval,

$$R(r) = \sum_{n=0}^\infty a_{2n} r^{2n}. \tag{A11}$$

We assume that $a_2 \neq 0$, which implies that $a_2 < 0$. Now for any $b > 0$ such that $b \le c$, we can write

$$N_0 = -\int_0^b r^{-1} R'(r) dr - \int_b^\infty r^{-1} R'(r) dr. \tag{A12}$$

We can evaluate the first integral on the right-hand side of Eq. (A12) by substituting for R its power series expansion and integrating term by term. This gives

$$\int_0^b r^{-1} R'(r) dr = \sum_{n=1}^\infty \frac{2n}{2n-1} a_{2n} b^{2n-1}. \tag{A13}$$

We can transform the second integral on the right-hand side of Eq. (A12) by integrating by parts. This yields

$$\int_b^\infty r^{-1} R'(r) dr = -b^{-1} R(b) + \int_b^\infty r^{-2} R(r) dr. \tag{A14}$$

Now $R(r) \le R(0)$; hence we can write

$$\int_b^\infty r^{-2} R(r) dr \le R(0) \int_b^\infty r^{-2} dr = b^{-1} R(0). \tag{A15}$$

Then from Eq. (A14) we have

$$\int_b^\infty r^{-1} R'(r) dr \le b^{-1}[R(0) - R(b)]. \tag{A16}$$

We now use the power series expansion of R [Eq. (A11)] to substitute for the term on the right-hand side of Eq. (A16), after which this equation, together with Eqs. (A12) and (A13), shows that

$$N_0 \ge -\sum_{n=1}^\infty \frac{a_{2n}}{2n-1} b^{2n-1}. \tag{A17}$$

Since $a_2 < 0$ we can write Eq. (A17) in the form

$$N_0 \ge -a_2 b \left(1 + \sum_{n=2}^\infty \frac{a_{2n}/a_2}{2n-1} b^{2n-2}\right). \tag{A18}$$

We now choose b so small that

$$\left| \sum_{n=2}^\infty \frac{a_{2n}/a_2}{2n-1} b^{2n-2} \right| < 1,$$

after which the result follows.

Note that the same result is obtained if $a_2 = a_4 = \cdots = a_{2j} = 0; j \ge 1$, and $a_{2j+2} < 0$.

A similar argument shows that $N_1 > 0$.

APPENDIX B

In order to obtain the short-wave approximation of $K(k_0, k_0)$ from Eq. (32), it is necessary to evaluate integrals of the form

$$I = \int_0^\infty e^{ik_0 r} D(r) J_0^{(n)}(k_0 r) dr \tag{B1}$$

for large k_0, with $n = 0, 1, 2$. Here $J_0^{(n)}$ is the nth derivative of the Bessel function of order zero, and the function $D(r)$ is assumed to be such that all the integrals appearing in the analysis exist.

We can transform the integral of Eq. (B1) by substituting for $J_0^{(n)}$ the expression

$$J_0^{(n)}(kr) = \frac{(-i)^n}{\pi} \int_0^\pi e^{-ikr\cos\theta} \cos^n\theta\, d\theta$$

(we drop the subscript on k) and reversing the order of integration. This yields

$$I = \frac{(-i)^n}{\pi} \int_0^\pi f(\theta) \cos^n\theta\, d\theta, \tag{B2}$$

where

$$f(\theta) = \int_0^\infty \exp[ikr(1 - \cos\theta)] D(r)\, dr. \tag{B3}$$

We note that

$$f(0) = \int_0^\infty D(r) dr$$

for all k; whereas if k is large, $f(\theta)$ is small for $\theta > 0$. Consequently, we expect that, for large k, most of the contribution to the integral of Eq. (B2) will come from the immediate neighborhood of $\theta = 0$. This suggests using a method similar to the stationary phase method to evaluate this integral.

We begin by writing Eq. (B2) in the form

$$I = [(-i)^n/\pi](I_1 + I_2), \tag{B4}$$

where

$$I_1 = \int_0^{\pi/2} f(\theta) \cos^n\theta\, d\theta, \tag{B5}$$

$$I_2 = \int_{\pi/2}^\pi f(\theta) \cos^n\theta\, d\theta. \tag{B6}$$

Now if $\pi/2 \le \theta \le \pi$ then $1 - \cos\theta \ge 1$. Hence in this interval we can get an asymptotic expansion for $f(\theta)$ by integrating by parts in Eq. (B3). This expansion shows that

$$f(\theta) = O[k^{-1}(1 - \cos\theta)^{-1}]$$

in this interval, and therefore [from Eq. (B6)] that

$$I_2 = O(k^{-1}). \tag{B7}$$

We consider next the integral I_1. We introduce the variable s, defined by

$$s^2 = 1 - \cos\theta, \quad 0 \le \theta \le \pi/2,$$

and write Eq. (B5) in terms of an integral over s. This yields

$$I_1 = 2^{1/2} \int_0^1 (1 - s^2)^n (1 - \tfrac{1}{2}s^2)^{-1/2} g(s) ds, \tag{B8}$$

where

$$g(s) = \int_0^\infty e^{iks^2 r} D(r) dr. \tag{B9}$$

By expanding $(1 - s^2)^n$ and $(1 - \frac{1}{2}s^2)^{-1/2}$ in power series we see that, for $0 \leq s \leq 1$,

$$(1 - s^2)^n(1 - \tfrac{1}{2}s^2)^{-1/2} = 1 + s^2p(s), \tag{B10}$$

where $p(s)$ is continuous in the interval $[0, 1]$. Hence we can write Eq. (B8) in the form

$$I_1 = 2^{1/2}\left(\int_0^1 g(s)ds + \int_0^1 s^2p(s)g(s)ds\right). \tag{B11}$$

Integrating by parts in Eq. (B9) shows that $g(s) = O(k^{-1}s^{-2})$, and therefore that $s^2g(s) = O(k^{-1})$. Consequently,

$$\int_0^1 s^2p(s)g(s)ds = O(k^{-1}). \tag{B12}$$

We now introduce t, a new variable of integration, defined by $t^2 = ks^2$. In terms of t we have

$$\int_0^1 g(s)ds = k^{-1/2}\int_0^{k^{1/2}} h(t)dt, \tag{B13}$$

where

$$h(t) = \int_0^\infty e^{it^2r}D(r)dr. \tag{B14}$$

Equation (B13) can be written

$$\int_0^1 g(s)ds = k^{-1/2}\left(\int_0^\infty h(t)dt - \int_{k^{1/2}}^\infty h(t)dt\right). \tag{B15}$$

The first integral on the right hand side of Eq. (B15) can be evaluated by substituting for $h(t)$ from Eq. (B14), reversing the order of integration, and noting that the resulting integral over t is tabulated. This yields

$$\int_0^\infty h(t)dt = (\pi/8)^{1/2}(1 + i)\int_0^\infty r^{-1/2}D(r)dr. \tag{B16}$$

The second integral on the right-hand side of Eq(B15) is easily shown to be of order $k^{-1/2}$ by first integrating by parts in Eq. (B14) and then substituting the resulting expression for $h(t)$ into the integral. Thus we find that

$$\int_0^1 g(s)ds = (\pi/8k)^{1/2}(1 + i)\int_0^\infty r^{-1/2}D(r)dr + O(k^{-1}). \tag{B17}$$

We now collect our results [Eqs. (B7), (B12), and (B17)] and, with the aid of Eqs. (B4) and (B11), we obtain finally

$$I = (-i)^n(4\pi k)^{-1/2}(1 + i)\int_0^\infty r^{-1/2}D(r)dr + O(k^{-1}). \tag{B18}$$

*Present address: NASA Ames Research Center, Moffett Field, California 94035.

[1]V. Twersky, J. Acoust. Soc. Amer. **29**, 209 (1957).
[2]V. Twersky, IRE Trans. Ant. Prop. **AP-5**, 81 (1957).
[3]M. A. Biot, J. Acoust. Soc. Amer. **44**, 1616 (1968).
[4]J. E. Burke and V. Twersky, J. Acoust. Soc. Amer. **40**, 883 (1966).
[5]V. Twersky, J. Opt. Soc. Amer. **52**, 145 (1962).
[6]T. B. A. Senior, Appl. Sci. Res. **8B**, 437 (1960).
[7]Y. P. Lysanov, Sov. Phys.-Acoust. **15**, 340 (1970).
[8]G. G. Zipfel, Jr., and J. A. DeSanto, J. Math. Phys. **13**, 1903 (1972).
[9]J. B. Keller (unpublished manuscript).
[10]J. B. Keller, Proc. Symp. Appl. Math. **16**, 145 (1964).
[11]P. Beckmann and A. Spizzichino, *The Scattering of Electromagnetic Waves from Rough Surfaces* (MacMillan, New York, 1963).
[12]L. Fortuin, J. Acoust. Soc. Amer. **47**, 1209 (1970).
[13]C. W. Horton, J. Acoust. Soc. Amer. **51**, 1049 (1972).
[14]U. Frisch in *Probabilistic Methods in Applied Mathematics*, edited by A. T. Barucha–Reid (Academic, New York, 1968), p. 114.
[15]S. O. Rice, Commun. Pure Appl. Math. **4**, 351 (1951), Eq. (5.2).
[16]I. Tolstoy and C. S. Clay, *Ocean Acoustics* (McGraw-Hill, New York, 1966), Eq. (6.44b).
[17]G. N. Watson, *Theory of Bessel Functions* (MacMillan, New York, 1945), p. 384.
[18]R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1953), Vol. I, p. 109.

A one-dimensional fermion model exhibiting an anomalous type of phase transition*

R. Moscheo†

Istituto di Fisica dell'Universitá and GNSM del CNR, Messina, Italy

R. P. Moya‡

Physics Department, Queen Mary College, University of London, England
(Received 9 July 1973)

We construct a one-dimensional fermion model (rigorously reducible to a mean field theory). We show that this model exhibits a second order phase transition associated with a spontaneous breakdown of continuous space translational symmetry in favor of a periodic symmetry. However, Landau and Lifshitz conjectured that a phase transition in which there is a spontaneous breakdown of Euclidean symmetry in favor of a crystallographic symmetry must be a transition of the first order. Thus we obtain a counterexample to this conjecture in the case of one dimension.

1. INTRODUCTION

It has been observed experimentally that the solid–liquid phase transition is always a transition of the first order. In particular, this transition is accompanied by a non zero latent heat. Furthermore, the solid–liquid transition is always associated with a spontaneous breakdown of Euclidean symmetry in favor of a crystallographic symmetry appropriate to the solid phase. It has been conjectured theoretically by Landau and Lifshitz[1] that a phase transition which may be characterized by a symmetry breakdown of this type must be a transition of the first order.[2] In this paper we put the above conjecture to the test by examining the equilibrium states of an exactly solvable one–dimensional fermion model. Our work is based on a rigorous treatment of the model using the C^*-algebraic approach[3] to quantum statistical mechanics. As the mathematical formulation is rather abstract, we will present here a heuristic definition of the model, the rigorous treatment being deferred to the next section. The model is essentially an infinite volume generalization of the model defined by the following finite volume Hamiltonian

$$H_v = H_{0,v} + H_{I,v}, \tag{1}$$

where $H_{0,v}$ is the free Hamiltonian for the finite volume v, and the interaction Hamiltonian $H_{I,v}$ is a "truncated" form of the following

$$g/v \int_v \int \Psi^*(x)\Psi^*(y)e^{2ik_F(x-y)}\Psi(x)\Psi(y)dxdy, \tag{2}$$

where g is a negative coupling constant, Ψ^* and Ψ denote the Fermi annihilation and creation operators, respectively, $k_F \neq 0$ denotes the Fermi momentum. Crudely speaking, the word "truncated" means that only those fermions with momentum within a region close to the Fermi surface interact. Motivated by Haag's[4] treatment of the B.C.S. model we will actually consider an infinite volume generalization of the model determined by the following Hamiltonian, rather than that obtained from H_v of equation (1),

$$H_{\pi,v} = H_{\pi,0,v} + gQ_\pi \int_v \Psi_\pi^*(x)\underline{v}^1\Psi_\pi(x)dx + gQ_\pi^* \int_v \Psi_\pi^*(x)\underline{v}^2\Psi_\pi(x)dx, \tag{3}$$

where this Hamiltonian is to be regarded as an operator with dense domain in some Hilbert space $\mathfrak{H}_\pi$, corresponding to the representation π. $H_{\pi,0,v}$ represents the free Hamiltonian and Q_π is defined as follows:

$$Q_\pi = \lim_{v\to\infty} \frac{1}{v} \int_v \Psi_\pi^*(x)\underline{v}^2\Psi_\pi(x)dx \tag{4}$$

on a suitable dense domain in the representation space. The representation π is to be chosen such that this limit exists. $\underline{v}^1$ and $\underline{v}^2$ are "truncated" forms, in the above-mentioned sense, of $\exp(2iK_Fx)$ and $\exp(-2iK_Fx)$, respectively.

Following arguments similar to those used by Haag[4] for the B.C.S. model, we see that Q_π belongs to the center of the von Neumann algebra[3] $\pi(\mathfrak{A})''$, where $\mathfrak{A}$ denotes the algebra of observables for the system. We emphasize at this point that the foregoing heuristic description was designed to give some insight into the abstract mathematical definitions of the next section. In particular, we have set this description out for finite volume, whereas the rigorous treatment will be given for an infinite volume; at no stage will we resort to the so-called thermodynamic limit.

It may be seen that our model is in some sense similar to the one-dimensional electron–phonon model studied by Fröhlich[5] in connection with the theory of super conductivity. The manner in which we formulate our model is such that it reduces rigorously to a mean field theory.[6] The amplitude of this mean field essentially plays the role of an order parameter. Further, the mean field is periodic. We shall adopt the pragmatic point of view that the equilibrium states for our model are determined by the Kubo–Martin–Schwinger (K.M.S.)[7,8] conditions. We mention here that these conditions are a natural generalization to infinite volume of the usual finite volume Gibbs equilibrium presciption. As a result we shall show that our model exhibits a phase transition of the second order associated with a spontaneous breakdown of continuous space translational symmetry in favor of a periodic symmetry. The transition is of the second order in the sense that the above-mentioned order parameter tends continuously to zero as the temperature tends continuously to the critical temperature from below. Thus we provide a counter example to Landau and Lifshitz's conjecture in the case of one dimension.

Our paper is set out as follows; in Sec. 2 we present the mathematical definition of the model. Our formulation leads to a "self-consistency" equation for the order parameter. In Sec. 3 we discuss the solutions of this equation. We also prove the existence and determine the order of the phase transition. Section 4 contains our conclusions. Appendix A is devoted to the mathematical calculations leading to the self-consistency equation.

2. THE MATHEMATICAL DEFINITION OF THE MODEL PRELIMINARIES

Definition: Let $\mathcal{K}$ be a complex Hilbert space with inner product given by $\langle \ldots, \ldots \rangle$

Then we define a representation of the canonical anticommutation relations (C.A.R.) over $\mathcal{K}$ to be a linear mapping

$$f \to \Psi(f), \forall f \in \mathcal{K},$$

from $\mathcal{K}$ onto operators $\Psi(f)$ on a Hilbert space $\mathfrak{H}$, such that

$$\{\Psi(f), \Psi(g)\}_+ = \{\Psi(f)^*, \Psi(g)^*\}_+ = 0$$

and

$$\{\Psi(f)^*, \Psi(g)\}_+ = \langle f, g \rangle, \quad \forall f, g \in \mathcal{K}.$$

In the following we shall take $\mathcal{K}$ to be single particle Hilbert space $\mathcal{L}^2(\mathbb{R})$ of square integrable functions on the real line $\mathbb{R}$. In particular, we have the usual Fock[3] representation of the C.A.R. over $\mathcal{K}$, i.e.,

$$f \to \Psi_F(f), \quad \forall f \in \mathcal{K},$$

where the $\Psi_F(f)^*$ and $\Psi_F(f)$ are operators satisfying the above anticommutation relations on the Fock–Hilbert space $\mathfrak{H}_F$. Now define $\tilde{\mathfrak{A}}$ to be the algebra generated by polynomials in the $\Psi_F(f)^*$, $\Psi_F(g)$, $\forall f, g \in \mathcal{K}$. We then take as C^*-algebra of observables $\mathfrak{A}$, the norm closure of $\tilde{\mathfrak{A}}$.

Space Translations

Define $y \in \mathbb{R} \to U_y$, a mapping of the additive group of reals into the set of unitary operators on $\mathcal{K}$ by the following:

$$U_y f(x) = f(x - y), \quad \forall f \in \mathcal{K}, \quad x, y \in \mathbb{R}.$$

Define $y \in \mathbb{R} \to \alpha_y$, a mapping of the additive group of reals into Aut$\tilde{\mathfrak{A}}$ by

$$\alpha_y(\Psi_F(f)) = \Psi_F(U_y f), \quad \forall f \in \mathcal{K}, \quad y \in \mathbb{R}.$$

We may extend α to a one-parameter group of strongly continuous automorphisms of $\mathfrak{A}$ in the usual manner. We shall denote this extension by α also.

We now present several definitions which we will require in order to define the dynamics for our model.

Definition: We define the free single particle Hamiltonian $\tilde{H}_0$ on the single particle Hilbert space as follows

$$\widehat{\tilde{H}_0 f}(k) = \omega(k)\hat{f}(k), \quad \forall f \in \mathcal{S}(\mathbb{R}),$$

where $\mathcal{S}(\mathbb{R})$ denotes the Schwartz space of infinitely differentiable functions of fast decrease on $\mathbb{R}$, $\hat{}$ denotes Fourier transformation, and $\omega(k) = C|k|$, $0 < C < \infty$. It may easily be shown that $\tilde{H}_0$ is essentially self adjoint on $\mathcal{S}(\mathbb{R})$ and so possesses a unique self-adjoint extension; denote this extension by $\tilde{H}_0$ also. We shall denote $\tilde{H}_0 - \mu$ by H_0, with μ the chemical potential in the range $(-\infty, cK_F]$, K_F the fermi-momentum, $K_F > 0$.

Definition: Let $Q \in \mathbb{C}$ (the complex plane), $K_F \in \mathbb{R}^+$, $K_F > 0$, $G_1 = (0, 2K_F]$, $G_2 = (-2K_F, 0]$, and let χ_{G_1}, χ_{G_2} denote the characteristic functions for the regions G_1 and G_2, respectively. We define the bounded self-adjoint operator v^Q on $\mathcal{K}$ as follows:

$$\widehat{V^Q f}(K) = Q\widehat{V^1 f}(K) + \bar{Q}\widehat{V^2 f}(K)$$

with

$$\widehat{V^1 f}(K) = \chi_{G_1}(K)\hat{f}(K - 2K_F), \qquad (5)$$

$$\widehat{V^2 f}(K) = \chi_{G_2}(K)\hat{f}(K + 2K_F), \quad \forall f \in \mathcal{K}$$

where $\bar{}$ denotes complex conjugation. These definitions make precise what we have previously referred to as "truncated." It follows easily from the definitions that

$$U_y V^1 = e^{-2iK_F y} V^1 U_y$$

and

$$U_y V^2 = e^{2iK_F y} V^2 U_y, \quad \forall y \in \mathbb{R}.$$

Also if $y \to Q_y$, a mapping of $\mathbb{R}$ into $\mathbb{C}$ is defined by

$$Q_y = e^{-2iK_F y} Q,$$

then

$$U_y V^Q U_{-y} = V^{Q_y}. \qquad (6)$$

Definition: We define H_Q to be the self-adjoint closure of $H_0 + gV^Q$ with $g \in \mathbb{R}^-$, $Q \in \mathbb{C}$. We shall also define the mapping $t \to T_t^Q$ of $\mathbb{R}$ into the unitary operators of $\mathcal{K}$ by

$$T_t^Q = \exp\{iH_Q t\}, \quad t \in \mathbb{R}.$$

Notation

Denote by Γ the set of all bounded open measurable subsets Λ of $\mathbf{R}$ and by $V(\Lambda)$ the Lesbegue measure (volume) of the region Λ.

The model

Having specified the observables it remains to specify the dynamics and a set of physical states; in our case equilibrium states. We will define the dynamics and the equilibrium states in a "self-consistent" manner. We do this by defining a set of states $\mathcal{E}$ which satisfy the following three conditions. Then we evaluate the set $\mathcal{E}$ explicitly and will interpret $\mathcal{E}$ as the set of equilibrium states for our model.

The set $\mathcal{E}$: Let $\mathcal{E}$ be the set of states for the C^*-algebra $\mathfrak{A}(\mathcal{K})$ which satisfies the following three conditions for temperature β^{-1} and chemical potential μ, $\mu \leq cK_F$. Denote by $(\mathfrak{H}_\omega, \pi_\omega, \Omega_\omega)$ the G.N.S.[3] triple associated with ω.

(1) There exists a strongly continuous one parameter group τ_t, $t \in \mathbb{R}$ of automorphisms of $\pi_\omega(\mathfrak{A}(\mathcal{K}))''$, the bicommutant of[9] $\pi_\omega(\mathfrak{A}(\mathcal{K}))$.

(2) $\tilde{\omega}$ the canonical extension of ω to $\pi_\omega(\mathfrak{A}(\mathcal{K}))''$ must satisfy the K.M.S. conditions with respect to τ_t at temperature β^{-1}, $\beta > 0$, and chemical potential μ, $\mu \leq cK_F$, c and $K_F \neq 0$.

(3) Let $\omega = \int \omega_\sigma d\mu_{\omega_\sigma}$ be the central decomposition of ω and let $(\mathfrak{H}_{\omega_\sigma}\pi_{\omega_\sigma}\Omega_{\omega_\sigma})$ denote the G.N.S. triple associated with ω_σ. It may be shown that $\tilde{\omega}_\sigma$ is an extremal K.M.S. state. Now define the mapping $t \to \hat{\tau}_t$ of $\mathbb{R}$ into Aut$\pi_{\omega_\sigma}(\tilde{\mathfrak{A}}(\mathcal{K}))$ by

$$\hat{\tau}_t(\Psi_\sigma(f)) = \Psi_\sigma(T_t^{Q_\sigma} f), \quad \forall f \in \mathcal{K},$$

with $\Psi_\sigma(f) = \pi_{\omega_\sigma}(\Psi_F(f))$. Our third condition is that τ_t, for the representation π_{ω_σ}, coincides with $\hat{\tau}_t$ on

$\pi_{\omega_\sigma}(\mathfrak{A}(\mathcal{K}))$. It then follows from Condition (2) that we may uniquely extend $\hat{\tau}_t$ to $\pi_{\omega_\sigma}(\mathfrak{A}(\mathcal{K}))''$. Furthermore, to complete Condition (3), we require that

$$\overline{Q}_\sigma = \lim_{V(\Lambda)\to\infty} \frac{1}{V(\Lambda)} \sum_\alpha \omega_\sigma(\Psi_F(f_\alpha)^* \Psi_F(V^1 f_\alpha)), \tag{7}$$

the limit being independent of the choice of complete orthonormal basis $\{f_\alpha\}$ for the region Λ, $\Lambda \in \Gamma$.

Conditions 1 and 2 essentially correspond to our characterization of ω, $\omega \in \mathcal{E}$, as an equilibrium state. Condition 3 corresponds to our formulation of the dynamics for the heuristically defined Hamiltonian of equation (3).

Condition 3 tells us that the time evolution for the representation π_{ω_σ} is a quasifree time evolution in the sense of Rocca-Sirugue-Testard.[10] A simple application of the techniques developed in the paper by Rocca-Sirugue-Testard[10] leads us to the following result:

$$\omega_\sigma(\Psi_F(f)^* \Psi_F(g)) = \langle f, G(H_{Q_\sigma}) g\rangle, \qquad \forall f, g \in \mathcal{K} \tag{8}$$

where $G(H_{Q_\sigma}) = [1 + \exp(\beta H_{Q_\sigma})]^{-1}$ (the fermi-factor).

Using Eq. (8) we may re-write Eq. (7) as

$$\overline{Q}_\sigma = \lim_{V(\Lambda)\to\infty} \frac{1}{V(\Lambda)} \sum_\alpha \langle f_\alpha, G(H_{Q_\sigma}) V^1 f_\alpha\rangle \tag{9}$$

with $\{f_\alpha\}$ a complete orthonormal basis for the region Λ. In Appendix A we express the RHS of Eq. (9) explicitly as a function of Q_σ, with the consequence that Eq. (9) may now be regarded as an equation for Q_σ.

Thus from Appendix A

$$\overline{Q}_\sigma \gamma = \overline{Q}_\sigma \int_0^{2K_F} \frac{\sinh\beta[\epsilon(K)_-^2 + g^2|Q_\sigma|^2]^{1/2}}{\{\cosh\beta\epsilon(K)_+ + \cosh\beta[\epsilon(K)_-^2 + g^2|Q_\sigma|^2]^{1/2}\}[\epsilon(K)_-^2 + g^2|Q_\sigma|^2]^{1/2}} dK, \tag{10}$$

where $\epsilon(K)_\pm = \frac{1}{2}\{\epsilon(K) \pm \epsilon(K - 2K_F)\}$, $\epsilon(K) = C|K| - \mu$, and $\gamma = -2K_F/g$, i.e., $\gamma \in \mathbb{R}^+$ when $g \in \mathbb{R}^-$.

3. THE PHASE TRANSITION

We now examine the solutions of Eq. (10). Explicitly, Eq. (10) reduces to

$$\overline{Q}_\sigma \gamma = \overline{Q}_\sigma \int_0^{2K_F} \frac{\sinh\beta(C^2(K - K_F)^2 + g^2|Q_\sigma|^2)^{1/2}}{\{\cosh\beta(CK_F - \mu) + \cosh\beta(C^2(K - K_F)^2 + g^2|Q_\sigma|^2)^{1/2}\}(C^2(K - K_F)^2 + g^2|Q_\sigma|^2)^{1/2}} dK,$$

If we now choose $\mu = CK_F$, then

$$\overline{Q}_\sigma \gamma/2 = \overline{Q}_\sigma \int_0^{K_F} \frac{\tanh\beta/2(C^2y^2 + g^2|Q_\sigma|^2)^{1/2}}{(C^2y^2 + g^2|Q_\sigma|^2)^{1/2}} dy. \tag{11}$$

Equation (11) has the trivial solution $Q_\sigma = \overline{Q}_\sigma = 0$, which is valid for all β. We now seek a nontrivial solution. To do this we define the function F as

$$F: \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$$

with

$$F(\beta, \eta) = \int_0^{K_F} \frac{\tanh\beta/2(C^2y^2 + g^2\eta)^{1/2}}{(C^2y^2 + g^2\eta)^{1/2}} dy, \qquad \beta, \eta \in \mathbb{R}^+.$$

There will exist a nontrivial solution to Eq. (11) if the function F attains the value $\gamma/2$ for suitably chosen β and η. To show that such a nontrivial solution exists, we list five properties of the function F that will make this apparent.

(1) $F(\beta, \eta)$ is separately continuous in β and η.

(2) $F(0, \eta) = 0, \quad \forall \eta \in \mathbb{R}^+$.

(3) $F(\infty, \eta) = \frac{1}{C} \ln \frac{(C + \sqrt{C^2 + g^2K_F^2\eta})}{CgK_F\eta^{1/2}}$, then $F(\infty, 0) = \infty$.

(4) $F(\beta, \eta)$ is a monotonically increasing function of β, $\forall \eta, \beta \in \mathbb{R}^+$.

(5) $F(\beta, \eta)$ is a monotonically decreasing function of η, $\forall \eta, \beta \in \mathbb{R}^+$.

Properties 2, 3, and 4 imply that there exists a β_c, $\beta_c \in \mathbb{R}^+$, $\beta_c \neq 0$ such that

$$F(\beta_c, 0) = \gamma/2$$

and that this equation uniquely defines β_c. Furthermore, 2 to 5 imply that if $\beta \leq \beta_c$, then $F(\beta, \eta) \leq \gamma/2$, $\forall \eta \in \mathbb{R}^+$ so that we only have the trivial solution for Eq. (11).

Now, if $\beta > \beta_c$ properties 2 to 5 imply that there exists a unique nontrivial solution to the following equation for η

$$F(\beta, \eta) = \gamma/2,$$

denote this solution by $\eta(\beta)$.

Thus, Eq. (11) has a unique nontrivial solution for $|Q_\sigma|$. Then using Properties 1 to 5 we see that the mapping $\beta \to \eta(\beta)$ of $\mathbb{R}^+ \to \mathbb{R}^+$ is continuous with the additional property.

$$\lim_{\beta\to\beta_c^+} \eta(\beta) = 0. \tag{12}$$

We also have that $\eta(\beta)$, $\beta \geq \beta_c$ is monotonically increasing function of β, with

$$\max_{\beta\geq\beta_c} \eta(\beta) = \eta(\infty),$$

where $\eta(\infty)$ is the unique solution of the following equation:

$$\gamma/2 = \frac{1}{C} \int_0^1 \left(y^2 + \frac{K_F^2 g^2}{C^2}\eta(\infty)\right)^{-1/2} dy.$$

Thus we have shown that Eq. (11) has a unique nontrivial solution for $|Q_\sigma|$ if $\beta > \beta_c$; furthermore, the solution is independent of σ. Therefore, we have

$$Q_\sigma = \hat{Q} \exp i\alpha(\sigma), \qquad \text{with } \alpha(\sigma) \in \mathbb{R}, \tag{13}$$

where $\hat{Q}$ is the solution of Eq. (11) for given β, $\beta > \beta_c$. Equation (13) implies that we may put the set $\{\omega_\sigma\}$ into a one-to-one correspondence with the unit circle S^1. The existence of a nontrivial solution to Eq. (10) demonstrates that our model exhibits a phase transition with critical temperature β_c^{-1}. It follows from the foregoing

that the transition is of the second order with $\hat{Q}$ regarded as an order parameter. That this interpretation is reasonable follows from the arguments leading to Eq. (12). In particular, $\hat{Q}$ tends continuously to zero as the temperature tends continuously to β_c^{-1} from below.

We now enquire as to the action of space translations on the states ω_σ. In the proof of Eq. (8) one shows that the ω_σ are in fact quasifree and so to determine the action of space translations we only need consider the two point functions.

In particular,

$$\begin{aligned}\alpha_y^*\omega_\sigma(\Psi_F(f)^*\Psi_F(g)) &\equiv \omega_\sigma(\alpha_y(\Psi_F(f)^*\Psi_F(g)))\\ &= \langle f, U_y G(H_{Q_\sigma})U_{-y}g\rangle = \langle f, G(H_{Q_{\sigma(y)}})g\rangle\\ &= \omega_{\sigma(y)}(\Psi_F(f)^*\Psi_F(g)), \qquad \forall f, g \in \mathcal{K},\end{aligned}$$

where we have used Eq. (6) and $Q_{\sigma(y)} = e^{-2iK_F y}Q_\sigma$, i.e., the set $\{\omega_\sigma\}$ is invariant under space translations but not pointwise invariant. It is easy to see from the above equations that the states ω_σ are invariant under a periodic symmetry with period π/K_F.

4. CONCLUSIONS

We concluded after Eq. (13) that our model undergoes a second order phase transition at the critical temperature β_c^{-1}. Furthermore, we have established the following facts about the set of states $\mathcal{E}$. Firstly, if $\beta \le \beta_c$, $\mathcal{E}$ contains just one element, namely the free state corresponding to $Q_\sigma = 0$. It may easily be seen that this state is invariant under space translations. Secondly, if $\beta > \beta_c$, $\mathcal{E}$ contains many states. In addition, we found that the set $\mathcal{E}$ was setwise but not pointwise invariant under space translations. Thus we have a spontaneous breakdown[3] of continuous space translational symmetry. We have shown that the extremal elements of the convex set $\mathcal{E}$ are however invariant under a periodic symmetry of period π/K_F. It is interesting to note that for $\beta > \beta_c$ the set $\mathcal{E}$ contains only two space translationally invariant states. These are the free state corresponding to the trivial solution of Eq. (10), and the following state $\omega = K_F/\pi \int_0^{\pi/K_F} \omega_{\sigma(y)}dy$

The situation we have described in the previous paragraph demonstrates that our model exhibits a second order phase transition associated with a spontaneous breakdown of continuous space translational symmetry in favour of a periodic symmetry. Thus we have established a counter example to Landau and Lifshitz's conjecture in the case of one dimension.

We point out that our methods are not restricted to the particular form of "free" Hamiltonian we use. In fact, we have established the existence of the phase transition and the symmetry breakdown for a wide class of "free" Hamiltonians; this class includes $\omega(K) = K^2/2$. However, it then becomes more difficult to determine the order of the transition, in particular to obtain Eq. (12).

ACKNOWLEDGMENTS

The authors are indebted to Dr. G. L. Sewell for his constant help and encouragement throughout the course of this work. One of us (R.P.M.) would also like to thank the Science Research Council and Queen Mary College, Drapers Company for Research Studentships. The other one of us (R.M.) wants to thank Professor D. Martin and Professor J. Valatin for their kind hospitality during the months from March 1971 to December 1972 at the Physics Department of Queen Mary College.

APPENDIX A

We evaluate

$$\lim_{V(\Lambda)\to\infty} \frac{1}{V(\Lambda)} \sum_\alpha \langle f_\alpha, G(H_\sigma)f_\alpha\rangle.$$

Firstly we shall consider

$$G(H_\sigma)V^1 f(x), \qquad \text{with } f \in \mathcal{K}.$$

Definitions: Let

$$\begin{aligned} g_x^1(K) &= e^{-iKx}, \quad K \in G_1, & g_x^1(K)' &= e^{-iK'x}, \quad K \in G_1,\\ &= 0, \quad K \notin G_1, & &= 0, \quad K \notin G_1,\\ g_x^2(K) &= e^{-iKx}, \quad K \in G_2, & g_x^2(K)'' &= e^{-iK''\kappa}, \quad K \in G_2,\\ &= 0, \quad K \notin G_2, & &= 0, \quad K \notin G_2,\end{aligned}$$

with $K' = K - 2K_F$ and $K'' = K + 2K_F$.

Then

$$\begin{aligned} G(H_\sigma)V^1 f(x) &= \int_{G_1} \overline{g_x^1(K)}\, \widehat{G(H_\sigma)}\hat{f}(K - 2K_F)dK\\ &= (g_x^1, \widehat{G(H_\sigma)}\hat{f}_{2K_F}).\end{aligned}$$

with $\hat{f}_{2K_F} \equiv \hat{f}(K - 2K_F)$ and $\widehat{G(H_\sigma)f}(K) \equiv \widehat{G(H_\sigma)}\hat{f}(K)$, $\forall f \in \mathcal{K}$.

Case I: $K \in G_1$

Definition 1.a:

$$\lambda_{K_\pm}^1 = \frac{\epsilon(K) + \epsilon(K')}{2} \pm [\tfrac{1}{4}(\epsilon(K) - \epsilon(K'))^2 + g^2|Q_\sigma|^2]^{1/2}$$

and

$$\begin{aligned}\lambda_K^1 &= \lambda_{K_-}^1 \quad \text{if} \quad K \in (0, K_F]\\ &= \lambda_{K_+}^1 \quad \text{if} \quad K \notin (0, K_F].\end{aligned}$$

Definition 1.b:

$$\Psi_K = g_x^1(K) + A^1(K)g_x^1(K)',$$

where

$$A^1(K) = \frac{\lambda_K^1 - \epsilon(K)}{g\bar{Q}_\sigma} = \frac{g\bar{Q}_\sigma}{\lambda_K^1 - \epsilon(K')}$$

Case II: $K \in G_2$.

Definitions 2.a:

$$\lambda_{K_\pm}^2 = \frac{\epsilon(K) + \epsilon(K'')}{2} \pm [\tfrac{1}{4}(\epsilon(K) - \epsilon(K''))^2 + g^2|Q_\sigma|^2]^{1/2}$$

and

$$\begin{aligned}\lambda_K^2 &= \lambda_{K_-}^2 \quad \text{if} \quad K \in (-K_F, 0]\\ &= \lambda_{K_+}^2 \quad \text{if} \quad K \notin (-K_F, 0].\end{aligned}$$

Definition 2.b:

$$\Psi_K = A^2(K)g_x^2(K) + g_x^2(K)'',$$

where

$$A^2(K) = \frac{\lambda_K^2 - \epsilon(K'')}{g\bar{Q}_\sigma} = \frac{gQ_\sigma}{\lambda_K^2 - \epsilon(K)}.$$

From the definition of the g functions we see that

$$\widehat{V^{Q_\sigma}g^1_x}(K) = g\overline{Q}_\sigma g^1_x(K)', \qquad \widehat{V^{Q_\sigma}g^1_x}(K)' = gQ_\sigma g^1_x(K),$$

$$\widehat{V^{Q_\sigma}g^2_x}(K) = g\overline{Q}_\sigma g^2_x(K)'', \qquad \widehat{V^{Q_\sigma}g^2_x}(K)'' = g\overline{Q}_\sigma g^2_x(K),$$

with

$$\widehat{V^{Q_\sigma}f}(K) = \widehat{V^{Q_\sigma}\hat{f}}(K), \qquad \forall f \in \mathcal{K}.$$

A simple calculation verifies that

$$\hat{H}_\sigma \Psi_K = \lambda^1_K \Psi_K, \qquad K \in G_1,$$

$$\hat{H}_\sigma \Psi_K = \lambda^2_K \Psi_K, \qquad K \in G_2.$$

It follows simply from the above definitions that

$$\Psi_{K'} = g^1_x(K) + A^2(K')g^1_x(K)', \qquad K \in G_1.$$

Therefore,

$$g^1_x(K) = (A^2(K')\Psi_K - A^1(K)\Psi_{K'})(A^2(K') - A^1(K))^{-1}.$$

It was our desire to invert the relation between the Ψ's and the g's that led to the particular choice of sign in the Definitions 1a and 2a.

Consider now $\widehat{G(H_\sigma)}g^1_x(K)$. We have

$$\begin{aligned}\widehat{G(H_\sigma)}g^1_x(K) &= \{A^2(K')G(\lambda^1_K)\Psi_K - A^1(K)G(\lambda^2_{K'})\Psi_{K'}\}[A^2(K') - A^1(K)]^{-1} \\ &= \{\{A^2(K')G(\lambda^1_K) - A^1(K)G(\lambda^2_{K'})\}g^1_x(K) + A^2(K')A^1(K) \\ &\quad \times \{G(\lambda^1_K) - G(\lambda^2_{K'})\}g^1_x(K)'\}[A^2(K') - A^1(K)]^{-1}.\end{aligned}$$

Therefore,

$$\widehat{(G(H_\sigma)g^1_x, \hat{f}_{2K_F})} = \int_{G_1} \xi(K)\{G(\lambda^1_K) - G(\lambda^2_{K'})\}\, e^{iK'x}\hat{f}(K')dK + \text{the other term},$$

where

$$\overline{\xi(K)} = A^2(K')A^1(K)[A^2(K') - A^1(K)]^{-1}.$$

Therefore,

$$\langle f_\alpha, G(H_\sigma)V^1 f_\alpha\rangle = \int_\Lambda \int \bar{f}_\alpha(x)f_\alpha(y)[\int_{G_1} \xi(K)\{G(\lambda^1_K) - G(\lambda^2_{K'})\}e^{iK'(x-y)}dK]_{dx\,dy} + \text{the other term}.$$

Now

$$\sum_\alpha \langle f_\alpha, G(H_\sigma)V^1 f_\alpha\rangle = \int_\Lambda \left(\int_{G_1} \xi(K)\{G(\lambda^1_K) - G(\lambda^2_{K'})\}\, dK\right)dx + \text{the other term}.$$

Then

$$\lim_{V(\Lambda)\to\infty} \frac{1}{V(\Lambda)} \sum_\alpha \langle f_\alpha, G(H_\sigma)V^1 f_\alpha\rangle = \int_{G_1} \xi(K)\{G(\lambda^1_K) - G(\lambda^2_{K'})\}dK$$

the other term going to zero in the limit.

We have

$$\int_{G_1} \xi(K)\{G(\lambda^1_K) - G(\lambda^2_{K'})\}dK = \frac{\pi g}{K_F}\int_0^{2K_F} \overline{Q}_\sigma \frac{\{G(\lambda^1_{K_-}) - G(\lambda^1_{K_+})\}}{\lambda^1_{K_-} - \lambda^1_{K_+}}\, dK.$$

$$= -\frac{\pi g}{2K_F}\int_0^{2K_F} \frac{\sinh\beta(\epsilon(K)^2_- + g^2|Q_\sigma|^2)^{1/2}\, dK}{[\cosh\beta\epsilon(K)_+ + \cosh\beta(\epsilon(K)^2_- + g^2|Q_\sigma|^2)^{1/2}](\epsilon(K)^2_- + g^2|Q_\sigma|^2)^{1/2}}$$

with $\epsilon(K)_\pm = \frac{1}{2}[\epsilon(K) \pm \epsilon(K - 2K_F)]$, $\epsilon(K) = C|K| - \mu$.

*Work performed during a visit by R. Moscheo to the Physics Department, Queen Mary College, London.

†Holder of a fellowship granted by the Comitato Regionale Richerche, Nucleari della Regione Siciliana.

‡Present address: Department of Mathematical Physics, The University of Birmingham, P. O. Box 363, Birmingham 15, England.

[1]L. D. Landau and E. E. Lifshitz, *Statistical Physics* (Pergamon, London, 1958), see in particular pp. 260 and 430.

[2]See Ref. 1 for the definition of first and second order transition.

[3]G. G. Emch, *Algebraic Methods in Statistical Mechanics and Quantum Field Theory* (Wiley, New York, 1972). D. Ruelle, *Statistical Mechanics* (Benjamin, New York, 1969). (All the technical terms used in this paper may be found in these books or in the references therein.)

[4]R. Haag, Nuovo Cimento 25, 287 (1962).

[5]H. Fröhlich, Proc. Roy. Soc. 223, 296 (1954).

[6]R. H. Brout, *Phase Transitions* (Benjamin, New York, 1965).

[7]R. Haag, N. M. Hugenholtz, and M. Winnink Commun. Math. Phys. 5, 215 (1967).

[8]R. P. Moya, "Equilibrium states for the infinite ideal Bose gas," J. Math. Phys. (to be published). In this paper we found anomalous K. M. S. states for the ideal Bose gas; a simple check shows that such states do not occur here.

[9]G. L. Sewell, "States and Dynamics of Infinitely Extended Physical Systems," (preprint). We refer to this paper for further justification of Condition (1).

[10]F. Rocca, M. Sirugue, and D. Testard, Commun. Math. Phys. 13, 317 (1969).

Conformal Killing vector fields on timelike two-surfaces*

J. C. du Plessis

University of the Witwatersrand, Johannesburg, South Africa
(Received 4 November 1973)

A topological description of Killing vector fields as well as conformal Killing vector fields on a 2-space–time is given, and the block diagram extension technique is generalized.

1. INTRODUCTION

Studies[1–5] of static timelike 2-surfaces have contributed much to the understanding of some physically important stationary space—times such as the Kerr—Newman family.[1,2] This paper aims at a unified qualitative theory of stationary and conformally stationary timelike 2-surfaces.

In Sec. 2 the necessary preliminaries, notably some ideas from the qualitative theory of vector fields as initiated by Poincaré,[6] are briefly described and a result is proved which enables one to obtain the index of a critical point[6,7] in a simple geometrical manner. In Sec. 3 a qualitative description of a general conformal Killing vector field ξ and its horizon[8] is given. It is shown, for example, that the isolated critical points of ξ, which are also bifurcate points[3] of the horizon, are of a particularly elementary type and can readily be classified, that ξ has no homotopically trivial cycles, and that horizon segments, which are also isolated if they contain isolated critical points, cannot join critical points of the same nonzero index. If ξ is a Killing vector field, each isolated critical point is a simple saddle, and each arcwise connected component of the Killing horizon can contain at most one bifurcate point, so that horizon polygons do not exist. On the other hand, conformal Killing horizon polygons may exist.

In Sec. 4 these results are used to show that a conformal Killing vector field is structurally stable on each disc in a 2-space—time M (in the sense of Andronov and Pontrjagin[9]) if and only if each critical point is simple, while a Killing vector field is structurally stable if and only if each critical point is isolated.

In Sec. 5 a generalization is given of the well-known "block diagram" extension technique.[1–5]

2. PRELIMINARIES

Let M be a connected orientable 2-manifold. Then the bundle $F(M)$ of frames over M has two connected components and any two frames at $p \in M$ in the same component of $F(M)$ are related by a matrix transformation with positive determinant. An orientation of M is a labelling of these components by "+" and "−", respectively; thus one can speak of *positive* and *negative* frames. Given a metric tensor g on M, the angle α mod 2π that a vector v makes with a vector u is uniquely determined, since g determines the angle up to sign, while the sign in turn is fixed by the sign of the frame (u, v).

An orientation of M also determines a concordant orientation of the boundaries of all 2-cells in M as follows. Let $\dot{e}$ be a simple closed differentiable curve (i.e., a Jordan curve) bounding a 2-cell e and consider a parametrization $S^1 \to \dot{e}$ with corresponding tangent field t and unit inward normal n (i.e., n points *into* e, where "into" is defined in terms of the exponential mapping). Then t defines an orientation of $\dot{e}$ and one says $\dot{e}$ is *positively oriented* if at each point on $\dot{e}$ the frame (t, n) is positive. In well-known manner this orientation of $\dot{e}$ now determines a unique positive orientation of the boundary of any other 2-cell f in M, even if $\dot{f}$ is not differentiable (in which case "oriented" means simply "directed").

Let $A \subset M$ be an open set and consider a cell-decomposition of the closure $\overline{A}$. The positive orientation of the boundary of each 2-cell in this decomposition induces a unique positive orientation of the topological boundary $\dot{\overline{A}}$.

Let v be a vector field on M. The zeros of v will be called *critical points*. A critical point p is *simple* if

$$\det\left(\frac{\partial v^i}{\partial x^j}\right) \neq 0 \quad \text{at } p. \tag{2.1}$$

Let J be a positively oriented Jordan curve bounding a 2-cell and containing no critical points. Let J be parametrized in the positive sense by $\theta \in [0, 2\pi]$ and let $\alpha(\theta)$ be the angle that v makes with the tangent to J at θ. Since $\alpha(\theta)$ is determined up to multiples of 2π, the requirement that α be continuous on $[0, 2\pi]$ uniquely determines the function $\Delta(\theta) = \alpha(\theta) - \alpha(0)$. Define the *index*[6,7] of J to be

$$i(J; v) \overset{\text{def}}{=} (2\pi)^{-1}\Delta(2\pi) + 1. \tag{2.2}$$

The index is an integer and is independent of the orientation of M. The index of an isolated critical point p is defined by

$$i(p; v) \overset{\text{def}}{=} i(C; v), \tag{2.3}$$

where C is a small Jordan curve about p enclosing no other critical point except p. The index of p is independent of the choice of C. A noncritical point is assigned index 0. The index of a simple critical point can only be 0 or ± 1.[10] A simple critical point with index -1 is a *saddle point*.[10] Bendixson[11] showed that for a critical point p

$$i(p; v) = 1 + \tfrac{1}{2}(e - h), \tag{2.4}$$

where e is the number of elliptic sectors and h the number of hyperbolic sectors[12] about p. Poincaré[6] proved that if J is a Jordan curve bounding a 2-cell which contains isolated critical points $p_1 \cdots p_n$, then

$$i(J; v) = \sum_{\alpha=1}^{n} i(p_\alpha; v). \tag{2.5}$$

We shall now derive a result which is useful for

evaluating the index of a critical point and which will be used repeatedly in the next section.

Let (u, v) be a dyad of vector fields on an oriented 2-manifold M and suppose each of u and v has only isolated critical points. Consider the open set

$$M_+ = \{p \in M : (u_p, v_p) \text{ is positive}\}. \tag{2.6}$$

The *degenerate boundary* B is defined to be the topological boundary of the closure of M_+, i.e., $B = \overline{M}_+$. Since (u, v) is degenerate on B, it can be partitioned

$$B = C \cup B_+ \cup B_-, \tag{2.7}$$

where C is the collective set of critical points of u and v,

$$B_\pm = \{p \in B - C : \hat{u}_p = \pm \hat{v}_p\}, \tag{2.8}$$

and $\hat{u}_p$, $\hat{v}_p$ are unit vectors parallel to u_p, v_p, respectively. It will be assumed that $\overline{M}_+$ admits a triangulation which induces locally finite triangulations of $\overline{B}_+$ and $\overline{B}_-$. Then each of B_+ and B_- is the union of a collection of arcs in M, only finitely many of which intersect some neighborhood of each $p \in M$. Also, positive orientation of B is well-defined so that the B_+ and B_- arcs may be regarded as directed arcs.

Definition: Let J be a Jordan curve free of critical points, transversal to B and bounding a 2-cell e. Let B have positive orientation. Then the *orientation degree* of J is defined by

$$\epsilon(J) = (\text{number of } B_+ \text{ arcs leaving } e) - (\text{number of } B_+ \text{ arcs entering } e). \tag{2.9}$$

For a point p, define $\epsilon(J) = \epsilon(C)$, where C is a small Jordan curve (as above) about p.

If (u, v) is degenerate only on B, then not only is B independent of the orientation of M, but (unlike, for example, positive orientation of a cell boundary) the *positive orientation of B is also independent of the orientation of M*. In general, of course, the degenerate boundaries yielded by the two orientations of M may not even be homeomorphic. However, one readily sees that the orientation degree of J is independent of the orientation of M.

If only B and the partitioning (2.7) is known, one can still determine $\epsilon(J)$ correctly up to sign by simply directing the B-arcs intersecting J alternately in and out of the 2-cell e as one proceeds around J. The sign indeterminacy then amounts to an indecision as to which of the two vector fields is u and which is v.

Lemma 2.1: Let J be as above. Then

$$\epsilon(J) = i(J;v) - i(J;u).$$

In particular, for a point $p \in M$,

$$\epsilon(p) = i(p;v) - i(p;u).$$

Proof: Let J intersect B_+ at points $p_1 \cdots p_n$ and let p_α be given by $\theta = \theta_\alpha \neq 0$, where J is parametrized by $\theta \in [0, 2\pi]$ in the positive sense. Then u and v are parallel at p_α. If $\delta(\theta)$ is the angle that $v(\theta)$ makes with $u(\theta)$ at θ, one says v crosses u in the positive (respectively negative) sense at p_α if $\delta(\theta)$ is monotonically increasing (respectively decreasing) on an open interval about $\theta = \theta_\alpha$. It readily follows from (2.2) that[13]

$$i(J;v) - i(J;u) = m - n,$$

where m and n denote the total number of times v crosses u in the positive and negative sense, respectively, at points in $J \cap B_+$. Now consider an orientation preserving homeomorphism h of a neighborhood of $\overline{e}$ into the plane and choose the orientation on M such that it agrees under h with the standard positive orientation of the plane. Then $M_+ \cap J$ is the set of points on J for which $\delta(\theta) \bmod 2\pi \in (0, \pi)$ and each B arc γ is so oriented that, in the planar picture, "M_+ lies to the left of γ" as one proceeds along γ. It follows v crosses u in the positive sense at p_α if the directed B_+ arc γ through p_α leaves e at p_α, while v crosses u in the negative sense if γ enters e at p_α. Hence $\epsilon(J) = m - n$. □

It follows that B_+ (and also B_-, by symmetry) is the union of arcs in M with endpoints in the critical set C. For if p is a noncritical point, then by the lemma, $\epsilon(p) = 0$, so that incoming and outgoing B_+ arcs may be paired off. If M is compact, each arc has its endpoints (possibly coincident) in C, whence

$$0 = \sum_{p \in C} \epsilon(p) = \sum_{p \in C} i(p;v) - \sum_{p \in C} i(p;u). \tag{2.10}$$

Thus the *total index* of a vector field on M is a constant. This constant is of course the Euler–Poincaré characteristic of M.

As another application, consider a vector field u on the plane of the form

$$u = G(x)\frac{\partial}{\partial x} + F(x, y)\frac{\partial}{\partial y}, \tag{2.11}$$

where G has isolated roots. Then the dyad $(u, \partial/\partial y)$ is degenerate at the zeros of G, so B will consist of directed lines $x = c_\alpha$, where each c_α is some root of G. Thus at most one of these lines passes through a given point p, with one segment directed towards p, the other away from p. Hence $\epsilon(p) = 0, \pm 1$ and by Lemma 2.1, the index of an isolated critical point can assume only the values 0 or ± 1.

3. CONFORMAL KILLING VECTOR FIELDS

Let (M, ds^2) be an oriented time-oriented 2-space–time (or timelike 2-surface). Then M is parallelizable since there exist future directed null vector fields l and n with sign $(l, n) = +1$ everywhere, and each of l and n is uniquely determined up to a positive factor. Let ξ be a conformal Killing vector field satisfying the conformal Killing equations

$$g_{ik}\xi^k{}_{,j} + g_{jk}\xi^k{}_{,i} + g_{ij,k}\xi^k = \phi g_{ij}, \tag{3.1}$$

for some scalar field ϕ. The *l-horizon* H_l is defined by

$$H_l = \{p \in M : g_{ij}\xi^i l^j = 0 \text{ at } p\}, \tag{3.2}$$

and H_n is defined similarly. These horizons are clearly independent of the choice of l and n. Since l and n are nowhere tangent, the critical set of ξ is precisely $H_l \cap H_n$. The set $H = H_l \cup H_n$, on which ξ is null, is called the *Killing horizon*.[8]

Proposition 3.1: Let ξ be a conformal Killing vector field on an oriented time-oriented 2-space—time, and suppose ξ has only isolated critical points.

(i) If a component λ of the l-horizon contains a critical point, then λ is a maximal integral curve of l and is isolated. The critical points with nonzero index on λ have indices alternately $+1$ and -1 along λ. Similarly for the n-horizon. A component of the l-horizon intersects a component of the n-horizon in at most one point.

(ii) None of l, n, and ξ have homotopically trivial cycles. In particular, no critical point is a center.[14]

(iii) A critical point p of ξ has no elliptical sectors while the number of hyperbolic sectors may be 0, 2, or 4.

(iv) The seperatrixes[15] of a saddle point of ξ are contained in the Killing horizon, and no seperatrix goes from a saddle to a saddle.

Proof: (i) Since l has no zeros there are local coordinates $\{u, x\}$ such that $l = \partial/\partial x$. Then

$$ds^2 = E du^2 + 2F\, du\, dx, \tag{3.3}$$

and one of the conformal Killing equations implies

$$\xi = f(u)\frac{\partial}{\partial u} + g(x,u)\frac{\partial}{\partial x}. \tag{3.4}$$

Hence the index of a critical point p of ξ may assume only the values 0, ± 1. Also

$$g_{ij}\xi^i l^j = Ff, \tag{3.5}$$

so H_l is locally given by $f(u) = 0$. Thus H_l is generated by maximal integral curves of l, since these are locally given by $u = \text{const}$. On the other hand, H_l is also generated by maximal integral curves of ξ. Let λ and η be the respective maximal integral curves of l and n through a critical point p, so that $\lambda \subset H_l$ and $\eta \subset H_n$. Since p is isolated there is a neighborhood V of p in which H_l and H_n intersect only at p. Since η is transversal to λ at p and since l has no critical points, the integral curves of l intersecting the curve $\eta \cap V$ will trace out a neighborhood W of λ. Then $W \cap H_l = \lambda$, for if $q \in H_l \cap W$ but $q \notin \lambda$, then the maximal integral curve λ_q of l through q is contained in H_l, which is not possible since λ_q intersects η in V at a point other than p. Moreover, λ and η cannot cross at more than one point, since the sign of (l, n) at adjacent crosspoints will have to be opposite, contradicting sign $(l, n) = +1$ everywhere.

Now consider the dyad (l, ξ) with positively oriented degenerate boundary $B \subset H_l$. If λ contains critical points p_α with nonzero index, it follows from the above and Lemma 2.1 that λ is an isolated connected component of B, and the p_α divide λ into an alternating sequence of B_+ and B_- arcs concordantly oriented along λ. Consequently, the orientation degrees $\epsilon(p_\alpha)$ of the p_α are alternately $+1$ and -1 along λ. By Lemma 2.1, $\epsilon(p_\alpha) = i(p_\alpha; \xi)$ since $i(p_\alpha; l) = 0$, so that the indices of the p_α alternate likewise.

(ii) If J is a homotopically trivial cycle of l, then $i(J, l) = 1$ and by Eq. (2.5), l must have a critical point-contradiction. Let J be a homotopically trivial cycle of ξ. Then $i(J; \xi) = 1$ and $i(J; l) = 0$ so that by Lemma 2.1, the orientation degree of J with respect to the diad (l, ξ) is $\epsilon(J) = 1$. Hence J intersects $B \subset H_l$ so that $J \subset H_l$, since H_l is generated by maximal integral curves of ξ. Thus l is tangent to ξ along J, so J is also a cycle of l-contradiction.

(iii) Suppose p has an elliptic sector. Then there is an integral curve γ of ξ initiating and terminating at p and such that $\bar{\gamma} = \gamma \cup \{p\}$ bounds a disc e. Since l has no zeros any integral curve of l entering e must leave it again, so there is an integral curve λ not through p intersecting γ at least twice. Then sign (l, ξ) is opposite at adjacent crosspoints which implies (l, ξ) is degenerate at some point $q \neq p$ on γ, i.e., $q \in H_l$ so that $\bar{\gamma} \subset H_l$. By (i) the loop $\bar{\gamma}$ is a cycle of l-contradiction. The number of hyperbolic sectors follows from the formula (2.4) of Bendixson.

(iv) The seperatrixes of a saddle point p are precisely the four integral curves of ξ which extend to p (excluding the constant curve at p). From (i) it follows that p is the intersection of isolated components λ and η of H_l and H_n, respectively, both being nontrivial curves generated by maximal integral curves of ξ. Thus $\lambda \cup \eta$ must contain the four seperatrixes. If two saddles are joined by a seperatrix, there would be adjacent critical points on a component of H_l or H_n, both with index -1, which contradicts (i). □

It follows from (i) that a null conformal Killing vector field ξ cannot have isolated critical points.

We now consider a Killing vector ξ, satisfying Eqs. (3.1) with $\phi = 0$.

Theorem 3.2: Let ξ be a Killing vector field on a 2-space—time. Each isolated critical point of ξ is a simple saddle.

Proof: The Killing equations imply

$$(Ff)_u = -(Fg)_x, \tag{3.6}$$

where the subscripts indicate partial derivatives with respect to u and x, respectively. From Eq. (3.4),

$$\det\left|\frac{\partial \xi^i}{\partial x^j}\right| = f_u g_x. \tag{3.7}$$

Let p with coordinates $\{0, 0\}$ be an isolated critical point of ξ. From Eqs. (3.4) and (3.6),

$$f_u = -g_x, \quad \det\left|\frac{\partial \xi^i}{\partial x^j}\right| = -f_u^2 \quad \text{at } p, \tag{3.8}$$

so that p is simple if and only if $f_u \neq 0$ at p. Suppose p is not simple, i.e., $f(0) = f_u(0) = 0$, so that by Eq. (3.6), $F(x, 0)g(x, 0) = \text{const}$, and since F has no zeros and $g(0, 0) = 0$, it follows $g(x, 0) = 0$. Thus $\xi = 0$ on the line $u = 0$, so p is not an isolated critical point. Hence p must be simple.

For the dyad (l, ξ) the set M_+ is locally given by $f < 0$, so that the degenerate boundary B is locally given by $u = 0$. Since p is simple, $f_u = -g_x \neq 0$ at p so that $f(u)$ and $g(x, 0)$ change sign across $u = 0$ and $x = 0$, respectively, and the sign changes are in the opposite sense. Hence if $f(u) < 0$ for small $u > 0$, then $g(x, 0) > 0$ for small $x > 0$, so that by Eq. (3.4), the positive x axis near p is a B_+ segment terminating at p, while if $f(u) < 0$ for small $u < 0$, then the negative x axis near p is a B_+ segment

terminating at p. In both cases $\epsilon(p)=-1$, and by Lemma 2.1, $i(p;\xi)=-1$. But p is simple, so it is a saddle. □

An illustration is provided by the Kruskal[4] extension of Schwarzschild 2-space—time in which there is a single bifurcate point[3] of the Killing horizon, which is a simple saddle of ξ.

Since critical points with index 0 or +1 do not exist, it follows from Proposition 3.1. (i) that a component of the Killing horizon can contain at most one bifurcate point, and that horizon polygons do not exist. For example, in block diagram extensions (cf. last section), a typical "block" is a region of space—time bounded by four horizon segments. It follows that at most two vertices of a block can be regular points and that there is always a pair of opposite vertices which are either singularities or infinities.

On the other hand, a conformal Killing vector field may give rise to horizon polygons.

4. STRUCTURAL STABILITY

Let $\mathcal{J}^1(M)$ be the space of C^1 vector fields on a 2-manifold M with the fine topology so that basic neighborhoods are defined as follows: Let $\{V_\alpha\}$ be a locally finite covering of M by coordinate neighborhoods. For given $\epsilon>0$ and compact set $\Omega\subset M$ the neighborhood $N(\xi;\Omega,\epsilon)$ of $\xi\in\mathcal{J}^1(M)$ consists of all $\xi^*\in\mathcal{J}^1(M)$ such that on each set $V_\alpha\cap\Omega$ the components of ξ^* and their first partial derivatives differ by less than ϵ from those of ξ, while $\xi=\xi^*$ outside Ω.

The notion of structural stability of vector fields on a disc was introduced by Andronov and Pontrjagin.[9] We shall call a vector field on M structurally stable if it is structurally stable in the sense of Andronov and Pontrjagin on each disc $D\subset M$. More precisely:

Definition[16]: A vector field $\xi\in\mathcal{J}^1(M)$ is structurally stable if for each closed disc $D\subset M$ and each $\epsilon>0$ one can find $\delta>0$ such that for each $\xi^*\in N(\xi;D,\delta)$ there is an ϵ-homeomorphism of M onto itself which maps integral curves of ξ onto integral curves of ξ^*.

The following characterization theorem was first stated by Andronov and Pontrjagin.[9] A proof was given by De Baggis.[17] The theorem is stated here in slightly different form because of the slight adaptation of the definition above.

Theorem 4.1: A vector field $\xi\in\mathcal{J}^1(M)$ is structurally stable if and only if

(i) each critical point is simple,[18]

(ii) each homotopically trivial cycle γ has stability index

$$\int_\gamma(\mathrm{div}\xi)\,ds\neq 0,$$

(iii) no seperatrix goes from a saddle to a saddle.

Comparison with Proposition 3.1 and Theorem 3.2 leads directly to the following result.

Theorem 4.2: A Killing vector field on a 2-space—time is structurally stable if and only if each critical point is isolated. A conformal Killing vector field is structurally stable if and only if each critical point is simple.

5. ANALYTIC EXTENSIONS

On a 2-space—time with Killing vector field ξ one can always find local coordinates x,y such that $\xi=\partial/\partial y$ and $\det(g_{ij})=-1$. The metric then takes the form

$$ds^2=G^{-1}(F^2-1)\,dx^2+2F\,dx\,dy+G\,dy^2, \tag{4.1}$$

where F and G are functions of x only. It will be assumed that F and G are analytic on an interval I. The norm square of ξ is G and the Gaussian curvature is $K=\frac{1}{2}G''$. This suggests that the singular appearance of the horizons, given by $G=0$, is not real. Suppose G has roots $x=a_i$ in I. The lines $x=a_i$ then divide $I\times R$ into strips $W_i=(a_i,a_{i+1})\times R$, on each of which ds^2 is regular analytic. The problem now is to find a way of "joining" W_{i-1} to W_i along a nonsingular "seam." The static case, given by $F=0$ (i.e., when $y\to -y$ is an isometry), has been discussed in detail by Walker and Godfrey,[5] following work by several authors[1-4] on the Schwarzschild, Kerr, and Reissner—Nordström metrics.

It will be assumed that $G'(a_i)\neq 0$ for each i, so that for given i one can write

$$G(x)=(x-a_i)G_0(x),\quad \alpha=G_0(a_i)\neq 0. \tag{4.2}$$

Since $1/G_0(x)$ is analytic on an interval I_i about $x=a_i$, one can write

$$\begin{aligned}&1/G_0(x)=\alpha^{-1}+(x-a_i)H'(x),\\ &F(x)/G_0(x)=\beta/\alpha+(x-a_i)E'(x),\quad \beta=F(a_i),\end{aligned} \tag{4.3}$$

for $x\in I_i$, where H and E may be taken as bounded analytic functions on I_i. We now introduce generalized Kruskal[4] coordinates u,v, given by the double valued transformation

$$uv=(x-a_i)\exp[\alpha H(x)],\quad u/v=\pm\,|x-a_i|^{\beta}\exp\{\alpha[y+E(x)]\}, \tag{4.4}$$

where the sign is that of $(x-a_i)$. In terms of differentials

$$\begin{aligned}&u^{-1}\,du=\tfrac{1}{2}\alpha[dy+G^{-1}(F+1)\,dx],\\ &v^{-1}\,dv=-\tfrac{1}{2}\alpha[dy+G^{-1}(F-1)\,dx],\end{aligned} \tag{4.5}$$

so that the metric takes the form

$$ds^2=\chi\,du\,dv,\quad \chi(x)=-4\alpha^{-2}G_0(x)\exp[-\alpha H(x)]. \tag{4.6}$$

The transformation (4.4) is analytic everywhere on $U_i=I_i\times R$ except where $x=a_i$, and has the effect of mapping a strip of W_i adjacent to $x=a_i$ into the first and also the third quadrant of the u,v plane, while a strip of W_{i-1} adjacent to $x=a_i$ is mapped into the second and fourth quadrants. Also $uv=0$ at $x=a_i$ and uv is an analytic function of x on I_i with $d(uv)/dx\neq 0$ at $x=a_i$. Thus x in turn is an analytic function of uv on a neighborhood of $uv=0$, so that the conformal factor χ in (4.6) is nonzero and analytic in uv on a neighborhood of $uv=0$. Finally,

$$\xi=\tfrac{1}{2}\alpha\left(u\frac{\partial}{\partial u}-v\frac{\partial}{\partial v}\right), \tag{4.7}$$

so that ξ has a simple saddle at the origin and the Kill-

ing horizon, given by the coordinate axes $u=0$ and $v=0$, is nonsingular.

The space—time may now be extended by patching together the given coordinate neighborhoods W_i and the Kruskal neighborhoods U_i, using (4.4). The double valuedness of the transformation (4.4) entails that the universal covering space of a space—time which has been maximally extended in this manner will contain either zero, one, or countably infinite horizon bifurcate points, depending on whether G has zero, one, or more than one root. Maximal extensions which are not simply connected may contain any finite or countably infinite number of bifurcate points. For examples, see Refs. 1—5.

In some cases, one may also use the above procedure to extend nonstationary space—time on which there is a conformal Killing vector field ξ. One then starts with a metric conformally related to a stationary metric of the form (4.1).

*This work has been sponsored in part by a grant from the South African Council for Scientific and Industrial Research.

[1]B. Carter, Phys. Rev. **141**, 1242 (1966).
[2]R. Penrose, in *Batelle Recontres*, edited by C.M. de Witt and J.A. Wheeler (Benjamin, New York, 1968).
[3]R.H. Boyer and R.W. Lindquist, J. Math, Phys. 8, 265 (1967).
[4]M.D. Kruskal, Phys. Rev. **119**, 1743 (1943).
[5]M. Walker, J. Math, Phys. **11**, 2280 (1970); B.B. Godfrey, J. Math. Phys. **12**, 606 (1971).
[6]H. Poincaré, J. Math. Pures Appl. (3)**7**, 375 (1881); (3)8, 251 (1882); (4)1, 167 (1885); (4)2, 151 (1886); *Oeuvres complètes, Vol. I.* (Gauthier-Villars, Paris), pp. 3—84, 90—161, 167—221.
[7]S. Lefschetz, *Differential Equations: Geometric Theory* (Interscience, New York), 2nd ed., p. 214.
[8]B. Carter, J. Math. Phys. **10**, 70 (1969).
[9]A.A. Andronov and L.S. Pontrjagin, Doklady Akad. Nauk **14**, 247 (1937).
[10]See Ref. 7, p. 181.
[11]I. Bendixon, Acta Mathematica **24**, 1 (1901). See also Ref. 7, p. 222.
[12]See Ref. 7, p. 219.
[13]See Ref. 6, p. 29 or Ref. 7, p. 195.
[14]See Ref. 7, p. 187.
[15]See Ref. 7, p. 223.
[16]Because of the definition of neighborhoods, it is not necessary to demand that discs D be chosen such that their boundaries are transversal to ξ.
[17]H. F. De Baggis, in *Contributions to the Theory of Non-linear Oscillations*, edited by S. Lefschetz (Princeton U. P., Princeton, N.J., 1952), Vol. II, p. 37.
[18]A simple critical point must be isolated, so there can be only finitely many of these on any closed disc.

Renormalized number operators

James D. Fabrey

Department of Mathematics, University of North Carolina, Chapel Hill, North Carolina 27514
(Received 19 October 1973)

A number operator for a Weyl system is called renormalized essentially if it is obtained from the total number operator by subtraction of a (possibly infinite) constant (in exponentiated form). Necessary and sufficient conditions for the existence of a renormalized number operator are obtained.

1. INTRODUCTION

Given a representation of the canonical commutation relations (CCR), a number operator (in the sense of Chaiken[1]) is a self-adjoint operator N which satisfies the commutation relations

$$Na^*(f)=a^*(f)(N+I) \tag{1}$$

in bounded (exponentiated) form. Here $a^*(f)$ is the creation operator for the wavefunction f, and (1) is to hold for all f. If such an operator N exists, then the representation of the CCR, or *Weyl* system, is called a *particle* representation or particle *Weyl* system.

It has been shown[1] that a particle representation always has a *normalized* number operator N (one whose spectrum is a subset of the integers). If Ψ_n is an eigenvector for N with eigenvalue n, then (1) implies that

$$Na^*(f)\Psi_n=(n+1)a^*(f)\Psi_n. \tag{2}$$

Thus, even if n is negative, one has the interpretation that $a^*(f)\Psi_n$ has one more particle than Ψ_n does.

An example of a number operator is the total number operator in the Fock representation:

$$N=\sum_{j=1}^{\infty}N(f_j)=\lim_{k\to\infty}N(M_k), \tag{3}$$

where $\{f_j\}$ is an orthonormal basis of test functions, $N(f_j)=a^*(f_j)a(f_j)$ measures the number of particles with wavefunction f_j, and $N(M_k)=\sum_{j=1}^{k}N(f_j)$ measures the number of particles with wavefunctions in the finite-dimensional space M_k spanned by $\{f_1,\ldots,f_k\}$. Moreover, (3) is independent of basis[2] and this characterizes the Fock representation.[2,3]

Another example is provided by (irreducible) *direct-product* representations.[1] They either have no number operator or they admit a number operator of the form

$$N=\sum_{j=1}^{\infty}(N(f_j)-m_j)=\lim_{k\to\infty}(N(M_k)-n_k) \tag{4}$$

expressed in bounded form, where $n_k=\sum_{j=1}^{k}m_j$. We may view N as the difference between the total number operator $\lim_{k\to\infty}N(M_k)$ and the constant $\lim_{k\to\infty}n_k$. Unless the representation is Fock, neither of these limits exists and one *renormalizes* the total number operator by subtracting an infinite constant. Nevertheless, the particle interpretation (1) and (2) remains valid. Roughly speaking, one subtracts the infinite constant $\lim_{k\to\infty}n_k$ because there is a cyclic vector which has exactly m_j particles with wavefunction f_j (for all j) and, therefore, an infinite number of particles.

Another example of a representation with a cyclic vector having an infinite number of particles is provided by the *exponential* representations[4-8] that arise in connection with the $(\phi^4)_3$ model. They are *locally Fock* in the sense that the number operator $N(B_k)$, which measures the number of particles with wavefunctions supported within the ball B_k of radius k, exists and is the total number operator (3) for the representation restricted to test functions supported within B_k. The following question naturally arises. If one removes the cutoff $(k\to\infty)$, can the total number operator be renormalized by subtracting an infinite constant:

$$N=\lim_{k\to\infty}[N(B_k)-n_k], \tag{5}$$

where $\lim_{k\to\infty}n_k=\infty$? This question is not completely resolved. However, if the answer is affirmative, there may be a shift in domain; this would mean that there is a dense set of analytic vectors for all $N(B_k)$, and none of these vectors is in the domain of N.

We say that a number operator N is renormalized if

$$e^{itN}=s\text{-}\lim_{k\to\infty}\exp[it(N_k-n_k)]\quad(t\text{ real}) \tag{6}$$

holds for some sequence of integers $\{n_k\}$, where (a) $N_k=N(M^{(k)})$ for some increasing sequence of finite-dimensional subspaces $M^{(k)}$ or (b) $N_k=N(B^{(k)})$ for some increasing sequence of balls $B^{(k)}$ if the Weyl system is locally Fock. It turns out that for locally Fock representations, (a) holds whenever (b) does.

Remark 1: The existence of (6) does not, in general, suffice to define a number operator, i.e., an operator for which (1) also holds. Indeed, Chaiken[3] has constructed (discontinuous) irreducible representations in which the right side of (6) exists in the case (a) with $M^{(k)}=M_k=\{f_1,\ldots,f_k\}$ and $n_k=0$, but the limit is dependent on the choice of basis; a number operator does not exist for any such choice. In this connection it should be noted that in order to *construct* a renormalized number operator N, both (6) and (1) must hold. Of course, one need only verify (6) to show that a *given* number operator N is renormalized.

Remark 2: For continuous Weyl systems on separable test function spaces, the existence of (6) implies (1) so we do permit any increasing sequence of subspaces $M^{(k)}$ rather than just those of the form $M^{(k)}=M_k=\{f_1,\ldots,f_k\}$ for some choice of basis. Of course, we can always choose a basis such that $M^{(k)}=M_{k'}$, where $\{k'\}$ is a subsequence of $\{k\}$. We do not require that the limit in (6) be defined for more than one choice of $M^{(k)}$ or $B^{(k)}$. However, if two renormalized number operators are defined in this way for an *irreducible* representation, it turns out that they differ by an integer so that they gen-

erate the same automorphism of the Weyl algebra

$$W(f) \to W[\exp(it)f] = \exp(itN)W(f)\exp(-itN). \quad (7)$$

Note: The equality in (7) is the exponentiated version of (1) which serves as our definition of number operator in the sequel.[1,3]]

We regard two number operators which differ by a constant as *essentially* equal. They have the same eigenvectors with constant difference in eigenvalues. Singling out one such number operator corresponds to selecting a state with "zero" particles. Thus, the limit in (6) is essentially independent of choice of $M^{(k)}$ or $B^{(k)}$ when the limit exists. This assertion is no longer valid for reducible Weyl systems.

However, bounded functions of renormalized number operators are clearly observables, which commute with all invariant subspaces for the Weyl system, and the corresponding restrictions induce number operators. Thus, an observable number operator could be a direct sum of renormalized number operators on irreducible subspaces. Two such operators need only be essentially equal on each subspace. (Chaiken[1] gives examples of nonobservable number operators for an infinite free nonrelativistic Bose gas.)

Theorems 1 and 2 provide necessary and sufficient conditions for the existence of a renormalized number operator. Theorem 1 provides a short proof of partial results of Chaiken[1], who uses the special structure of product representations to obtain more concrete necessary and sufficient conditions that *any* number operator exist; then all number operators turn out to be renormalized over subspaces of the form $M^{(k)} = M_k$. It would be of interest to apply these theorems to representations which lack this product structure [such as exponential representations associated with the $(\phi^4)_3$ model].

2. MAIN RESULTS

See Ref. 5 for notation and definitions concerning Weyl systems $W(f)$, $f \in J$, on a Hilbert space H. We shall assume that the space J of test functions is separable. We proceed with some introductory lemmas.

Lemma 1: If $W(f)$, $f \in J$, is a locally Fock Weyl system, then condition (b) implies condition (a).

Proof. Suppose (b) holds. Since $W(f)$, f supported within $B^{(k)}$, is Fock, the main result of Chaiken[3] asserts that for any orthonormal basis $\{f_j^{(k)}\}$ with support within $B^{(k)}$

$$e^{itN(B^{(k)})} = s\text{-}\lim_{j\to\infty} e^{itN(M_j^{(k)})}$$

where $M_j^{(k)} = \{f_1^{(k)}, \ldots, f_j^{(k)}\}$. Thus, given a vector $\Psi \in H$ and $\epsilon > 0$, there exists a j_k such that

$$\|(e^{itN(B^{(k)})} - e^{itN(M_{j_k}^{(k)})})\Psi\| \leq \epsilon/2.$$

We may choose the bases so that $M_{j_k}^{(k)}$ is increasing as $k \to \infty$. By (b), choose k sufficiently large so that

$$\|\{\exp(itN) - \exp[itN(B^{(k)})]\}\Psi\| \leq \epsilon/2.$$

Then

$$\|[\exp(itN) - \exp(itN_k)]\Psi\| \leq \epsilon$$

for $N_k = N(M_{j_k}^{(k)})$ and k sufficiently large. The lemma is proved.

In the terminology of Ref. 1 a number operator is *normalized* if its spectrum is either $\{0, \pm 1, \pm 2, \cdots\}$ or $\{0, 1, 2, \cdots\}$.

Lemma 2: A renormalized number operator is essentially equal to a normalized number operator.

Proof: By (6), with $t = 2\pi$,

$$\exp(2\pi iN) = s\text{-}\lim_{k\to\infty} \exp[2\pi i(N_k - n_k)] = I,$$

since the n_k are integers and the spectrum of N_k is the set of nonnegative integers. Therefore, the spectrum of N is a subset of the integers. By the reasoning of Ref. 1, $N - n$ is normalized for some integer n.

QED

We require that the n_k in (6) be integers solely for convenience (see Ref. 1).

Lemma 3: If (6) and (7) hold with the n_k not necessarily integers, then N is essentially equal to a renormalized number operator.

Proof: Let $t = 2\pi$ in (6) to obtain

$$\exp(2\pi iN) = s\text{-}\lim_{k\to\infty} \exp[2\pi i(N_k - n_k)] = s\text{-}\lim_{k\to\infty} \exp(-2\pi i n_k) I. \quad (8)$$

Since $\exp(2\pi iN)$ is unitary, the right side of (8) is a complex number $\exp(-ic)$ of modulus one. It may be that $n_k \to \infty$, but (8) implies that $n_k \to c$ (modulo 2π), i.e., there exists a sequence of integers $\{m_k\}$ such that

$$n_k - c - m_k \to 0 \quad (k \to \infty).$$

Thus,

$$\exp[it(N - c)] = s\text{-}\lim_{k\to\infty} \exp[it(N_k - m_k)],$$

so that $N - c$ is a renormalized number operator.

QED

The following lemma shows that when (6) and (7) hold, the resulting number operator N is essentially independent of the choice of N_k if the representation is irreducible. Moreover, if an irreducible representation has a renormalized number operator, then every number operator is essentially renormalized.

Lemma 4: An irreducible Weyl system $W(f)$, $f \in J$, has essentially at most one number operator.

Proof: If N and N' are two number operators for W, then two applications of (7) yield

$$[\exp(itN)\exp(-itN'),\ W(f)] = 0 \quad (f \in J).$$

Since W is irreducible,

$$\exp(itN)\exp(-itN') = a(t)I, \quad (9)$$

and the left side of (9) is unitary so that $|a(t)| = 1$. Thus,

$$\exp(itN) = a(t)\exp(itN'). \quad (10)$$

Since both sides of (10) are continuous, one-parameter unitary groups, it follows that

$$a(t_1+t_2)=a(t_1)a(t_2)$$

and $a(t)$ is continuous. Therefore, $a(t)=\exp(-ict)$ for some real constant c, and $N=N'-c$. This completes the proof.

We remark that if the right side of

$$U(t)=s\text{-}\lim_{k\to\infty}\exp[it(N_k-n_k)] \tag{11}$$

exists, then $U(t)$ is a one-parameter unitary group. However, the existence of a number operator further requires that $U(t)$ be strongly continuous at $t=0$ (so that it have a self-adjoint generator N) and that (7) be satisfied (so that N is indeed a number operator). We now show that for *continuous* Weyl systems the latter condition is automatic. [Recall that a Weyl system $W(f)$, $f\in J$, is *continuous* if the mapping

$$f\to(\phi,W(f)\Psi)\quad(\phi,\Psi\in H)$$

is continuous—on all of J, not just on rays $\{tf\}$.]

Lemma 5: If $U(t)$ exists, then it is a one-parameter unitary group. If $W(f)$, $f\in J$, is a continuous Weyl system on a Hilbert space H and $U(t)$ is strongly (or weakly) continuous at $t=0$ with self-adjoint generator N, then W is a renormalized particle representation with renormalized number operator N.

Proof: Suppose $U(t)$ exists. Since $\exp[it(N_k-n_k)]$ is a one-parameter unitary group for each k and the product is continuous with respect to the strong topology restricted to the unit ball in $B(H)$, it follows that $U(t)$ is also a one-parameter unitary group.

Let $J_0\subset J$ be either

Case (a) the algebraic span of $\{M^{(k)}\}_{k=1}^{\infty}$; or

Case (b) all functions in J of compact support.

Then in either case, for $f\in J_0$,

$$\exp(itN)W(f)\exp(-itN)=s\text{-}\lim_{k\to\infty}\exp[it(N_k-n_k)]W(f)\times\exp[-it(N_k-n_k)]. \tag{12}$$

By Proposition 3.1 of Ref. 3 the right side of (12) is independent of k for k sufficiently large and it equals $W[\exp(it)f]$. Thus, N is a number operator for the representation $W(f)$, $f\in J$. The lemma follows from Lemma 4.3 of Ref. 1 and the density of J_0 in J. The lemma is proved.

For *cyclic* continuous representations, it suffices to check convergence and continuity of $U(t)$ {equicontinuity of $\exp[it(N_k-n_k)]$ at a cyclic vector}. [Compare this with a result for Fock representations (Remark 2, p. 79, of Ref. 3).]

Theorem 1: Suppose Ω is a cyclic vector for a continuous Weyl system $W(f)$, $f\in J$. Then W has a renormalized number operator if and only if

$$U(t)\Omega=\lim_{k\to\infty}\exp[it(N_k-n_k)]\Omega \tag{13}$$

converges for some choice of N_k and $t\to U(t)\Omega$ is continuous {equivalently, $\exp[it(N_k-n_k)]\Omega$ is equicontinuous} at $t\to 0$.

Proof: Necessity is trivial. Let us now suppose that $U(t)\Omega$ exists. Then for $f\in J$ the limit

$$\begin{aligned}U(t)W(f)\Omega&=\lim_{k\to\infty}\exp[it(N_k-n_k)]W(f)\Omega\\&=\lim_{k\to\infty}W[\exp(it)f]\exp[it(N_k-n_k)]\Omega\\&=W[\exp(it)f]U(t)\Omega\end{aligned} \tag{14}$$

exists. [Property (7) for number operators has been used to "push" $\exp[it(N_k-n_k)]$ past the Weyl operators in (14).] By the Weyl relations and linearity, $U(t)A\Omega$ exists for all polynomials A in the operators $W(f)$. Since W is cyclic, the set of such $A\Omega$ is dense in H. Therefore, $U(t)$ is everywhere defined.

If $t\to U(t)\Omega$ is continuous at $t=0$, then by strong continuity of $W[\exp(it)f]$, which is the composite of the continuous maps $t\to\exp(it)f$ and $g\to W(g)$, and by (14), $U(t)W(f)\Omega$ and $U(t)A\Omega$ are continuous at $t=0$ for all polynomials A in the operators $W(f)$. Since $\|U(t)\|=1$ for all t and the set of such $A\Omega$ is dense in H, it is easy to verify that $U(t)$ is strongly continuous at $t=0$.

The theorem then follows from Lemma 5.

Corollary: Continuous discrete direct-product representations[1] are renormalized particle representations.

Proof: Let $\phi=\otimes_j\phi_{m_j}$ be a cyclic vector, where ϕ_j is the jth Hermite function and m_j are bounded. Let $M^{(k)}=\{e_i\}_{i=1}^{k}$ and $n_k=\sum_j m_j$. Then

$$U(t)=\lim_{k\to\infty}\exp[it(N_k-n_k)]\phi=\otimes_j\exp[it(N_s-m_j)]\phi_{m_j}=I$$

exists (since ϕ_{m_j} is an eigenvector for the number operator N_s with eigenvalue m_j in the Schrödinger representation) and is trivially continuous at $t=0$. QED

Of course, the results of Chaiken utilize product structure and are therefore much deeper. Theorem 3.3 of Ref. 1 establishes that the only direct product representations with number operators are the discrete representations (no continuity required) and they are renormalized particle representations. The key step is the existence of the strong limit of $\exp(itN_k)$ (which nevertheless must be renormalized because it is not strongly continuous at $t=0$).

Weyl systems may well be cyclic, yet limits of expectation values may be more manageable than strong limits (for example, exponential representations[6]). Unfortunately, the weak limit of $\exp[it(N_k-n_k)]$ defines an operator $U(t)$ which need not even be unitary. However, we can formulate the following theorem that includes two conditions which are together sufficient for the existence of a number operator.

Theorem 2: Let $W(f)$, $f\in J$, be a continuous Weyl system on a Hilbert space H. Then W is a renormalized particle representation if and only if the limit $U(t)$ in (11) exists and is weakly (or strongly) continuous {equivalently, $\exp[it(N_k-n_k)]$ is equicontinuous} at $t=0$ on a dense set $D\subset H$ for same choice of N_k and n_k. In this case the self-adjoint generator N of $U(t)$ is a number operator for W.

If $U(t)$ exists, then a sufficient condition that it be weakly continuous at $t=0$ on D is that $D\subset\cap_k D(N_k)$ for each $\Psi\in D$ and

$$(\Psi,|N_k-n_k|\Psi)\le K=K(\Psi) \tag{15}$$

is uniformly bounded ($k=1,2,\cdots$).

The W^*-algebra generated by the operators $\exp(itN_k)$ is unitarily equivalent to a multipication algebra $\mathcal{M}$ in some measure space $L_2(X,\mu)$ and $\exp(itN_k)$ is represented by $\exp[itn_k(x)]$ for some real, measurable functions $n_k(x)$, $k=1,2,\cdots$. For any choice of N_k and n_k, there exists a subsequence $\{k'\}$ of $\{k\}$ such that the weak limit

$$U(t)=w\text{-}\lim_{k\to\infty}\exp[it(N_{k'}-n_{k'})] \tag{16}$$

exists. If there exists a choice of N_k and n_k for which $n_k(x)-n_k$ converges almost everywhere, then the convergence in (16) is strong. If, moreover, (15) holds, then a number operator exists.

Proof: Necessity in the first paragraph is trivial, and sufficiency is essentially a restatement of Lemma 5.

If (15) holds, then

$$\begin{aligned}|(\Psi,\{\exp[it(N_k-n_k)]-I\}\Psi)| &= |(\Psi,\{\textstyle\int_0^t i(N_k-n_k)\\ &\quad\times\exp[is(N_k-n_k)]\,ds\}\Psi)|\\ &\le K(\Psi)t\end{aligned}$$

so that $\exp[it(N_k-n_k)]$ is equicontinuous at $t=0$ for all $\Psi\in D$.

The first two assertions of the third paragraph are an immediate consequence of the spectral theorem, Stone's theorem, and the fact that the N_k commute with each other. Then (16) follows from the fact that $\exp[itn_k(x)]$ belong to the unit ball in $L_\infty(X,\mu)$, which is w^*-compact. The last assertion follows from the Lebesgue dominated convergence theorem. Indeed, we obtain a.e. convergence of $\exp[itn_k(x)-n_k]$ to a measurable function $g(x)$ of magnitude 1, and multiplication by $g(x)$ defines a unitary operator $A\in\mathcal{M}$. Moreover,

$$\begin{aligned}\|(\exp[it(N_k-n_k)]-A)\Psi\| &= \int|\exp[it(n_k(x)-n_k)]\\ &\quad-g(x)\|\Psi(x)|^2\,d\mu\\ &\to 0\quad(k\to\infty).\end{aligned}$$

The theorem is proved.

Remark: It is quite possible that $D\subset\cap_k D(N_k)$ yet $D\cap D(N)=\{0\}$. It would be of interest to know if this is the case for exponential representations.

[1]J. Chaiken, Commun. Math. Phys. 8, 168 (1968).
[2]G.-F. Dell'Antonio and S. Doplicher. J. Math. Phys. 8, 663 (1967).
[3]J. Chaiken, Ann. Phys. 42, 23 (1967).
[4]J.-P. Eckmann and K. Osterwalder, Helv. Phys. Acta 44, 884 (1971).
[5]J. Fabrey, Commun. Math. Phys. 19, 1 (1970).
[6]J. Fabrey, J. Math. Phys. 14, 909 (1973).
[7]J. Glimm, Commun. Math. Phys. 10, 1 (1968).
[8]J Kunofsky, U.C.L.A. Thesis (1973).

The three-particle S matrix*

Roger G. Newton†

Department of Physics, Indiana University, Bloomington, Indiana 47401
(Received 27 November 1973)

It is proved that a reduced T matrix, defined by factoring out of the three-particle S matrix the product of the three two-particle S matrices, is a *compact* operator on the energy shell, in spite of the double-scattering singularity. As a result there exists a discrete, complete set of eigenphase shifts at every energy.

1. INTRODUCTION

The calculations of three-particle scattering amplitudes are well known to be beset by disconnected-diagram difficulties that are solved in a variety of ways.[1–3] What is much less generally appreciated[4] is that the solution of these mathematical problems still leaves untouched certain other related difficulties that arise in the three-particle S matrix itself on the energy shell. These have to do with the possibility of *double scattering*.[5–7] The part of the T matrix that describes three free particles coming in and emerging freely has infinities at those momenta at which real, on-the-energy-shell double scattering is possible. These are, at a given total energy, the momenta compatible with conservation of energy and momentum in a two-particle collision, followed by a collision of one of these two particles with the third. One of the consequences of this behavior of the T matrix is that the total three-particle collision "cross section" is always infinite. The main reason why this disconcerting result fails to violate our intuition as egregiously as it might is that the three-particle-to-three-particle "cross section" is a rather unintuitive object [of dimension (distance)^5] anyway.

The main purpose of the present paper is to prove that, the double scattering infinity not withstanding, the reduced T matrix, defined by factoring out from the S matrix the three single-scattering S matrices, as an integral operator on the energy shell is *compact*. This result is proved in Sec. 5. In Sec. 6 we discuss some of the consequences, primarily spectral expansions in terms of eigenphase shifts. Section 2 contains the connections between the T operators, the Faddeev equations, and other "connected" formulations.

2. T OPERATORS

Let H_0 be the kinetic-energy operator of the three particles in the center-of-mass system; V_i the potential of the interaction of particles j and k, $i\neq j\neq k$; and

$$G_0^+(E)=(E-H_0+i0)^{-1}, \tag{2.1}$$

$$G_i^+(E)=(E-H_0-V_i+i0)^{-1}, \tag{2.2}$$

$$G^+(E)=(E-H+i0)^{-1}, \tag{2.3}$$

where $H=H_0+\sum V_i$. The two-particle T operators are defined by

$$T_iG_0^+=V_iG_i^+, \tag{2.4}$$

and Faddeev's τ operators by

$$\tau_iG_0^+=V_iG^+, \tag{2.5}$$

$$\tau_{ij}G_j^+=V_iG^+,\quad i\neq j,\quad \tau_{ii}G_i^+=V_i(G^+-G_i^+). \tag{2.6}$$

The resolvent equations

$$G^+=G_i^++G_i^+\sum_{j\neq i}V_jG^+ \tag{2.7}$$

left-multiplied by V_i then yield Faddeev's equations

$$\tau_i=T_i+T_iG_0^+\sum_{j\neq i}\tau_j, \tag{2.8}$$

$$\tau_{ij}=T_i(1-\delta_{ij})+T_iG_0^+\sum_{l\neq i}\tau_{lj}. \tag{2.9}$$

We may also write

$$\tau_i=V_i\Big(1-\sum_j U_j\Big)^{-1},\quad U_j=G_0^+V_j, \tag{2.10}$$

$$\tau_{ij}=\tau_i(1-U_j)-V_i\delta_{ij},$$

and then obtain

$$\tau_i=T_i\Big(1-\sum_{j\neq i}U_j-\sum_{j\neq i}U_j\Gamma_i\Big)^{-1}$$

$$=T_i(1-\Gamma_{-i}\Gamma_i)^{-1}(1+\Gamma_{-i}), \tag{2.11}$$

where

$$\Gamma_i=G_0T_i=U_i(1-U_i)^{-1}, \tag{2.12}$$

$$\Gamma_{-i}=\Big(1-\sum_{j\neq i}U_j\Big)^{-1}\sum_{j\neq i}U_j, \tag{2.13}$$

so that we get

$$\Gamma_{-i}=\sum_{j\neq i}\Gamma_j(1-\Gamma_k\Gamma_j)^{-1}(1+\Gamma_k),\quad i\neq k\neq j. \tag{2.14}$$

Equations (2.8) and (2.9) and the resolvents in (2.11) and (2.14) are "connected."[8]

The T operators are given in terms of the τ operators as follows:

$$T_{00}=\sum_i\tau_i,$$

$$T_{i0}=\sum_{j\neq i}\tau_j,\quad T_{0i}=\sum_j\tau_{ji}, \tag{2.15}$$

$$T_{ij}=\sum_{l\neq i}\tau_{lj}+V_j(1-\delta_{ij}).$$

The notation here is such that T_{ij} is the T operator for a collision in which initially the pair $j=(k,l)$, $k\neq j\neq l$, is bound and finally the pair $i=(l,m)$, $l\neq i\neq m$, is bound; T_{0i} is the T operator for the breakup reaction of pair i colliding with particle i; T_{i0} is the T operator for three free incident particles, with a bound state of pair i in the final state; T_{00} is the T operator for three free particles both in the initial and final states.

The corresponding T matrix elements are given by

$$\mathbf{T}_{00}=(\Psi_0,T_{00}\Psi_0),$$

$$\mathbf{T}_{0i}=(\Psi_0,T_{0i}\Psi_i),\quad \mathbf{T}_{i0}=(\Psi_i,T_{i0}\Psi_0), \tag{2.16}$$

$$\mathbf{T}_{ij}=(\Psi_i,T_{ij}\Psi_j),$$

where Ψ_0 is the free state, in the coordinate representation, in the center-of-mass system, and Ψ_j is the tensor product of a bound state for pair j and free motion for particle j relative to the center-of-mass of pair j.

It was shown in Ref. 9 that time-reversal invariance implies that the T matrix satisfies the reciprocity theorem. The completeness of the bound states of H, plus those states $\Psi^{(j)}$ that correspond asymptotically to Ψ_j, plus those states $\Psi^{(0)}$ that correspond asymptotically to Ψ_0, leads to the unitarity equations, which we write in abbreviated form as

$$\mathbf{T}_{\alpha\beta} - \mathbf{T}^*_{\beta\alpha} = -2\pi i \sum_\gamma \int \mathbf{T}_{\alpha\gamma}\mathbf{T}^*_{\beta\gamma}, \tag{2.17}$$

where α, β, and γ take on the values 0, 1, 2, 3. If we want to be more explicit we have to introduce specific variables and normalizations, as we shall now do.

We shall use the following coordinates and momenta, and their cyclic permutations[10]:

$$\mathbf{r}_1 = (2\mu_1)^{1/2}(\mathbf{R}_2 - \mathbf{R}_3) \equiv (2\mu_1)^{1/2}\mathbf{R}_{23}, \tag{2.18}$$

$$\rho_1 = (2/\overline{\mu}_1)^{1/2}(m_2\mathbf{R}_2 + m_3\mathbf{R}_3) = m_1(2/\overline{\mu}_1)^{1/2}\mathbf{R}_1,$$

$$\mathbf{k}_1 = (2\overline{\mu}_1)^{-1/2}\mathbf{p}_1 = (2\overline{\mu}_1)^{-1/2}(\mathbf{p}_2 + \mathbf{p}_3), \tag{2.19}$$

$$\mathbf{q}_1 = (\mu_1/2)^{1/2}\left(\frac{\mathbf{p}_2}{m_2} - \frac{\mathbf{p}_3}{m_3}\right),$$

$$\mu_1 = m_2m_3/(m_2+m_3), \quad \overline{\mu}_1 = m_1(m_2+m_3)/M, \tag{2.20}$$

$$M = m_1 + m_2 + m_3$$

in terms of the particle coordinates $\mathbf{R}_i$ and momenta $\mathbf{p}_i$ in the center-of-mass system. The momentum $\mathbf{q}_i$ is conjugate to $\mathbf{r}_i$, $\mathbf{k}_i$ is conjugate to ρ_i, and the three-particle Schrödinger equation reads

$$\left[-\nabla_6^2 + \sum_i V_i(\mathbf{r}_i)\right]\Psi = E\Psi, \tag{2.21}$$

where ∇_6^2 is the six-dimensional Laplacean

$$\nabla_6^2 = \frac{\partial^2}{\partial r_i^2} + \frac{\partial^2}{\partial \rho_i^2} \tag{2.22}$$

and we write simply $V_i(\mathbf{r}_i)$ for $V_i(\mathbf{R}_i) = V_i[(2\mu_i)^{-1/2}\mathbf{r}_i]$.

Changes from one coordinate system to another are accomplished by

$$\mathbf{r}_j = a_{ji}\mathbf{r}_i + b_{ji}\rho_i, \qquad \mathbf{q}_j = a_{ji}\mathbf{q}_i + b_{ji}\mathbf{k}_i,$$
$$\rho_j = -b_{ji}\mathbf{r}_i + a_{ji}\rho_i, \quad \mathbf{k}_j = -b_{ji}\mathbf{q}_i + a_{ji}\mathbf{k}_i, \tag{2.23}$$

where for $i,j = 1,2;\ 2,3;\ 3,1$; and $k \neq i \neq j$,

$$a_{ij} = a_{ji} = -(\mu_i\mu_j)^{1/2}/m_k, \quad -b_{ij} = b_{ji} = (\mu_j/\overline{\mu}_i)^{1/2} \tag{2.24}$$

so that $a_{ij}^2 + b_{ij}^2 = 1$. We denote the six-dimensional coordinate vector $(\mathbf{r}_i, \rho_i)$ by $\mathbf{R}$, and the six-dimensional momentum vector $(\mathbf{q}_i, \mathbf{k}_i)$ by $\mathbf{K}$. Then $E = K^2 = |K|^2 = \mathbf{q}_i^2 + \mathbf{k}_i^2$ and the volume elements are $(d\mathbf{K}) = (d\mathbf{q}_i)(d\mathbf{k}_i)$ in momentum space and $(d\mathbf{R}) = (d\mathbf{r}_i)(d\rho_i)$ in configuration space. Our notation for the solid-angle element, for example in $\mathbf{r}$-space, is $(d\hat{\mathbf{r}})$, so that $(d\mathbf{r}) = (d\hat{\mathbf{r}})r^2\,dr$. The inner product $\mathbf{K}\cdot\mathbf{R} = \mathbf{q}_i\cdot\mathbf{r}_i + \mathbf{k}_i\cdot\rho_i$ is also invariant.

It is notationally convenient to think of $T_{\alpha\beta}$ as a 4×4 matrix[11] (with α and β taking on the values 0,1,2,3) whose elements are integral operators on the energy shell, with integral kernels given by the T matrix elements. We shall call it simply T. Then the element kernels are given by $\mathbf{T}_{00}(E;\mathbf{K},\mathbf{K}')$, $\mathbf{T}_{i0}(E;q_{in}^2,\mathbf{k}_i;\mathbf{K}')$, $\mathbf{T}_{0i}(E;\mathbf{K};q_{in}'^2,\mathbf{k}_i')$, and $\mathbf{T}_{ij}(E;q_{in}^2,\ \mathbf{k}_i;q_{jm}'^2,\mathbf{k}_j')$, and the unitarity equation (2.17) reads simply

$$T(E) - T^\dagger(E) = -2\pi i T(E)T^\dagger(E). \tag{2.25}$$

The measure used in the matrix multiplication, according to (3.8) and (3.9), is the following: On 0-elements it is

$$\int (d\mathbf{K})\delta(E-K^2)\cdots \tag{2.26}$$

and on i-elements, $i=1,2,3$, it is

$$(2\mu_i)^{-3/2}\sum_n \int (d\mathbf{k}_i)\delta(E - q_{in}^2 - k_i^2)\cdots, \tag{2.27}$$

with the sum running over the two-body bound states of the pair i.

3. COMPACTNESS OF THE REDUCED T MATRIX

In order to remove from the 3-3 part of the S matrix the single-scattering terms, i.e., the disconnected parts, it is most convenient to define a *reduced* scattering operator. This can be done in a variety of ways and the resulting operator depends on the order in which it is done, but the essential properties of that operator do not.

We have the scattering operator for 3-3 scattering,

$$S_{00} = 1 - 2\pi i\delta_0 T_{00}, \tag{3.1}$$

where $\delta_0 = \delta(E - H_0)$ and

$$T_{00} = \sum_i \tau_i = \sum_i T_i + \sum_{i\neq j} T_i G_0^+ T_j + \tau'. \tag{3.2}$$

Equation (2.8) shows that

$$\tau' = \sum_{i\neq j,\,k\neq j} T_i G_0^+ T_j G_0^+ \tau_k. \tag{3.3}$$

Factoring out $S_1S_2S_3$, where

$$S_i = 1 - 2\pi i\delta_0 T_i \tag{3.4}$$

is the two-particle S operator for the pair j, l, $j\neq i\neq l$, gives[12]

$$S_{00} = S_1S_2S_3S_{00}' \tag{3.5}$$

with

$$S_{00}' = 1 - 2\pi i\delta_0 T_{00}'. \tag{3.6}$$

The reduced T operator is given by

$$T_{00}' = S_3^\dagger S_2^\dagger S_1^\dagger (T_D + T_{00}''), \tag{3.7}$$

in terms of

$$T_{00}'' = \tau' - (2\pi i)^2 T_1\delta_0 T_2\delta_0 T_3, \tag{3.8}$$

$$T_D = \sum_{i<j} (T_i G_0^- T_j + T_j G_0^+ T_i). \tag{3.9}$$

The double-scattering difficulties reside in the operator T_D.

The matrix element of a typical double-scattering term is given by the kernel of the operator $T_1G_0^+T_2$,

$$(\mathbf{K}|\,T_1G_0^+T_2\,|\mathbf{K}') = \int (d\mathbf{K}'')\frac{T_1(E;\mathbf{K},\mathbf{K}'')T_2(E;\mathbf{K}'',\mathbf{K}')}{E-K^2+i0}$$
$$= \frac{t_1(E-k_1^2;\mathbf{q}_1',\mathbf{q}_1'')t_2(E-k_2^2;\mathbf{q}_2',\mathbf{q}_2'')}{b_{12}^3(E-k_1''^2-q_1''^2+i0)}, \tag{3.10}$$

where $\mathbf{q}_1''$ and $\mathbf{q}_2''$ are determined by the double-scattering conditions

$$\mathbf{k}_1 = a_{12}\mathbf{k}_2' - b_{12}\mathbf{q}_2'',$$
$$\mathbf{k}_2' = a_{12}\mathbf{k}_1 + b_{12}\mathbf{q}_1''. \tag{3.11}$$

Now

$$b_{12}^2(E-k_1^2-q_1''^2) = b_{12}^2E - k_1^2 - k_2'^2 + 2a_{12}\mathbf{k}_1\cdot\mathbf{k}_2', \tag{3.12}$$

which vanishes at those values of $\mathbf{k}_1$ and $\mathbf{k}_2'$ for which double scattering can occur on the energy shell. For those values of $\mathbf{k}_1$ and $\mathbf{k}_2'$, therefore, the 3-3 amplitude becomes infinite. What is more, the infinity is not square integrable as a function of either $\mathbf{k}_1$ or $\mathbf{k}_2'$. This means two things: (a) the operator $T_1G_0^+T_2$, of which (3.10) is the integral kernel, is not in the Hilbert-Schmidt class (hereafter called HS); (b) the contribution (3.9) to the 3-3 scattering cross section is such that the total 3-3 scattering cross section is infinite.

We now want to consider T_D as an operator on the energy shell, i.e., as an operator on the space $\mathcal{E} = L^2(\hat{\mathbf{K}})$ of square integrable functions on the sphere $|\mathbf{K}| = K$ in six dimensions. We intend to show that, as such, $T_DT_D^\dagger$ is in the Hilbert–Schmidt class. One of the terms in that product is $T_1G_0^-T_2\delta_0T_2^\dagger G_0^+T_1^\dagger$, and its integral kernel is given by

$$I = (\mathbf{K}|\,T_1G_0^-T_2\delta_0T_2^\dagger G_0^+T_1\,|\mathbf{K}')$$
$$= \int (d\hat{\mathbf{k}}_2'')\int (d\hat{\mathbf{q}}_2'')\tfrac{1}{2}q_2''(\mathbf{K}|\,T_1G_0^-T_2\,|\mathbf{K}'')(\mathbf{K}''|\,T_2^\dagger G_0^+T_1^\dagger\,|\mathbf{K}')$$
$$= \int_0^K dk_2''k_2''^2q_2''$$
$$\times\int (d\hat{\mathbf{k}}_2'')\frac{N(\mathbf{K},\mathbf{K}',\mathbf{k}_2'')}{b_{12}^4(E-i\epsilon-k_1^2-q_1'''^2)(E+i\epsilon-k_1'^2-q_1'^{\nu 2})}, \tag{3.13}$$

where $\mathbf{q}_1$ and $\mathbf{q}_1'^{\nu}$ are given by the equations

$$\mathbf{k}_2'' = a_{21}\mathbf{k}_1 - b_{21}\mathbf{q}_1''' = a_{21}\mathbf{k}_1' - b_{21}\mathbf{q}_1'^{\nu},$$

and the numerator is given by

$$N(\mathbf{K},\mathbf{K}',\mathbf{k}_2'') = \tfrac{1}{2}b_{12}^{-2}\int (d\hat{\mathbf{q}}_2'')t_1(E-k_1^2;\mathbf{q}_1,\mathbf{q}_1''')t_2(E-k_2''^2;\mathbf{q}_2''',\mathbf{q}_2'')$$
$$\times t_2^*(E-k_2''^2;\mathbf{q}_2'^{\nu},\mathbf{q}_2'')t_1^*(E-k_1'^2;\mathbf{q}_1',\mathbf{q}_1'^{\nu})$$

with $\mathbf{q}_2'''$ and $\mathbf{q}_2'^{\nu}$ determined from the equations

$$\mathbf{k}_1' = a_{12}\mathbf{k}_2'' - b_{12}\mathbf{q}_2'^{\nu},\quad \mathbf{k}_1 = a_{12}\mathbf{k}_2'' - b_{12}\mathbf{q}_2'''.$$

It is understood that $E = k_1^2+q_1^2 = k_1'^2+q_1'^2 = k_2''^2+q_2''^2$.

Thus I is of the form

$$I = \int_0^K dk_2''k_2''^2q_2''J(k_2'')/a'b', \tag{3.14}$$

where

$$J(k_2'') = \int (d\hat{\mathbf{n}})\frac{N(\hat{\mathbf{n}})}{(1+\mathbf{a}\cdot\hat{\mathbf{n}})(1+\mathbf{b}\cdot\hat{\mathbf{n}})} \tag{3.15}$$

with

$$\mathbf{a} = 2a_{12}k_2''\mathbf{k}_1/a',\quad \mathbf{b} = 2a_{12}k_2''\mathbf{k}_1'/b',$$
$$a' = b_{12}{}^2E - i\epsilon - k_1{}^2 - k_2''^2,\quad b' = b_{12}{}^2E + i\epsilon - k_1'^2 - k_2''^2. \tag{3.16}$$

For generic values of $\mathbf{a}$ and $\mathbf{b}$ the two denominators in (3.15) vanish simultaniously (for $\epsilon = 0$) at two values of $\hat{\mathbf{n}}$ that are symmetric with respect to the plane of $\mathbf{a}$ and $\mathbf{b}$. We handle each half-space separately. Let $\hat{\mathbf{n}}_1$ be the value of $\hat{\mathbf{n}}$ in one half space, S_1, for which both denominators vanish. Then

$$J = \int_{S_1}(d\hat{\mathbf{n}})\frac{N(\hat{\mathbf{n}})-N(\hat{\mathbf{n}}_1)}{(1+\mathbf{a}\cdot\hat{\mathbf{n}})(1+\mathbf{b}\cdot\hat{\mathbf{n}})} + \int_{S_2}(d\hat{\mathbf{n}}_2)\frac{N(\hat{\mathbf{n}})-N(\hat{\mathbf{n}}_2)}{(1+\mathbf{a}\cdot\hat{\mathbf{n}})(1+\mathbf{b}\cdot\hat{\mathbf{n}})}$$
$$+[N(\hat{\mathbf{n}}_1)+N(\hat{\mathbf{n}}_2)]\tfrac{1}{2}\int (d\hat{\mathbf{n}})(1+\mathbf{a}\cdot\hat{\mathbf{n}})^{-1}(1+\mathbf{b}\cdot\hat{\mathbf{n}})^{-1}.$$

If the two-particle T matrices, and hence N, satisfy Lipschitz conditions near the energy shell, then the first two integrals converge even for $\epsilon = 0$ and we need not consider them further. They will in any event be better behaved than the third. Hence we consider the third,

$$J' = \int (d\hat{\mathbf{n}})(1+\mathbf{a}\cdot\hat{\mathbf{n}})^{-1}(1+\mathbf{b}\cdot\hat{\mathbf{n}})^{-1}. \tag{3.17}$$

If we write $\mathbf{a} = (a+i\xi)\hat{\mathbf{a}}$, $\mathbf{b} = (b+i\eta)\hat{\mathbf{b}}$ and we assume that $a > b > 0$, then we get in the limit, as $\xi \to 0$ and $\eta \to 0$,

$$\chi J'/2\pi = \begin{cases} L, & \text{if } a<1,\ b<1,\\ L+i\pi\,\mathrm{sgn}\xi, & \text{if } a>1,\ b<1,\\ L+i\pi(\mathrm{sgn}\xi-\mathrm{sgn}\eta), & \text{if } a>1,\ b>1,\\ \qquad \mathbf{a}\cdot\mathbf{b}>1+[(a^2-1)(b^2-1)]^{1/2},\\ Q+\pi|\,\mathrm{sgn}\xi-\mathrm{sgn}\eta\,|, & \text{if } a>1,\ b>1,\\ \qquad b^2<\mathbf{a}\cdot\mathbf{b}<1+[(a^2-1)(b^2-1)]^{1/2},\\ L+i\pi(\mathrm{sgn}\xi+\mathrm{sgn}\eta), & \text{if } a>1,\ b>1,\\ \qquad \mathbf{a}\cdot\mathbf{b}<1-[(a^2-1)(b^2-1)]^{1/2},\\ Q-\pi|\,\mathrm{sgn}\xi+\mathrm{sgn}\eta\,|, & \text{if } a>1,\ b>1,\\ \qquad b^2>\mathbf{a}\cdot\mathbf{b}>1-[a^2-1)(b^2-1)]^{1/2}, \end{cases} \tag{3.18}$$

where

$$L = \ln\left|\frac{\mathbf{a}\cdot\mathbf{b}-1-\chi}{\mathbf{a}\cdot\mathbf{b}-1+\chi}\right|,$$

$$Q = 2\tan^{-1}\frac{\chi}{a^2-\mathbf{a}\cdot\mathbf{b}} + 2\tan^{-1}\frac{\chi}{|\,b^2-\mathbf{a}\cdot\mathbf{b}\,|},$$

$$\chi = |(\mathbf{a}\cdot\mathbf{b}-1)^2-(a^2-1)(b^2-1)|^{1/2}$$
$$= |(\mathbf{a}-\mathbf{b})^2-(\mathbf{a}\times\mathbf{b})^2|^{1/2}.$$

Consequently, there are two kinds of integrals in I that have to be tested, one from the constants and Q (which is bounded) in $J'\chi$,

$$I_1 = \int_0^K \frac{dk_2''k_2''q_2''}{a'b'\chi}, \tag{3.19}$$

and the other from L,

$$I_2 = \int_0^K \frac{dk_2''k_2''q_2''L'}{a'b'},\quad L' = L/\chi. \tag{3.20}$$

We first consider I_1, which may be written

$$I_1 = \frac{1}{4a_{12}|\mathbf{k}_1-\mathbf{k}_1'|}\int_0^{K_2}\frac{dx(K^2-x)^{1/2}}{(x^2-2cx+d^2)^{1/2}},$$

where $x=k_2''^2$ and

$$c=b_{12}^2E+k_1k_1'\alpha+a_{12}^2(k_1^2+k_1'^2)\beta,$$
$$d^2=b_{12}^4E^2+k_1^2k_1'^2+2b_{12}^2Ek_1k_1'\alpha,$$

α and β being defined by

$$(\mathbf{k}_1-\mathbf{k}_1')\cdot(\hat{\mathbf{k}}_1k_1'-\hat{\mathbf{k}}_1'k_1)=(\mathbf{k}_1-\mathbf{k}_1')^2\alpha,$$
$$(\mathbf{k}_1\times\mathbf{k}_1')^2/(\mathbf{k}_1-\mathbf{k}_1')^2=\tfrac{1}{2}(k_1^2+k_1'^2)\beta.$$

It is easy to see that no matter what values $\mathbf{k}$, and $\mathbf{k}_1'$ take on, we always have

$$|\alpha|\leq 1,\quad 0\leq\beta\leq 1.$$

Now

$$\frac{1}{K}\int_0^{K^2}\frac{dx(K^2-x)^{1/2}}{(x^2-2cx+d^2)^{1/2}}\leq\int_0^{K^2}\frac{dx}{(x^2-2cx+d^2)^{1/2}}$$
$$=\ln\left[\left(\left|\frac{c+d}{c-d}\right|\right)^{1/2}\left(\frac{1+(1\pm y^2)^{1/2}}{y}\right)\right]$$
$$\text{if } d\gtrless c,$$

where $y=(|c^2-d^2|)^{1/2}/|K^2-c|$. Consequently, it is clear that

$$\frac{1}{K}\int^{K^2}\frac{dx(K^2-x)^{1/2}}{(x^2-2cx+d^2)^{1/2}}\leq\text{const}\left|\ln|c^2-d^2|\right| \tag{3.21}$$

and, therefore,

$$|I_1|\leq\frac{\text{const}}{|\mathbf{k}_1+\mathbf{k}_1'|}\ \left|\ln|c^2-d^2|\right|. \tag{3.22}$$

The right-hand side of this inequality is square integrable with respect to $\mathbf{k}_1$ and $\mathbf{k}_1'$, in the ball $k_1\leq K$, $k_1'\leq K$.

Next we test I_2 of (5.20) for square integrability. Using Schwarz's inequality, we have

$$|I_2|^2=\left|\int_0^K dk_2''k_2''^2q_2''L'/a'b'\right|^2$$
$$\leq\left(\int_0^K dk_2''k_2''^2q_2''^2\right)\int_0^K dk_2''k_2''^2|L'|^2/a'^2b'^2$$
$$=\text{const}\int_0^K dk_2''k_2''|L'|^2/a'^2b'^2.$$

At this point we exchange orders of integration, deferring the k_2''-integral and performing the integration over $\mathbf{k}_1$ and $\mathbf{k}_1'$ first. (This exchange in the order of integration is justified by the Fubini–Tonelli theorem.[13]) The first step is the integration over the directions of $\mathbf{k}_1'$, i.e., of $\mathbf{b}$, using a z axis in the direction of $\mathbf{a}$. We introduce the abbreviations $A=1-a^2$, $B=1-b^2$, $t=(1-\mathbf{a}\cdot\mathbf{b})/(AB)^{1/2}$, where $a=|\mathbf{a}|$, $b=|\mathbf{b}|$, and obtain

$$L'=(AB)^{-1/2}|t^2-1|^{-1/2}\ln\frac{t+(t^2-1)^{1/2}}{t-(t^2-1)^{1/2}}.$$

Hence

$$\int(d\hat{\mathbf{b}})|L'|^2=\frac{2\pi}{ab}(AB)^{1/2}\int_{-t_0}^{t_0}dt|L'|^2$$
$$=\frac{2\pi}{ab(AB)^{1/2}}\int_{-t_0}^{t_0}\frac{dt}{|t^2-1|}\left|\ln\left|\frac{t+(t^2-1)^{1/2}}{t-(t^2-1)^{1/2}}\right|\right|^2$$

where $t_0=ab(AB)^{-1/2}$. The t integral being uniformly bounded with respect to t_0, we have

$$\int(d\hat{\mathbf{k}}_1)(d\hat{\mathbf{k}}_1')|L'|^2\leq\frac{\text{const}}{ab(AB)^{1/2}}$$

and hence

$$\int_{k_1<K,k_1'<K}(d\mathbf{k}_1)(d\mathbf{k}_1')|I_2|^2\leq\text{const}\int_0^K dk_2''k_2''^2\int_0^K dk_1k_1^2$$
$$\times\int^K\frac{dk_1'k_1'^2}{a'^2b'^2ab(AB)^{1/2}}=\text{const}\int_0^K dk_2''$$
$$\times\left(\int_0^K\frac{dk_1k_1}{[k_1^2-(b_{12}q_2''-a_{12}k_2'')^2]^{1/2}[k_1^2-(b_{12}q_2''+a_{12}k_2'')^2]^{1/2}}\right)^2.$$

The integral over k_1 is of the form

$$\int_0^{x_0}\frac{dx}{|(x-\alpha)(x-\beta)|^{1/2}}\leq\text{const}\left|\ln|\alpha-\beta|\right|$$
$$=\text{const}\left|\ln 4a_{12}b_{12}q_2''k_2''\right|$$

by the same argument as in (5.21), with $\alpha=(b_{12}q_{12}''-a_{12}k_2'')$, $\beta=(b_{12}q_2''+a_{12}k_2'')^2$. We therefore get

$$\int_{k_1<K,k_1'<K}(d\mathbf{k}_1)(d\mathbf{k}_1')|I_2|^2\leq\text{const}\int_0^K dk_2''[\ln(k_2''q_2'')]^2$$
$$\leq\text{const}. \tag{3.23}$$

Thus we have shown that I of (5.13) is square integrable on the energy shell. The same argument applies to the term $T_1G_0^-T_3\delta_0T_3^\dagger G_0^+T_2^\dagger$ and to all terms of the form $T_iG_0^\pm T_j\delta_0T_j^\dagger G_0^\mp T_k^\dagger$ with $i\neq j$, $k\neq j$.

Let us now consider a term such as $T_1G_0^-T_2\delta_0T_i^\dagger G_0^+T_j^\dagger$ with $i\neq 2$. The second denominator in (5.13) now becomes

$$b_{ij}^2(E-k_j'^2-q_j''^2)=b_{ij}^2E-k_i''^2-k_j'^2+2a_{ij}\mathbf{k}_i''\cdot\mathbf{k}_j'$$

with $\mathbf{k}_i''=a_{i2}\mathbf{k}_2''-b_{i2}\mathbf{q}_2''$. Hence it is given by

$$b_{ij}^2E-a_{i2}^2k_2''^2-b_{i2}^2q_2''^2-k_j'^2+2a_{ij}\mathbf{k}_2''\cdot\mathbf{k}_j'$$
$$+2b_{i2}\mathbf{q}_2''\cdot(a_{i2}\mathbf{k}_2''-a_{ij}\mathbf{k}')$$

where $q_2''^2=E-k_2''^2$. In contrast to (3.12) this depends on the direction of $\mathbf{q}_2''$, which is integrated over. Thus, the singularity produced by the vanishing of the denominator is changed from a pole into a logarithmic one, and since we have shown that the possibility of simultaneous vanishing of the two denominators still keeps the function square integrable even when that causes a double pole, the term $T_1G_0^-T_2\delta_0T_j^\dagger G_0^+T_i^\dagger$ will also be square integrable. Thus, all terms of the form $T_iG_0^\pm T_j\delta_0T_kG_0^\mp T_l$ are in HS. Moreover, the same applies to $T_iG_0^-T_j\delta_0T_k^\dagger G_0^-T_l^\dagger$ and $T_iG_0^+T_j\delta_0T_k^\dagger G_0^+T_l^\dagger$. Equations (3.18) show that there are no singularities worse than we have examined. The double-scattering term T_D in S_{00}', as an operator from $\mathcal{E}$ to $\mathcal{E}$, is therefore in HS.

Next we examine the explicit, "real" triple-scattering term in (3.8)

$$(\mathbf{K}|T_1\delta_0T_2\delta_0T_3|\mathbf{K}')$$
$$=\int(d\hat{\mathbf{q}}_1'')t_1(E-k_1^2;\mathbf{q}_1,\mathbf{q}_1'')t_2(E-k_2''^2;\mathbf{q}_2'',\mathbf{q}_2''')$$
$$\times t_3(E-k_3'^2;\mathbf{q}_3''',\mathbf{q}_3')\,\delta(q_3'^2-q_3'''^2)q_1''/2b_{23}{}^3, \tag{3.24}$$

where

$$b_{23}\mathbf{q}_3'''=b_{21}\mathbf{q}_1''+a_{23}\mathbf{k}_3'-a_{21}\mathbf{k}_1,$$
$$\mathbf{q}_2'''=a_{23}\mathbf{q}_3'''+b_{23}\mathbf{k}_3',$$
$$\mathbf{q}_2''=a_{21}\mathbf{q}_1''+b_{21}\mathbf{k}_1, \tag{3.25}$$
$$\mathbf{k}_2''=a_{21}\mathbf{k}_1-b_{21}\mathbf{q}_1'',\quad q_1''=q_1.$$

One of the angle-integrations left can be done by means of the δ function. Using a coordinate system in which $a_{23}\mathbf{k}_3' - a_{21}\mathbf{k}_1$ lies along the z axis, we get

$$(\mathbf{K}|\, T_1\delta_0 T_2\delta_0 T_3\,|\mathbf{K}') = \int d\varphi \frac{t_1(q_1^2;\mathbf{q}_1,\mathbf{q}_1'')t_2(q_2''^2;\mathbf{q}_2'',\mathbf{q}_2''')t_3(q_3'^2;\mathbf{q}_3''',\mathbf{q}_3')}{2b_{23}|a_{23}\mathbf{k}_3' - a_{21}\mathbf{k}_1|},$$

where

$$(b_{21}\mathbf{q}_1'' + a_{23}\mathbf{k}_3' - a_{21}\mathbf{k}_1)^2 = b_{23}^2 q_3'^2$$

now determines the angle between $\mathbf{q}_1''$ and $a_{23}\mathbf{k}_3' - a_{21}\mathbf{k}_1$, and φ is the azimuthal angle of $\mathbf{q}_1''$ about that axis. Provided, therefore, that the two-body T matrices t_1, t_2, t_3, on the energy shell, are bounded functions, the above is square integrable over $\mathbf{K}$ and $\mathbf{K}'$ with $|\mathbf{K}| = |\mathbf{K}'| = K$, and $T_1\delta_0 T_2\delta_0 T_3$, as an operator from $\mathcal{E}$ to $\mathcal{E}$, is HS.

Finally, there is the operator τ', given by (3.3). According to (2.11) we have

$$\tau' = \sum_{\substack{i\neq j\\ l\neq j}} T_i G_0^+ T_j G_0^+ T_l C_l, \tag{3.26}$$

where

$$C_l = (1 - \Gamma_{-l}\Gamma_l)^{-1}(1 + \Gamma_{-l}) = 1 + \Gamma_{-l}(1 - \Gamma_l\Gamma_{-l})^{-1}\Gamma_l + (1 - \Gamma_{-l}\Gamma_l)^{-1}\Gamma_{-l} \tag{3.27}$$

and Γ_l and Γ_{-l} are given by (2.12) and (2.14).

The kernel of $T_1 G_0^+ T_2 G_0^+ T_3$ is

$$(\mathbf{K}|\, T_1 G_0^+ T_2 G_0^+ T_3\,|\mathbf{K}') = \int (d\mathbf{q}_1'') \frac{t_1(q_1^2;\mathbf{q}_1,\mathbf{q}_1'')t_2(E - k_2''^2;\mathbf{q}_2'',\mathbf{q}_2''')t_3(q_3'^2;\mathbf{q}_3''',\mathbf{q}_3')}{(q_1^2 - q_1'' + i\epsilon)(E - q_3'''^2 - k_3'^2 + i\epsilon)}. \tag{3.28}$$

We must now take $\mathbf{K}$ on the energy shell, but not $\mathbf{K}'$. Thus, we consider $T_1 G_0^+ T_2 G_0^+ T_3$ as an operator from the whole Hilbert space $\mathcal{H} = L^2(R^2)$ to the energy shell $\mathcal{E} = L^2(\hat{\mathbf{K}})$, and we want to test if its kernel (3.28) is square integrable.

First we note that

$$T_1 G_0^+ T_2 G_0^+ T_3 = u_1 \Omega Y u_3, \tag{3.29}$$

where

$$\Omega = 1 + v_1 G_1^+ u_1, \quad Y = v_1 G_2^+ V_2 G_3^+ v_3,$$
$$u_1 = |V_1|^{1/2}, \quad V_1 = v_1 u_1.$$

It was shown in Ref. 9 that if the potentials are bounded and decrease at infinity faster than $r^{-3-\epsilon}$, $\epsilon > 0$, and if there are no zero-energy resonances or zero-energy bound states in the two-body systems, then, as an operator from $\mathcal{H}$ to $\mathcal{H}$, Y is HS. The operator Ω is bounded. Hence ΩY is in HS. Now the question is if the kernel of $u_1\Omega Y$, as a mapping from $\mathcal{H}$ to $\mathcal{E}$ is square integrable. That is a question of *pointwise* convergence of the integrals at $q = (E - k^2)^{1/2}$, as becomes clear when we write

$$\int (d\mathbf{K}) \cdots = \tfrac{1}{2}\int_0^\infty dE \int_{k^2<E} (d\mathbf{k})(d\hat{\mathbf{q}})(E - k)^{1/2} \cdots,$$

which is an integral over "all energy shells." A function $f(\mathbf{K}) = f(E, \hat{\mathbf{q}}, \mathbf{k})$ that is square integrable over E need not be finite at a specific value of E. However, suppose that $f(\mathbf{q},\mathbf{k})$ is square integrable.

$$\int (d\mathbf{q})(d\mathbf{k})|f(\mathbf{q},\mathbf{k})|^2 < \infty.$$

Then it follows that

$$g(\mathbf{q},\mathbf{k}) = \int (d\mathbf{q}')u(\mathbf{q} - \mathbf{q}')f(\mathbf{q}',\mathbf{k})$$

is square integrable on every energy shell if $u(\mathbf{q})$ is square integrable, because by Schwarz's inequality

$$|g(\mathbf{q},\mathbf{k})|^2 \leq \int (d\mathbf{q}')|u(\mathbf{q}')|^2 \int (d\mathbf{q}'')|f(\mathbf{q}'',\mathbf{k})|^2.$$

This argument shows both that $u_1\Omega Y$ is an HS operator from $\mathcal{H}$ to $\mathcal{E}$, and that $u_1\Omega Y u_3$ is an HS operator from $\mathcal{E}$ to $\mathcal{E}$ (provided that the assumption of Ref. 9 are satisfied; square integrability of u_1 and u_3, i.e., absolute integrability of V_1 and V_3, are then assured).

There remains the factor C_3 of (3.26) and (3.27). The "1" term in it has already been covered, since $T_1 G_0^+ T_2 G_0^+ T_3$ has been shown to be HS from $\mathcal{E}$ to $\mathcal{E}$. The remaining terms in (3.27) are both of the form Au_i, where u_i is square integrable and A is a bounded operator from $\mathcal{H}$ to $\mathcal{H}$, provided that E does not have a value that coincides with a bound state.[14] Hence, the above argument implies that $T_1 G_0^+ T_2 G_0^+ T_3 C_3$ is HS from $\mathcal{E}$ to $\mathcal{E}$ Therefore, we have shown that τ' is HS from $\mathcal{E}$ to $\mathcal{E}$, i.e., its kernel is square integrable on the energy shell (unless E is the energy of a bound state in the continuum), and so is T_{00}''.

Now the operators S_i being unitary, it follows from the unitarity of S that S_{00}' is bounded. Hence, $\delta_0 T_{00}' = (1 - S_{00}')/2\pi i$ is bounded and thus T_{00}', as an operator from $\mathcal{E}$ to $\mathcal{E}$ is bounded, and so is $T_D + T_{00}''$. but T_{00}'' has been shown to be HS on $\mathcal{E}$, and hence bounded. Therefore, T_D is a bounded operator from $\mathcal{E}$ to $\mathcal{E}$. Since it was shown that $T_D\delta_0 T_D^\dagger$ is HS (from $\mathcal{E}$ to $\mathcal{E}$), it follows that on the energy shell, T_D is compact.[15] Therefore, so is T_{00}'.

4. SOME CONSEQUENCES

Let us discuss some of the implications of the compactness of the reduced T operator (3.7) on the energy shell. The most important consequence is that the spectrum of T_{00}' (other than zero) is discrete, i.e., a point spectrum only, each (except possibly the point zero) with finite degeneracy, and accumulating at zero. If we form a super-Hilbert space[11] $\overline{\mathcal{H}}$ as a direct sum of the four channel spaces and call S the scattering operator (on the energy shell) on this space, then the reduced S matrix is given by

$$\mathbf{S}' = \mathbf{S}_3^{-1}\mathbf{S}_2^{-1}\mathbf{S}_1^{-1}\mathbf{S} = 1 - 2\pi i\mathbf{T}', \tag{4.1}$$

where $\mathbf{S}_i$ is given by

$$(\mathbf{S}_i)_{\alpha\beta} = S_i\delta_{0\alpha}\delta_{0\beta} + (1 - \delta_{0\alpha})\delta_{\alpha\beta}. \tag{4.2}$$

Then $\mathbf{S}$, $\mathbf{S}'$, and $\mathbf{S}_i$ are all unitary, and $\mathbf{T}'$ is compact. Hence, the spectrum of $\mathbf{S}'$ is discrete, of finite degeneracy,[16] and accumulates at the point 1. Since $\mathbf{S}'$ is unitary, the eigenfunctions furthermore form a complete orthogonal set on $\overline{\mathcal{H}}$. So, even though above the break-up threshold there is a continuity of energy distributions between the particles at each fixed total energy, we may define an infinite *discrete* set of real eigenphase shifts δ_n by calling the eigenvalues of $\mathbf{S}'$,

$\exp(2i\delta_n)$. This makes the eigenvalues of $\mathbf{T}'$ equal to $-(1/\pi)\exp(i\delta_n)\sin\delta_n$. As $n\to\infty$, we must have $\sin\delta_n\to 0$ in such a way that

$$\sum_n d_n \sin^4\delta_n < \infty, \tag{4.3}$$

where d_n is the degeneracy[17] of δ_n. This follows from the fact that on $\mathcal{E}$

$$tr(\mathbf{T}'\mathbf{T}'^{\dagger})^2 < \infty$$

as shown in Sec. 3, and unitarity.

If we call the eigenfunctions of $\mathbf{T}'$ on the energy shell $\gamma_n^{(\alpha)}(\mathbf{K})$, then

$$\begin{aligned}
&-\pi\int(d\mathbf{K}')\delta(E-K'^2)\mathbf{T}'_{00}(E;\mathbf{K},\mathbf{K}')\gamma_n^{(0)}(\mathbf{K}')\\
&-\pi\sum_{j,m}\int(d\mathbf{k}'_j)\delta(E-q'^2_{jm}-k'^2_j)\mathbf{T}'_{0j}(E;\mathbf{K},q'^2_{jm},\mathbf{k}'_j)\gamma_n^{(j)}(q'^2_{jm},\mathbf{k}'_j)\\
&\quad=\exp[i\delta_n(E)]\sin\delta_n(E)\gamma_n^{(0)}(\mathbf{K}),
\end{aligned}\tag{4.4}$$

$$\begin{aligned}
&-\pi\int(d\mathbf{K}')\delta(E-K'^2)\mathbf{T}_{j0}(E;q^2_{jm},\mathbf{k}_j;\mathbf{K}')\gamma_n^{(0)}(\mathbf{K}')\\
&-\pi\sum_{i,l}\int(d\mathbf{k}'_i)\delta(E-q'^2_{il}-k'^2_i)\mathbf{T}_{ji}(E;q^2_{jm},\mathbf{k}_j;q'^2_{il},\mathbf{k}'_i)\gamma_n^{(i)}(q'^2_{il},\mathbf{k}'_i)\\
&\quad=\exp[i\delta_n(E)]\sin\delta_n(E)\gamma_n^{(j)}(q^2_{jm},\mathbf{k}_j).
\end{aligned}\tag{4.5}$$

They are mutually orthogonal with the weights (2.26) or (2.27), respectively. There then exist the representations, in the sense of strong convergence,[18]

$$-\pi\mathbf{T}'_{\alpha\beta}(E;\mathbf{K}',\mathbf{K})=\sum_1^\infty\gamma_n^{(\alpha)}(\mathbf{K}')\gamma_n^{(\beta)*}(\mathbf{K})\exp[i\delta_n(E)]\sin\delta_n(E), \tag{4.6}$$

or, more explicitly,

$$-\pi\mathbf{T}'_{00}(E;\mathbf{K}',\mathbf{K})=\sum_n\gamma_n^{(0)}(\mathbf{K}')\gamma_n^{(0)*}(\mathbf{K})\exp(i\delta_n)\sin\delta_n, \tag{4.7}$$

$$-\pi\mathbf{T}'_{0i}(E;\mathbf{K}',q^2_{in},\mathbf{k}_i)=\sum_n\gamma_n^{(0)}(\mathbf{K}')\gamma_n^{(i)*}(q^2_{in},\mathbf{k}_i)\exp(i\delta_n)\sin\delta_n, \tag{4.8}$$

$$\begin{aligned}
-\pi\mathbf{T}'_{i0}(E;q^2_{im},\mathbf{k}_{ij}\mathbf{K}')&=-\pi\mathbf{T}_{i0}(E;q^2_{im},\mathbf{k}_{ij}\mathbf{K}')\\
&=\sum_n\gamma_n^{(i)}(q^2_{im},\mathbf{k}_i)\gamma_n^{(0)*}(\mathbf{K}')\exp(i\delta_n)\sin\delta_n,
\end{aligned}\tag{4.9}$$

$$\begin{aligned}
-\pi\mathbf{T}'_{ij}(E;q'^2_{im},\mathbf{k}'_i;q'^2_{jl},\mathbf{k}_l)&=-\pi\mathbf{T}_{ij}(E;q'^2_{im},\mathbf{k}'_i;q'^2_{jl},\mathbf{k}_l)\\
&=\sum_n\gamma_n^{(i)}(q'^2_{im},\mathbf{k}'_i)\gamma_n^{(j)*}(q^2_{jl},\mathbf{k}_l)\\
&\quad\times\exp(i\delta_n)\sin\delta_n.
\end{aligned}\tag{4.10}$$

Equations (4.7) and (4.8) imply that

$$-\pi\mathbf{T}_{00}(E;\mathbf{K}',\mathbf{K})=\sum_n\gamma_n^{(0)'}(\mathbf{K}')\gamma_n^{(0)*}(\mathbf{K})\exp(i\delta_n)\sin\delta_n \tag{4.7'}$$

and

$$-\pi\mathbf{T}_{0i}(E;\mathbf{K}',\mathbf{K})=\sum\gamma_n^{(0)'}(\mathbf{K}')\gamma_n^{(0)*}(\mathbf{K})\exp(i\delta_n)\sin\delta_n, \tag{4.8'}$$

where

$$\gamma_n^{(0)'}=(S_1S_2S_3)\gamma_n^{(0)} \tag{4.11}$$

in which γ_n and all three S_i are on the energy shell. Below the break-up threshold, of course, both sides of (4.4) and the first term on the left of (4.5) vanish.

There are two relevant remarks to be made here. One is that the representations (4.7'), (4.8'), (4.9), and (4.10) do not manifestly obey reciprocity when time reversal invariance holds. It must be remembered that $\exp(2i\delta_n)$ is an eigenvalue of S' which depends on the side and order in which the three S_i have been factored out. Hence the eigenphase shifts depend on that, too.

The second remark is that since the values of $\mathbf{K}'$ at which the double-scattering divergences occur in $\mathbf{T}_{00}$ depend on $\mathbf{K}$, these cannot be divergences in the functions $\gamma_n^{(0)}$. They must, in (4.7) or (4.7'), manifest themselves as divergences in the expansion. In other words, the series in (4.7) and (4.7'), although strongly convergent, must be pointwise divergent at the double-scattering momenta.

ACKNOWLEDGMENT

I am happy to acknowledge useful discussions with Dr. R. Shtokhamer.

*A preliminary account of the main result of this paper was given at the International Conference on Mathematical Problems of Quantum Field Theory and Quantum Statistics, Moscow, U.S.S.R., December, 1972.

†Supported in part by the National Science Foundation and the U.S. Army Research Offices, Durham, North Carolina.

[1]L.D. Faddeev, Zh. Eksp. Teor. Fiz. 39, 1459 (1960) [Sov. Phys. JETP 12, 1014 (1961)]; Dokl. Akad. Nauk SSSR 138, 565 (1961) and 145, 301 (1962) [Sov. Phys. Dokl. 6, 384 (1961) and 7, 600 (1963)]; *Mathematical Aspects of the Three-Body Problem in Quantum Scattering Theory* (Davey, New York, 1965).

[2]S. Weinberg, Phys. Rev. 133, B232 (1964); C. van Winter, Kgl. Danske Videnskab. Selskab, *Mat.-fys. Skrifter* 2, No. 8 (1964) and No. 10 (1965); L. Rosenberg, Phys. Rev. 135, B715 (1964).

[3]R.G. Newton, J. Math. Phys. 8, 851 (1967).

[4]Some examples of this unawareness are J.A. Wright, Phys. Rev. 139, B137 (1965); D.D. Brayshaw, Phys. Rev. D6, 196 (1972).

[5]E. Gerjuoy, Proc. Phys. Soc. London, J. Phys. B 3, L92 (1970); E. Gerjuoy, Phil. Trans. Roy. Soc. London A 270, 197 (1971); J. Nuttall, J. Math. Phys. 12, 1896 (1971).

[6]S.P. Merkuriev, Theor. i Mat. Fiz. 8, 235 (1971).

[7]R.G. Newton, Ann. of Phys. 74, 324 (1972) and 78, 561 (1973).

[8]For more specific mathematical properties, see M. Rubin, R. Sugar, and G. Tiktopoulos, Phys. Rev. 159, 1348 (1967) and Refs. 1 and 9.

[9]R.G. Newton, J. Math. Phys. 12, 1552 (1971).

[10]The notation is the same as in Ref. 7.

[11]The idea of the super-Hilbert space here is the same as used by J.A. Wright, Ref. 4, and in more general abstract form by R. Coester, *Helv. Phys. Acta* 38, 7 (1965) and C. Chandler and A.G. Gibson, preprint.

[12]This has been done in the same way by J.A. Wright, Ref. 4.

[13]See, for example, R.G. Bartle, *The Elements of Integration* (Wiley, New York, 1966), p. 118.

[14]For a large class of interactions it has been proved that bound states embedded in the continuum above the break-up threshold cannot occur; see B. Simon, preprint.

[15]N.I. Akhiezer and I.M. Glazman, *Theory of Linear Operators in Hilbert Space*, (Unger, New York, 1961), Vol. 1, p. 57.

[16]Except possibly the point 1.

[17]If the interactions are invariant under rotation then the total angular momentum is conserved and the normal degeneracy of an eigenphase shift of angular momentum J is at least $2J+1$. There will in general be infinitely many eigenphase shifts of a given J, though.

[18]Ref. 15, p. 131.

Generalized free fields and the representations of Weyl group

J. Lukierski*

Institute of Physics, Catholic University of Rio De Janeiro (PUC), Brazil

K. Sienkiewicz

Institute of Theoretical Physics, University of Wroclaw, Wroclaw, Poland
(Received 15 February 1972; revised manuscript received 12 June 1972)

The transformation properties of generalized free fields under the transformations of the Weyl group, and particularly under the subgroup of dilatations, are discussed. It is shown that there exists, for any complex value d of the dimensionality parameter, a generalized free field, defining by means of its one-particle states the suitable irreducible representation of the Weyl group.

I. INTRODUCTION

First we recall briefly some group-theoretic results concerning the Weyl group and its unitary ray representations.[1-3]

The Weyl group consists of the following elements: dilatations $l(l>0)$, homogenous Lorentz transformations Λ, and translations a. Denoting an element of the Weyl group W by (a, Λ, d), the multiplication law of the group has the form:

$$(a', \Lambda', l')(a, \Lambda, l) = (a' + \Lambda' l' a, \Lambda' \Lambda, l' l). \tag{1.1}$$

It follows from (2.1 that

$$W = T \otimes DL_4, \quad DL_4 = L_4 \times S, \tag{1.2}$$

where T denotes the group of translations, L_4 the homogeneous Lorentz group, and S the one-parameter transformations of scale (dilatations). The commutators for the generators of the Weyl group are the following:

$$[\hat{M}_{\mu\nu}, \hat{M}_{\rho\tau}] = i(g_{\mu\rho}\hat{M}_{\nu\tau} + g_{\nu\tau}\hat{M}_{\mu\rho} - g_{\mu\tau}\hat{M}_{\nu\rho} - g_{\nu\rho}\hat{M}_{\mu\tau}), \tag{1.3a}$$

$$[\hat{M}_{\mu\nu}, \hat{P}_\rho] = i(g_{\mu\rho}\hat{P}_\nu - g_{\nu\rho}\hat{P}_\mu), \tag{1.3b}$$

$$[\hat{M}_{\mu\nu}, \hat{D}] = 0, \tag{1.3c}$$

$$[\hat{D}, \hat{P}_\mu] = i\hat{P}_\mu, \tag{1.3d}$$

$$[\hat{P}_\mu, \hat{P}_\nu] = 0. \tag{1.3e}$$

We obtain all representations of W by investigating all representations of the universal covering group $\overline{W}$. Every irreducible unitary ray representation of $\overline{W}$ can be found by the method of induced representations, given by Mackey,[4] which is in particular applicable to the semidirect product of a separable and locally compact group H and an Abelian group N. For the construction of irreducible representations of $H \otimes N$, induced on N, it is necessary to specify:

(a) The characters $\hat{N}$ on N.

(b) The orbits in $\hat{N}$ under the action of H.

In our case $N = T_4$ and $H = DL_4$. The group $\hat{T}_4$ of characters consists of the functions $\exp(ip_\mu x^\mu)$. The group DL_4 generates equivalence classes in $\hat{T}_4$; the invariant subspace with respect to the action of DL_4 in the space of eigenvalues P_μ is called an orbit. One gets the following six orbits:

$$O_1: \text{all } p_\mu = 0, \qquad O_4: p^2 < 0,$$

$$O_2: p_0 > 0,\ p^2 > 0, \quad O_5: p_0 > 0,\ p^2 = 0,$$

$$O_3: p_0 < 0,\ p^2 > 0, \quad O_6: p_0 < 0,\ p^2 = 0.$$

One can interpret the one-dimensional representation space corresponding to O_1 as a vacuum state, and subsequently the irreducible representation spaces related with orbits $O_2 - O_6$ as generated from the vacuum by means of the following free field operators[5]:

O_2, O_3—generalized free fields,
O_4—generalized free tachyon fields,
O_5, O_6—free massless fields.

The transformation properties of massless fields under the action of the Weyl group are well known.[7] If we perform the dilatation transformation

$$x \to x' = lx, \tag{1.4}$$

the components ϕ_A of a massless free field transform as follows[8]:

$$U(l)\phi_A(x)U^{-1}(l) = l^d \phi_A(lx), \tag{1.5}$$

where d denotes the *dimensionality parameter*. One obtains only the following two values of d:

$$d = 1 \text{ for bosons (helicity } \lambda = 0, \pm 1) \tag{1.5a}$$

$$d = \tfrac{3}{2} \text{ for fermions (helicity } \lambda = \pm\tfrac{1}{2}, \pm\tfrac{3}{2}).$$

In this paper we shall study the representations generated by the generalized free fields (orbits O_2 and O_3).[9] It is easy to see from the relation (1.3d) that

$$e^{i\alpha\hat{D}}\hat{P}_\mu\hat{P}^\mu e^{-i\alpha\hat{D}} = e^{-2\alpha}\hat{P}_\mu\hat{P}^\mu \tag{1.6}$$

and particularly, putting $l = e^\alpha$, one obtains for the eigenvalue m^2 of the Poincaré-invariant mass-squared operator $\hat{P}_\mu\hat{P}^\mu$:

$$m^2 \to m'^2 = m^2/l^2, \tag{1.7}$$

i.e., the only invariant support of the mass operator is the interval $(0, \infty)$. Choosing the vectors p_μ on the orbits O_2 and O_3 as follows:

$$O_2: p_\mu \sim (1, 0, 0, 0), \tag{1.8}$$

$$O_3: p_\mu \sim (-1, 0, 0, 0),$$

one gets for the orbits O_2 and O_3 the following little groups:

$$\mathcal{K}_2 \approx \mathcal{K}_3 \approx \overline{R}_3, \tag{1.9}$$

where $\bar{R}_3$ is a covering group of the three-dimensional rotation group R_3, i.e., one obtains the usual spin assignment ($s=0, \frac{1}{2}, 1, \ldots$). The irreducible representation of the Weyl group, characterized by an arbitrary complex dimensionality parameter d and a spin s, is generated from the vacuum by means of the components of generalized free field with spin:

$$O_2: \Phi^{(+)}_{d;\alpha}(x)=\int_0^\infty d\kappa^2(\kappa^2)^{(d/2-a_s)}\phi^{(+)}_\alpha(x;\kappa^2), \tag{1.10a}$$

$$O_3: \Phi^{(-)}_{d;\alpha}(x)=\int_0^\infty d\kappa^2(\kappa^2)^{(d/2-a_s)}\phi^{(+)}_\alpha(x;\kappa^2) \tag{1.10b}$$

where

$a_s=1$ for bosons ($s=0, 1, 2, \ldots$),

$a_s=\frac{5}{4}$ for fermions ($s=\frac{1}{2}, \frac{3}{2}, \ldots$),

and $\phi^{(+)}_\alpha$, $\phi^{(-)}_\alpha$ denote positive and negative frequency components of the so-called Licht field with spin.[10] The general rule for constructing any Licht field is to assume that in the usual massive theory the mass square is a continuous parameter. The field equations for Licht fields are the same as for usual massive free fields, but the commutator function is supplemented by a delta function $\delta(\kappa^2-\kappa'^2)$. Considering as an example the scalar case, one gets

$$(\Box-\kappa^2)\phi(x;\kappa^2)=0 \tag{1.11a}$$

and

$$[\phi(x;\kappa^2),\phi(x';\kappa'^2)]=i\Delta(x-x';\kappa^2)\delta(\kappa^2-\kappa'^2) \tag{1.11b}$$

The fields $\Phi^{(+)}_{d;\alpha}(x)$ and $\Phi^{(-)}_{d;\alpha}(x)$ are respectively the positive and negative frequency parts of real generalized free fields $\Phi_{d;\alpha}(x)\Phi^*_{d;\alpha}(x)$, where

$$\Phi_{d;\alpha}(x)=\Phi^{(+)}_{d;\alpha}(x)+\Phi^{(-)}_{d^*;\alpha}(x), \tag{1.11c}$$

$$\Phi_{d^*;\alpha}(x)=\Phi^{(+)}_{d^*;\alpha}(x)+\Phi^{(-)}_{d;\alpha}(x). \tag{1.11d}$$

If $d=d_1+id_2$, one gets (for $s=0$)

$$[\Phi_d(x),\Phi_d(0)]=[\Phi_{d^*}(x),\Phi_{d^*}(0)]$$
$$=i\int(\kappa^2)^{d_1-2}\Delta(x;\kappa^2)d\kappa^2. \tag{1.12}$$

The plan of our paper is the following: In Sec. 2 we introduce the representation spaces for orbits O_2 and O_3, and we shall relate them with the Hilbert space of "one-particle states"[6] of generalized free field.

In Sec. 3 we shall construct the generators of the Weyl group describing the transformation properties of the Licht fields. Because any real generalized free field can be written as the following linear form[11]

$$\Phi_{\xi;\alpha}(x)=\int d\kappa^2\,\xi(\kappa^2)\phi^{(+)}_\alpha(x;\kappa^2)+\text{h.c.}, \tag{1.13}$$

where $\xi(\kappa^2)\in S'(R_+)$ and $\xi(\kappa^2)=\xi_1(\kappa^2)+i\xi_2(\kappa^2)$. The generators for Licht fields are also the generators for generalized free fields. In the derivation of these generators we shall use the observation that the Lagrangian for free Licht fields[12] is invariant under the following transformation:

$$\phi(x;\kappa^2)\to\phi'(x;\kappa^2)=\phi(lx;\kappa^2/l^2). \tag{1.14}$$

Finally we derive in Sec. 3 the formulas (1.10). It should be mentioned that our results are consistent with the scaling properties of the Green's functions. Indeed, one can prove that the commutator (1.12) is proportional to the distribution $(x^2)^d_+$, and the causal propagator is proportional to $(x^2-i0)^d$. Similar scaling properties, implying dimensionality d, will have also the Green's functions of the generalized free field (1.11) with spin.

II. IRREDUCIBLE REPRESENTATION SPACES FOR THE WEYL GROUP (ORBITS O_2 AND O_3) AND GENERALIZED FREE FIELDS

Let us consider the realizations of the representations related with the orbits O_2 and O_3, on the Hilbert space $\mathcal{H}_s$ of square-integrable functions $\Psi(p_\mu,\lambda)$, where λ takes discrete values $\lambda=-s, -s+1, \ldots, +s$, and the scalar product is defined as follows:

$$<\Psi',\Psi>=\sum_{\lambda=-s}^{\lambda=s}\int_{\bar{V}_+\cup\bar{V}_-}\overline{\Psi'(p_\mu,\lambda)}\kappa(p^2)\Psi(p_\mu,\lambda)d^4p, \tag{2.1}$$

where $\kappa(p^2)=(p^2)^\alpha$ leads to the choice of the unitary representations for $d=\alpha+i\beta(-\infty<\beta<-\infty)$. Further we shall put $\kappa(p^2)=1$.

If supp $\Psi\in\bar{V}_+$, we can describe in $\mathcal{H}_s$ the representations related with orbit O_2; the choice supp $\Psi\in\bar{V}_-$ leads to the realizations of the representations associated with the orbit O_3.

In space $\mathcal{H}_s$ the algebra (1.3) of the generators of Weyl group can be represented by the differential operators (1.2). In particular we have

$$\hat{P}_\mu=p_\mu,\quad \hat{D}=p_\mu\frac{\partial}{\partial p_\mu}+2. \tag{2.2}$$

The eigenfunctions satisfying the equations

$$\hat{P}_\mu\Psi(p_\mu,\lambda;p^0_\mu)=p^0_\mu\Psi(p_\mu,\lambda;p^0_\mu), \tag{2.3a}$$

$$\hat{D}\Psi(p_\mu,\lambda;d)=d\Psi(p_\mu,\lambda;d) \tag{2.3b}$$

do not belong to $\mathcal{H}_s$. Indeed, one gets ($p^2=p_0^2-\mathbf{p}^2$)

$$\Psi(p_\mu,\lambda;p^0_\mu)=a\cdot\delta^4(p_\mu-p^0_\mu), \tag{2.4a}$$

$$\Psi(p_\mu,\lambda;d)=a'\cdot(p^2)^{(d/2)-1}, \tag{2.4b}$$

In order to get from the solutions (2.4a) the square-integrable functions it is sufficient to take the four-dimensional wave packet of eigenfunctions

$$\Psi(p_\mu,\lambda)=\int f(p^0_\mu)\Psi(p_\mu,\lambda;p^0_\mu)d^4p^0, \tag{2.5}$$

where $f\in L_2$. In the case of eigenfunctions (2.4b) *it is not sufficient* to take the wavepacket

$$\Psi(p^2,\lambda)=\int duF(u)\Psi(p_\mu,\lambda;u) \tag{2.6}$$

because of the infinite volume of the domain between two arbitrarily close hyperboloids p^2 and $p^2+\Delta p^2$. In order to get the element of $\mathcal{H}_s$ we should smear out (2.6) with the three-dimensional test function $\phi(\mathbf{p})$, belonging, for example, to $S(R_3)$.

The eigenfunction (2.4b) describes the sum of the following three irreducible representations[13]:

$$\Psi_\pm(p_\mu,\lambda;d)=a'\theta(\pm p_0)(p^2)^{(d/2)-1}_+, \tag{2.7a}$$

$$\Psi(p_\mu,\lambda;d)=a'(p^2)^{(d/2)-1}_-. \tag{2.7b}$$

Because for arbitrary $f(p^2)$

$$p_\mu\frac{\partial}{\partial p_\mu}f(p^2)=2p^2\frac{d}{dp^2}f(p^2) \tag{2.8}$$

and further, putting $(p^2)^\alpha_+=\theta(p^2)(p^2)^\alpha$, we get

$$p_\mu \frac{\partial}{\partial p_\mu}(p^2)_+^\alpha = 2p^2\delta(p^2)(p^2)^\alpha + 2\alpha(p^2)_+^\alpha. \quad (2.8')$$

The split into the part (2.7a) and (2.7b) (separation of orbit O_4 from O_2 and O_3) does not affect the eigenvalue equation only if Re $d>0$. On the other hand, the split into the positive and negative frequency parts (separation of orbits O_2 and O_3) implies in similar way that $Re\, d>\frac{1}{2}$. It is interesting to note, that the last condition is also required if we wish to obtain in the Kallen—Lehmann representation (1.12) the positive-definite spectral function.

The functions $\Psi(p_\mu,\lambda)\in\mathcal{H}_s$ can be interpreted as the wavefunction describing one-particle state[6] of generalized free field (1.13). For simplicity we shall consider here the scalar case ($s=0$). Let us smear out (1.13) with a test function $f\in S(R_4)$, and denote

$$\hat{\underline{\Psi}} = \int d^4x\, xf(x)\Phi_\xi(x). \quad (2.9)$$

Using the decomposition formula for the field (1.11)

$$\phi(x;\kappa^2) = \frac{1}{(2\pi)^{3/2}}\int \frac{d^3p}{(2\omega_\kappa)^{3/2}}\{e^{i(\mathbf{p}\mathbf{x}-\omega_\kappa t)}a(\mathbf{p};\kappa^2)+\text{H.C.}\}, \quad (2.10)$$

where

$$[a^+(\mathbf{p},\kappa^2),a(\mathbf{p}',\kappa'^2)] = \delta^3(\mathbf{p}-\mathbf{p}')\delta(\kappa^2-\kappa'^2), \quad (2.11)$$

and introducing the vacuum state as follows,

$$a(\mathbf{p},\kappa^2)|0\rangle = 0; \quad (2.12)$$

one can define the "one-particle" states with definite 4-momentum $p_\mu=(\mathbf{p},\omega_\kappa)$ by the formula

$$|\mathbf{p};\kappa^2\rangle = a^+(\mathbf{p},\kappa^2)|0\rangle \quad (2.13)$$

and one obtains

$$\langle \mathbf{p},\kappa^2|\mathbf{p}',\kappa'^2\rangle = \delta^4(p_\mu - p'_\mu)/2\omega_\kappa. \quad (2.14)$$

Writing

$$|\underline{\Psi}\rangle = \hat{\underline{\Psi}}|0\rangle = (2\pi)^{-3/2}\int_{V_+} d^4p\,\tilde{f}(\mathbf{p},p_0)\cdot(2p_0)^{1/2}\cdot\xi(p^2)|\mathbf{p};p^2\rangle, \quad (2.15)$$

where $\tilde{f}$ denotes the Fourier transform of f, one gets

$$\langle\underline{\Psi}|\Psi\rangle = \int_{V_+} d^4p\,\bar{\Psi}'(p)\Psi(p), \quad (2.16)$$

and

$$\Psi(p) = (2\pi)^{-3/2}\tilde{f}(\mathbf{p},p_0)\xi(p_0^2-\mathbf{p}^2). \quad (2.17)$$

The eigenfunction (2.4a) can be obtained *for any* choice of the continuous spectral function $\xi(\kappa^2)$ by the limit $\tilde{f}\to\delta^4(p-p_0)$. The eigenfunctions of the dilatation operator can be obtained by the choice $\xi(\kappa^2)=(\kappa^2)_+^{(d/2)-1}$ and $\tilde{f}=1$. Both limits obviously are leading outside of the class $S(R_4)$ of the test functions.

III. THE GENERATORS OF THE WEYL GROUP AND GENERALIZED FREE FIELDS WITH DEFINITE DIMENSIONALITY

The field equation (1.11a) and the commutation relations (1.11b) can be derived from the following Lagrangian[14]:

$$L_0 = \int d\kappa^2 \int d^4x\, L_0(x;\kappa^2), \quad (3.1)$$

where

$$L_0(x;\kappa^2) = -\tfrac{1}{2}[\partial_\mu\phi(x;\kappa^2)\partial^\mu\phi(x;\kappa^2) - \kappa^2\phi(x;\kappa^2)\phi(x;\kappa^2)]. \quad (3.2)$$

The Lagrangian (3.1) is obvious invariant with respect to the transformations of the Poincaré group (κ^2 is a c-number scalar). The invariance with respect to dilatations implies that the field operator $\phi(x;\kappa^2)$ transforms according to the transformation law (1.14).[15]

One can use the formalism, based on Noether theorem, which allows to express the generators of Weyl group as bilinear form in field operators $\phi(x;\kappa^2)$. The infinitesimal transformations of the Poincaré group are locally generated by the energy-momentum density $T^0{}_{\mu\nu}(x)$, which can be written as follows:

$$T^0{}_{\mu\nu}(x) = \int d\kappa^2 T^0{}_{\mu\nu}(x;\kappa^2), \quad (3.3a)$$

where

$$T^0{}_{\mu\nu}(x;\kappa^2) = \frac{\partial L_0}{\partial\phi_{,\mu}(x;\kappa^2)}\partial_\nu\phi(x;\kappa^2) + g_{\mu\nu}L_0(x;\kappa^2). \quad (3.3b)$$

Introducing

$$\hat{P}^0_\mu = \int d\kappa^2 \hat{P}^0_\mu(\kappa^2), \quad (3.4a)$$

where

$$\hat{P}^0_\mu(\kappa^2) = \int d^3x\; T^0_{0\mu}(x;\kappa^2), \quad (3.4b)$$

we have the relations

$$i^{-1}[\hat{P}^0_\mu(\kappa^2),\phi(x;\kappa^2)] = \partial_\mu\phi(x;\kappa^2)\delta(\kappa^2-\kappa'^2), \quad (3.5a)$$

$$i^{-1}[\hat{P}^0_\mu,\phi(x;\kappa^2)] = \partial_\mu\phi(x;\kappa^2). \quad (3.5b)$$

Let us introduce now the infinitesimal transformations (1.4) and (1.7),

$$x'_\mu = (1+\delta l)x_\mu,\quad \delta x_\mu = \delta l\cdot x_\mu, \quad (3.6)$$
$$\kappa'^2 = (1-2\delta l)\kappa^2,\quad \delta\kappa^2 = -2\delta l\cdot\kappa^2.$$

The transformation law (1.14) can be written in its infinitesimal form as follows:

$$\delta\phi(x;\kappa^2) = \delta l\left(x_\nu\frac{\partial}{\partial x_\nu}\phi(x;\kappa^2) - 2\kappa^2\frac{\partial}{\partial\kappa^2}\phi(x;\kappa^2)\right). \quad (3.7)$$

The invariance of the Lagrangian (3.1) implies the independence of the hypersurface σ of the following integral

$$\delta F(\sigma) = \int_\sigma d\sigma^\mu\int d\kappa^2\left(\frac{\partial L_0}{\phi_{,\mu}(x;\kappa^2)}\cdot\delta\phi(x;\kappa^2) + L_0(x;\kappa^2)\delta x_\mu\right) = \int_\sigma d\sigma^\mu\int d\kappa^2\left[T_{\mu\nu}(x;\kappa^2)\delta x^\nu + \frac{\delta L_0}{\partial\phi_{,\mu}(x;\kappa^2)}\left(\frac{\partial}{\partial\kappa^2}\phi(x;\kappa^2)\right)\right]\delta\kappa^2$$
$$= \int_\sigma d\sigma^\mu\{S^{(1)}_\mu(x) + S^{(2)}_\mu(x)\}\delta l \quad (3.8)$$

where

$$S^{(1)}_\mu(x) = \int d\kappa^2 T_{\mu\nu}(x;\kappa^2)x^\nu = T_{\mu\nu}(x)\cdot x^\nu, \quad (3.9a)$$

$$S^{(2)}_\mu(x) = -2\int d\kappa^2\cdot\kappa^2\cdot\frac{\partial L_0}{\delta\phi_{,\mu}(x;\kappa^2)}\cdot\frac{\partial}{\partial\kappa^2}\phi(x;\kappa^2). \quad (3.9b)$$

The formulas (3.8—9) are valid for any Lagrangian density, invariant under the infinitesimal transformation (3.6). Choosing in particular the form (3.2) and putting $d\sigma_\mu = (d^3x,0,0,0,)$, one gets

$$\delta F(t) = \hat{D}\delta l = \hat{D}_{\text{geom}}(t)\,\delta l + \hat{D}_{\text{mass}}(t)\,\delta l, \quad (3.10)$$

where $\delta F(t_1)=\delta F(t_2)$, the "geometric" part of the generator $\hat{D}$ is expressed by means of the formula

$$\hat{D}_{\text{geom}}(t)=\int T_{0m}(x)x^m d^3x, \tag{3.11a}$$

and

$$\hat{D}_{\text{mass}}(t)=-2\int d^3x\int d\kappa^2\cdot\kappa^2\partial_0\phi(x;\kappa^2)\cdot\frac{\partial}{\partial\kappa^2}\phi(x;\kappa^2). \tag{3.11b}$$

It is easy to check that

$$\delta\phi(x;\kappa^2)=\ i^{-1}[\hat{D},\phi(x;\kappa^2)]\delta l \tag{3.12}$$

which follows from the relations

$$[\phi(\mathbf{x},t;\kappa^2),\phi(\mathbf{x}',t;\kappa'^2)]=0, \tag{3.12'a}$$

$$[\partial_0\phi(\mathbf{x},t;\kappa^2),\phi(\mathbf{x}',t;\kappa'^2)]=i\delta^3(\mathbf{x}-\mathbf{x}')\delta(\kappa^2-\kappa'^2), \tag{3.12'b}$$

and[16]

$$[\phi(\mathbf{x},t;\kappa^2),(\partial/\partial\kappa^2)\phi(\mathbf{x}',t;\kappa'^2)]=0. \tag{3.12'c}$$

The relations (3.5b) and (3.12), integrated with the spectral function $\xi(\kappa^2)$, lead to the transformation properties of the generalized free fields. One gets from (3.12) that[17]

$$i^{-1}[\hat{P}_\mu,\Phi_\xi(x)]=\partial_\mu\Phi_\xi(x), \tag{3.13a}$$

$$i^{-1}[\hat{D},\Phi_\xi^{(\pm)}(x)]=(x_\mu\partial^\mu+\hat{d})\Phi_\xi^{(\pm)}(x), \tag{3.13b}$$

where

$$\hat{d}\Phi_\xi^{(\pm)}(x)=\frac{1}{i}[\hat{D}_{\text{mass}}(t)\,\Phi_\xi^{(\pm)}(x)]$$

$$=-2\int_0^\infty d\kappa^2\xi(\kappa^2)[\kappa^2\frac{\partial}{\partial\kappa^2}\phi^{(\pm)}(x;\kappa^2)]. \tag{3.14}$$

Let us write

$$\xi(\kappa^2)\cdot\kappa^2\frac{\partial}{\partial\kappa^2}\phi^{(\pm)}(x;\kappa^2)=\frac{\partial}{\partial\kappa^2}[\xi(\kappa^2)\cdot\kappa^2\phi^{(\pm)}(x;\kappa^2)]$$

$$-\phi^{(\pm)}(x;\kappa^2)\cdot(d/d\kappa^2)[\xi(\kappa^2)\cdot\kappa^2]. \tag{3.15}$$

One obtains

$$\hat{d}_\xi^{(\pm)}(x)=\Phi_{\xi'}^{(\pm)}(x)+\text{boundary term}, \tag{3.16}$$

where

$$\xi'(\kappa^2)=2\frac{d}{d\kappa^2}\kappa^2\xi(\kappa^2) \tag{3.17}$$

and[18]

$$\text{boundary term}=\{\xi(\kappa^2)\cdot\kappa^2\cdot\phi^{(\pm)}(x;\kappa^2)\}\Big|_a^b, \tag{3.18}$$

where supp $\xi=[a,b]$.

The generalized free field with definite dimensionality parameter is described by the following equation for the spectral function:

$$\xi'(\kappa^2)=d\cdot\xi(\kappa^2)=2d\cdot\left(\frac{d}{d\kappa^2}\kappa^2\xi(\kappa^2)\right) \tag{3.19}$$

and has the solution $\xi(\kappa^2)=(\kappa^2)^{(d/2)-1}$. We get, therefore, the result (1.10a—b) by considering, respectively, the positive (orbit O_2) and the negative (orbit O_3) frequencies. In order however, to write Eq. (3.19), one has to show that the boundary term vanishes, i.e.,

$$\lim_{\Lambda^2\to\infty}(\Lambda^2)^{d/2}\phi^{(\pm)}(x;\Lambda^2)=0. \tag{3.20}$$

The meaning and validity of the relation (3.20) is presented in the Appendix.

The generalization of our results to the case with spin does not present any particular difficulties. Let us consider for example the vector case ($s=1$). The generalized free field with spin one can be written as follows[19]:

$$\Phi_{\xi;\mu}(x)=\int\xi(\kappa^2)\phi_\mu(x;\kappa^2)d\kappa^2, \tag{3.21}$$

where

$$[(\Box-\kappa^2)g_{\mu\nu}-\partial_\mu\partial_\nu]\phi^\mu(x;\kappa^2)=0, \tag{3.22a}$$

$$[\phi_\mu(x;\kappa^2),\phi_\nu(x';\kappa^2)]=i(g_{\mu\nu}-\partial_\mu\partial_\nu/\kappa^2)\Delta(x-x';\kappa^2)\delta(\kappa^2-\kappa'^2), \tag{3.22b}$$

and, consequently, it satisfies the Lorentz condition

$$\partial^\mu\Phi_\mu(x)=0. \tag{3.23}$$

We have

$$[\Phi_{\xi;\mu}(x),\Phi_{\xi;\mu}(x')]=g_{\mu\nu}\int\xi(\kappa^2)\Delta(x-x';\kappa^2)d\kappa^2 \tag{3.24}$$

$$-\partial_\mu\partial_\nu\int[\xi(\kappa^2)/\kappa^2]\Delta(x-x';\kappa^2)d\kappa^2$$

$$=(g_{\mu\nu}-\partial_\mu\partial_\nu/\Box)\int\xi(\kappa^2)\Delta(x-x';\kappa^2)d\kappa^2.$$

One can check, by writing the Lagrangian leading to the relations (3.22a—b) that

$$U_l\phi_\mu(x;\kappa^2)U_l^{-1}=\phi_\mu(lx;\kappa^2/l^2) \tag{3.25}$$

and, choosing $\xi(\kappa^2)=(\kappa^2)^{(d/2)-1}$, one obtains

$$U_l\Phi_{\xi;\mu}(x)U_l^{-1}=l^d\Phi_{\xi;\mu}(lx). \tag{3.26}$$

The transformation properties of generalized free fields with higher integer spin are identical to the ones given by the formulas (3.25) and (3.26), provided that the corresponding Licht field $\phi_{\mu_1\ldots\mu_s}(x;\kappa^2)$ satisfies the subsidiary conditions[20]

$$\partial^\mu\phi_{\mu\mu_3\ldots\mu_s}(x;\kappa^2)=0,$$

$$\phi^\mu{}_{\mu\mu_3\ldots\mu_s}(x;\kappa^2)=0. \tag{3.27}$$

Let us consider now briefly the half-integer spins. The Lagrangian for Licht field with spin one-half looks as follows:

$$L_0(x)=\int d\kappa^2[\bar{\Psi}(x;\kappa^2)(\gamma_\mu\partial^\mu-\kappa)\Psi(x;\kappa^2)]. \tag{3.28}$$

The invariance with respect to the dilatations implies[21]

$$U_l\Psi(x;\kappa^2)U_l^{-1}=l^{1/2}\Psi(lx;\kappa^2/l^2). \tag{3.29}$$

Introducing

$$\Psi_d(x)=\int(\kappa^2)_+^{\left(\frac{1}{2}d-\frac{5}{4}\right)}\Psi(x;\kappa^2)d\kappa^2, \tag{3.30}$$

one gets

$$U_l\Psi_d(x)U_l^{-1}=l^d\Psi_d(lx). \tag{3.31}$$

The formula (3.29) characterizes the transformation properties of *any Licht field with half-integer spin* provided that the corresponding Licht field satisfies all the required subsidiary conditions.[22]

In such a case the relation between the dimensionality parameter and the power behavior of the Kallen-Lehmann spectral function [formulas (3.30) and (3.31)] also remains valid.

ACKNOWLEDGMENTS

One of the authors (J. L.) wishes to thank Professor J. A. Swieca for his hospitality at the Institute of Physics of PUC, where a portion of this work has been carried out.

APPENDIX

Because the field operator $\phi(x;\kappa^2)$ is an operator-valued distribution with respect to κ^2, we shall consider here the following operator:

$$\phi_\alpha(x;\kappa^2)=\int d\kappa'^2\alpha(\kappa^2-\kappa'^2)(\kappa'^2)^d\phi(x;\kappa^2), \tag{A1}$$

where $\alpha\in D$, $\alpha\geq 0$, supp $\alpha\in[-1,+1]$, and

$$\int d\kappa^2\alpha(\kappa^2)=1. \tag{A2}$$

The field operator $\phi_\alpha(x;\kappa_0^2)$ describes a generalized free field with the positive-definite spectral function (Re $d>-\frac{1}{2}$)

$$[\phi_\alpha(x;\kappa_0^2),\phi_\alpha(x';\kappa_0^2)]=i\int d\kappa'^2\alpha^2(\kappa_0^2-\kappa'^2)\times(\kappa'^2)^{2\mathrm{Re}\,d}\Delta(x-x';\kappa'^2). \tag{A3}$$

In order to show that the operator $\phi_\alpha(x;\kappa_0^2)$ vanishes strongly as an operator-valued distribution in S, i.e., that for any choice of α, satisfying the conditions specified above, we have

$$\lim_{\kappa_0^2\to\infty}||(f,\phi_\alpha(\kappa_0^2)||\to 0 \quad \text{for every } f\in S, \tag{A4}$$

we should prove that

$$\lim_{\kappa_0^2\to\infty}||(f,\phi_\alpha(\kappa_0^2))|P_N>||=0, \tag{A5}$$

where the state $|P_N>$ is created by the polynomials of the field $\phi(x;\kappa^2)$, smeared out with respect to *both* variables $\mathbf{x}$ and κ^2:

$$|P_N\rangle=\prod_{i=1}(f_i\otimes X_i,\phi)|0\rangle, \tag{A6}$$

where $f_i\in S(R_3)$, $X_i\in S(R_1)$, and

$$(f_i\otimes X_i,\phi)=\int d\kappa^2X_i(\kappa^2)\int d^3\mathbf{x}f_i(\mathbf{x})\phi(\mathbf{x},x_0;\kappa^2). \tag{A7}$$

Proof: In order to show (A5), one should observe that the norm (A5) is a product of the following three types of factors:

$$\langle 0|(f_i\otimes X_i,\phi)(f_j\otimes X_j,\phi)|0\rangle, \tag{A7a}$$

$$\langle 0|(f,\phi_\alpha(\kappa_0^2))(f_j\otimes X_j,\phi)|0\rangle, \tag{A7b}$$

$$\langle 0|(f,\phi_\alpha(\kappa_0^2))(f,\phi_\alpha(\kappa_0^2))|0\rangle. \tag{A7c}$$

The first factor is bounded, and the second one and the third one vanishes in the limit $\kappa_0^2\to\infty$. In order to show this statement for the term (A7b) let us observe that

$$\langle 0|(f,\phi_\alpha(\kappa_0^2)),(f_j\otimes X_j,\phi)|0\rangle = (2\pi)^{-3}\int d\kappa^2\alpha(\kappa_0^2-\kappa^2)\cdot x_j(\kappa^2)(\kappa^2)^{\mathrm{Re}\,d}\times\int\frac{d^3p}{2\omega_\kappa}\tilde f_j(\mathbf{p},\omega_\kappa), \tag{A8}$$

where $\omega_\kappa=(\mathbf{p}^2+\kappa^2)^{1/2}$. Because

$$\int\frac{d^3p}{2\omega_\kappa}\tilde f(\mathbf{p})\cdot\tilde f_j(\mathbf{B},\omega_\kappa) \leq(2\kappa)^{-1}\int d^3\mathbf{p}|\tilde f(\mathbf{p})|\,|\tilde f_j(p,\omega_\kappa)| \tag{A9}$$

and

$$|\tilde f_j(\mathbf{p},\omega_\kappa)|\leq A/(1+|\mathbf{p}|)^n(1+\omega_\kappa^2)^m\leq A/(1+|\mathbf{p}|)^n(1+\kappa^2)^m, \tag{A10}$$

one can write

$$\int\frac{d^3p}{2\omega_\kappa}\tilde f(\mathbf{p})\cdot\tilde f_j(\mathbf{p},\omega_\kappa)\leq\frac{A'}{\kappa(1+\kappa^2)^m} \tag{A11}$$

and, assuming $m>2\mathrm{Re}\,d-\frac{1}{2}$,

$$\langle 0|(f,\phi_\alpha(\kappa_0^2)),(f_j\otimes X_j,\phi)|0\rangle \leq\frac{A'}{(2\pi)^3}d\kappa^2\,\frac{|\alpha(\kappa_0^2-\kappa^2)||X_j(\kappa^2)^{\mathrm{Re}\,d}}{\kappa(1+\kappa^2)^m} \leq A''\int_{\kappa_0^2-1/2}^{\kappa_0^2+1/2}d\kappa^2\,\frac{(\kappa^2)^{2\mathrm{Re}\,d-1/2}}{(1+\kappa^2)^m}\xrightarrow[\kappa_0^2\to\infty]{}0. \tag{A12}$$

The proof that the expression (A7c) goes to zero in the limit $\kappa^2\to\infty$ can be given in similar way, under the assumption that $m>2\mathrm{Re}\,d-\frac{1}{4}$.

It is easy to show, using the decomposition of the $2N$-point VEV of the generalized free field into the finite sum of product of two-point functions, that every term in the decomposition of (A5) into the products of (A7) must contain either *once the term* (A7c) *or twice the term* (A7b). This property implies the relation (A5). QED

We see, therefore, that the eigenvalue equation [see (3.16) and (3.19)]

$$\hat d\Phi_\xi^{(\pm)}(x)=d\Phi_\xi^{(\pm)}(x) \tag{A13}$$

is valid strongly as an operator-valued equation in $S'(R_4)$, i.e.,

$$||(f,\hat d\Phi_\xi^{(\pm)}-d\Phi_\xi^{(\pm)})||=0, \tag{A14}$$

for any $f\in S$.

*On leave from the Institute of Theoretical Physics, University of Wroclaw, Wroclaw, Poland.

[1]U. Ottoson, Ark. Fys. **33**, 523 (1967).
[2]H. E. Moses, Ann. Phys. (N.Y.) **52**, 444 (1969).
[3]J. Mickelsson and J. Niederle, Czechoslovak Academy of Science, preprint, Prague (May 1970).
[4]G. W. Mackey, Bull. Am. Math. Soc. **69**, 628 (1963).
[5]For completeness one should make a remark that the representations defined on the orbits O_2, O_3 can be defined by means of *massless two-particle states*. Similarly one can relate the representations on the orbit O_4 with the superposition of massless particle and antiparticle states. This way of interpreting the representations of the Weyl group follows from the possibility of describing "one-particle states"[6] of a generalized free field in terms of conventional two-particle states [see for example A. Streit, Helv. Phys. Acta **39**, 65 (1965); J. Lukierski, Nuovo Cimento **60A**, 353 (1969); R. A. Brandt and O. W. Greenberg, J. Math. Phys. **10**, 1168 (1969); A. L. Licht, Max Planck Institute, preprint (1970)].
[6]We shall use the expression "one-particle state" for a state created from the vacuum by any linear functional of a generalized free field. We should, however remark, that these states do not posess all the characteristics usually attributed to the particle states.
[7]See, for example, L. Gross, J. Math. Phys. **5**, 687 (1964); G. Mack and I. Todorov, J. Math. Phys. **10**, 2078 (1969).
[8]It should be stressed that in covariant theory of massless particles with helicity $\lambda\geq 1$ one should consider a field-theoretic formulation using, for example, generalized Feymman or Landau gauges. In these gauges the dimensionality of the nonphysical degrees of freedom is equal to the dimensionality of the physical components, and the relation (1.5) can

be written for a covariant field operator ϕ_A, containing always some redundant nonphysical degrees of freedom (with lower helicities).

[9]The discussion can be easily extended to the representations related to O_4 by assuming that the mass spectrum of the generalized free field is purely space-like. Such a field can be called a generalized free tachyon field.

[10]A. L. Licht, Ann. Phys. (N.Y.) **34,** 165 (1965) (scalar case).

[11]It should be mentioned that the formula (1.13) is more general than the one occuring in Ref. 10. We obtain the conventional decomposition formula $\Phi_\xi(x) = \int d\kappa^2 \xi(\kappa^2)\phi(x;\kappa^2)$, if $\xi(\kappa^2)$ is real.

[12]Such a free Lagrangian was first used by W. Thirring [Phys. Rev. **126,** 1209 (1962)] in the field-theoretic description of the Zachariasen model.

[13]The Lorentz-invariant distributions $(p^2)^a_+$ are defined for example by Güttinger [see W. Güttinger, Fortschr. Phys. **14,** 485 (1966)].

[14]The case with spin, particularly the generalized free spinor field, will be considered at the end of Sec. 3.

[15]We see, therefore, that scalar Licht fields are also true scalars under the transformations of the dilatation group, i.e., in the formula $U_l\phi(x;\kappa^2)U_l^{-1} = l^d\phi[lx;(\kappa^2/l^2)]$, the dimensionality parameter d is equal to zero.

[16]The relation (3.12c) follows formally from the relation (3.12a) by differentiation with respect to κ'^2. The field operator $\phi_{(1)}(x;\kappa^2) = (d/d\kappa^2)\phi(x;\kappa^2)$ satisfies the following equation: $(\Box - \kappa^2)\phi_{(1)}(x;\kappa^2) = \phi(x;\kappa^2)$, which follows from differentiation with respect to κ^2 of Eq. (1.11a). We obtain $(\Box - \kappa^2)^2\phi_{(1)}(x;\kappa^2) = 0$.

[17]We write the relation (3.13b) for definite frequency part in order to incorporate in our discussion the case of representations of the Weyl group, with complex dimensionality parameter. Because the relation $\Phi^{(\pm)}(p) = \theta(\pm p_0)\Phi(p)$ is linear, one gets (3.13b) by performing the Fourier transform of (3.12) and multiplying by $\theta(\pm p_0)$.

[18]One obtains the formula (3.18) only in the case when the spectral function $\xi(\kappa^2)$ is continuous in the interval $[a, b]$. It will appear further that this is a case of interest for our considerations.

[19]We shall consider for simplicity here only real spectral functions $\xi(\kappa^2)$.

[20]Licht fields with higher integer spin have been discussed in other context by J. Lukierski, Fortschr. Phys. 21, 85 (1973).

[21]It should be mentioned that there is a possibility of using in (3.26) the mass instead of the mass square as the integration variable. In such a case the dimensionality of the Licht spinor field $\psi(x;\kappa)$ becomes $d = 1$ instead of $d = (1/2)$. If the integration in the free Lagrangian is performed over the variable κ, the dimensionality of a Licht field in comparison with the massless stable case [see (1.5a)] is smaller by (1/2); if the integration is over κ^2, it diminishes by 1. This can be deduced from the invariance properties of commutation relations containing respectively factors $\delta(\kappa - \kappa')$ or $\delta(\kappa^2 - \kappa'^2)$.

[22]The subsidiary conditions for the symmetric massive spinor–tensor field $\psi_{a;\mu_1\ldots\mu_s}$ have been given by Rarita and Schwinger [Phys. Rev. **60,** 61 (1941)]. The case of general spinor-tensor field $\psi_{a;\mu_1\ldots\mu_l[\mu_{l+1}\nu_{l+1}]\ldots[\mu_s,\nu_s]}$ have been discussed recently by Cukierda [see T. Cukierda, Nuovo Cimento Lett. **4,** 353 (1970)].

On algebraically irreducible representations of the Lie algebra $sl(2)$

D. Arnal and G. Pinczon

Universite de Dijon, Department de Mathematiques, 6, bd Gabriel, 21-Dijon, France
(Received 8 May 1972)

An algebraic study of the irreducible representations of the complex Lie algebra $sl(2)$ is presented in this article. This study generalizes a former series of works of W. Miller. Though the list is not complete, it gives hints as to the construction of a very vast family of representations.

INTRODUCTION

Though it is never stressed very clearly, Lie algebra representations (which are not necessarily integrable to representations of a corresponding Lie group) play an important role in mathematical physics. Let us mention some examples from physics.

(1) In potential scattering theories, an important contribution was made by T. Regge. His famous idea stressed the importance of the complex J. Plane and of the corresponding particle trajectories. If one wants to glue together infinitely many spins in an irreducible object, or in any case if one wants to have complex J-values, it is evident from elementary considerations that no group representation of $SU(2)$ can make it. One is then brought to the study of local (i.e., Lie algebra) representations of $SU(2)$.

(2) In elementary particle physics, we distinguish between external and internal symmetries: the external (i.e., Poincaré) symmetry group translates the postulate of the special theory of relativity, while the internal [e.g. $SU(3)$] "symmetry" reflects, e.g., the behavior of particles under strong interactions, their classification and quantum numbers, etc.

The external symmetry is traditionally represented in second quantized field theory by unitary representations of the Poincaré group, the group action being physically clear.

In very high energies, which correspond to very small distances, there is no physical reason to believe that the external symmetry acts in a group manner. In this case, therefore, even the external symmetry might very well act in a Lie algebra manner. For internal symmetries, we have no reason whatsoever to prefer the global to the local aspect, the "internal space" being fictitious.

Moreover, it is very well known that if one wants to have mass spectrum in an irreducible object, one cannot take a unitary representation of a Lie group containing the Poincaré group, but rather one has to take a representation of a Lie algebra containing the Poincaré Lie algebra, which can even be chosen so as to be integrable to the Poincaré part.

These examples, as well as many others, show the importance of the study of Lie algebra representations, both mathematically and physically.

In what follows, we shall give a detailed study of one of the simplest cases existing, that of *algebraically* irreducible representations of the simplest complex simple Lie algebra, $sl(2)$.

From this point of view, this work is a direct continuation of the elegant works of W. Miller. We divide our article in the following manner.

In Sec. 2, we give an equivalent formulation to the Miller hypothesis, and we then construct and classify the set of classes of irreducible representations satisfying the most general spectral hypothesis for the representative of the elements of the Lie algebra which strictly contains Miller representations and which is invariant under the action of the automorphisms of $sl(2)$.

In Sec. 3, we expound a detailed algebraic study of the enveloping algebra of $sl(2)$.

We utilize these results in Sec. 4 to study the action of the enveloping algebra automorphisms on the representations that we obtained in Sec. 2. By this generalization of the process of Sec. 2, we obtain a new series of representations.

In Sec. 5, we construct other representations that cannot be obtained by the previous technique.

1. NOTATION AND PRELIMINARY LEMMAS

Denote by (Y, F, G) a Weyl basis of $sl(2)$, Y being the regular element and F and G corresponding nilpotent elements,[1,2] with the commutation relations

$$[Y,F]=F,\quad [Y,G]=-G,\quad [F,G]=2Y.$$

$\mathcal{U}$ is the universal enveloping algebra of $sl(2)$ and Z the center of $\mathcal{U}$, which is written as

$$Z=\mathrm{C}[Q],\quad \text{where } Q=GF+Y+Y^2=FG-Y+Y^2$$
(Casimir element).

By a representation of $sl(2)$ we mean a linear representation in a complex vectorial space not reduced to $\{0\}$. By the universal property, every representation of $sl(2)$ is a representation of the associative algebra $\mathcal{U}$, and vice versa.

Definition 1.1: Let π and π' be representations of $sl(2)$ on V, V' respectively, π and π' are equivalent if there exists an isomorphism U of V and V' such that

$$U\circ\pi(X)=\pi'(X)\circ U \quad \text{for all } X\in sl(2).$$

Definition 1.2: A representation is (algebraically) irreducible if the only invariant subspaces of the representation space are $\{0\}$ and the space itself.

From the Poincaré—Birkhoff—Witt theorem, the representation space of an irreducible representation is (necessarily) of at most countable dimension.

Schur's well-known lemma for finite dimension is generalized to infinite dimension.[3,4]

Lemma 1.1: The commutator of an (algebraically) irreducible representation is reduced to scalars.

$\mathcal{U}$ being a noetherian algebra for which the only elements having inverses are the scalars,[2] one has the following lemma:

Lemma 1.2: Given a nonscalar element X of $\mathcal{U}$ and an element λ of C, there exists (at least) one irreducible representation π of $sl(2)$ such that $\pi(X)$ has the eigenvalue λ.

One easily obtains by induction the following formulas in $\mathcal{U}$:

Lemma 1.3: Let $P\in Z(X)$; one has

$$P(Y)F^n = F^nP(Y+n),$$
$$F^nP(Y) = P(Y-n)F^n,$$
$$P(Y)G^n = G^nP(Y-n),$$
$$G^nP(Y) = P(Y+n)G^n.$$

Let $\alpha\in Z[X]$ be the polynomial $\alpha = Q - X - X^2$. One has $GF = \alpha(Y)$, $FG = \alpha(Y-1)$, from which one deduces:

$$n\geq p:\quad G^nF^p = \alpha(Y+n-1)\cdots\alpha(Y+n-p)G^{n-p} = G^{n-p}\alpha(Y+p-1)\cdots\alpha(Y+1)\alpha(Y);$$
$$n\leq p:\quad G^nF^p = F^{p-n}\alpha(Y+p-n)\cdots\alpha(Y+p-1) = \alpha(Y)\cdots\alpha(Y+n-1)F^{p-n};$$
$$n\geq p:\quad F^pG^n = G^{n-p}\alpha(Y-n+p-1)\cdots\alpha(Y-n) = \alpha(Y-1)\cdots\alpha(Y-p)G^{n-p};$$
$$n\leq p:\quad F^pG^n = \alpha(Y-p)\cdots\alpha(Y-p+n)F^{p-n} = F^{p-n}\alpha(Y-n)\cdots\alpha(Y-1).$$

2. REPRESENTATIONS OF $sl(2)$ FOR WHICH THE REPRESENTATIVE OF ONE ELEMENT (AT LEAST) HAS AN EIGENVALUE

In this section (Y,F,G) is a fixed Weyl basis of $sl(2)$; $\pi_{sl(2)}$ designates the set of classes of representations which satisfy the property quoted in the title.

A. The Miller representations

In Refs. 5 and 6, Miller constructs (in particular) the irreducible representations π of $sl(2)$ which verify the following properties:

(a) The spectrum Λ of $\pi(Y)$ is countable, of the form $\Lambda = \{\lambda_s = \lambda_0 + s, s\in S\subset \mathbf{Z}\}$ ($\mathbf{Z}$ = integers).

(b) Each eigenspace V_{λ_s} is one-dimensional, and

$$V = \bigoplus_{s\in S} V_{\lambda_s}.$$

His result is (cf. [5])

Theorem 2.1: Every irreducible representation of $sl(2)$ satisfying conditions (a) and (b) is isomorphic to a representation from the following list:

(1) The representations $D(l,m_0)$, where l and m_0 are two complex numbers such that l is admissible (i.e., either $l+1/2 = \rho e^{i\varphi}$, $\rho>0$, $0\leq\varphi<\pi$ or $l=-1/2$), $0\leq \mathrm{Re}\, m_0 < 1$, and neither m_0+l nor m_0-l are integers. The spectrum

$$\Lambda = \{m_0+n,\ n\in\mathbf{Z}\}.$$

(2) The representations $\downarrow\vec{l}$ defined for all admissible l,

$$\Lambda = \{-l-1-n, \in \mathrm{N}\}\quad (\mathrm{N} = \text{integers}\geq 0).$$

(3) The representations $\downarrow\dot{l}$ defined for all admissible l different from $(h-1)/2 (\forall h\in \mathrm{N})$, with

$$\Lambda = \{l-n, n\in\mathrm{N}\}.$$

(4) The representations $\uparrow\vec{l}$ defined for all admissible l different from $(h-1)/2 (\forall h\in \mathrm{N})$, with

$$\Lambda = \{-l+n,\ n\in\mathrm{N}\}.$$

(5) The representation $\uparrow\dot{l}$ defined for all admissible l, with

$$\Lambda = \{l+1+n, n\in\mathrm{N}\}.$$

(6) The representations $D(2h)$ defined for all h, $2h\in\mathrm{N}$, with

$$\Lambda = \{h-n,\ n\in\mathrm{N},\ 0\leq n\leq 2h\}.$$

Let us give an equivalent formulation of the hypotheses of Miller.

Proposition 2.1: An irreducible representation π of $sl(2)$ satisfies hypotheses (a) and (b) if and only if $\pi(Y)$ has an eigenvalue.

Proof: Let V be the representation space, φ a nonnull vector such that $\pi(Y)\varphi = \lambda\varphi$. By Lemma 1.1, $\pi(Q) = qI$, $q\in\mathrm{C}$, and we have (using Lemma 1.3):

$$\pi(YF^s)\varphi = (\lambda+s)\pi(F^s)\varphi,\quad \pi(YG^s)\varphi = (\lambda-s)\pi(G^s)\varphi,$$
$$\pi(GF^s)\varphi = [q-(\lambda+s-1)(\lambda+s)]\pi(F^{s-1})\varphi,$$
$$\pi(FG^s)\varphi = [q-(\lambda-s+1)(\lambda-1)]\pi(G^{s-1})\varphi$$

for all integers s, $s\geq 1$. Thus, we have the invariance of the subspace generated by $\{\varphi, \pi(F^s)\varphi, \pi(G^s)\varphi,\ s$ an integer $\geq 1\}$. QED

B. Construction and classification of the irreducible representations of $sl(2)$ for which the representative of a regular element (at least) has an eigenvalue

We shall denote by

$\mathcal{A}' = \mathrm{Aut}[sl(2)]$, the group of the automorphisms of $sl(2)$;

i_μ, Ω_μ, the automorphisms of $sl(2)$ defined by

$$i_\mu(Y) = Y,\ i_\mu(F) = \mu F, i_\mu(G) = 1/\mu G,$$
$$\Omega_\mu(Y) = -Y,\ \Omega_\mu(F) = \mu G, \Omega_\mu(G) = 1/\mu F,$$

where $\mu\in \mathrm{C} - \{0\}$;

π_r, the set of the classes of infinite-dimensional irreducible representations of $sl(2)$ for which the representative of a regular element (at least) has an eigenvalue;

$\mathcal{D}$, the subset of π_r made up of the representative of type $D(l,m_0)$;

$\mathcal{L}^+$ (resp. $\mathcal{L}^-$), the subset of π_r made up of the representations of type $\uparrow\vec{l}$ (resp. $\downarrow\vec{l}$).

We designate by the same letter a class and one of its representatives.

Let π be a member of π_r; we denote by $O(\pi)$ the orbit

of π under the action of $\mathcal{A}'$.

Let us first show that the study of π_r is equivalent to that of the action of $\mathcal{A}'$ on $\mathcal{D}$, $\mathcal{L}^+$, and $\mathcal{L}^-$.

Proposition 2.2: π belongs to π_r if and only if there exists π' belonging to $\mathcal{D}$, $\mathcal{L}^+$, or $\mathcal{L}^-$ and A belonging to $\mathcal{A}'$ such that $\pi = \pi' \circ A$.

Proof: Let X be the regular element such that $\pi(X)$ has an eigenvalue, A_1 an automorphism of $sl(2)$ such that $A_1 Y \in \mathbf{C}X$ (cf. Ref. 2). The representation $\pi_1 = \pi \circ A_1$ satisfies the hypotheses of Proposition 2.1.

If π_1 is of type $\downarrow l$, then $\pi_2 = \pi_1 \circ \Omega_1 = \pi \circ A_1 \circ \Omega_1$ is $\downarrow l$; in this case one takes $A = A_1 \circ \Omega$, and $\pi' = \pi_2$. If not, π_1 and A_1 are appropriate. QED

A simple calculation proves the following propositions.

Propositions 2.3: When π belongs to $\mathcal{D}$ and X belongs to $sl(2)$, $\pi(X)$ has an eigenvalue if and only if $X = \lambda Y$, $\lambda \in \mathbf{C}$.

Proposition 2.4: When π belongs to $\mathcal{L}^+$ or $\mathcal{L}^-$ and X belongs to $sl(2)$, $\pi(X)$ has an eigenvalue if and only if

$$X = \lambda_0 Y + \lambda_1 F.$$

Remark: Such an element is regular if and only if $\lambda_0 \neq 0$.

Corollary: When π belongs to $\mathcal{L}^+$ or $\mathcal{L}^-$, X belongs to $sl(2)$, and $\pi(X)$ has an eigenvalue, X satisfies one and only one of the following assertions:

X regular and $\pi(X)$ diagonalizable;

X belongs to $\mathbf{C}F$, $\pi(X)$ has the single eigenvalue 0, which is of multiplicity 1.

We can now carry out a partition of π_r.

Proposition 2.5: Let π be a member of $\mathcal{D}$, π' a member of $\mathcal{L}^+$ or $\mathcal{L}^-$; one has $O(\pi) \cap O(\pi') = \emptyset$.

Proof: Let us suppose that the intersection is not empty; that is, that there exists an A in $\mathcal{A}'$ such that $\pi' = \pi \circ A$; then AF, which is nilpotent, would be satisfactory: $\pi(AF)$ has an eigenvalue, which is in contradiction with Proposition 2.3. QED

Proposition 2.6: Given π and π' belonging to $\mathcal{L}^+$ or $\mathcal{L}^-$,
(1) If π is $\downarrow l$ and π' is $\downarrow l'$ with $l \neq l'$, then $O(\pi) \cap O(\pi') = \emptyset$;
(2) If π is $\downarrow \bar{l}$ and π' is $\downarrow \bar{l}$ then $O(\pi) \cap O(\pi') = \emptyset$.

Proof:
(1) is evident since $AQ = Q$ for every A in $\mathcal{A}'$.
(2) Let us suppose that $\pi' = \pi \circ A$; then $\pi(AY)$ should have $-l-1$ as "maximum" eigenvalue. This is in contradiction with Proposition 2.4. QED

Let $\mathcal{D}'$ (resp. $\mathcal{L}'^+$, resp. $\mathcal{L}'^-$) be the union of the orbits under $\mathcal{A}'$ of the elements of $\mathcal{D}$ (resp. $\mathcal{L}^+$, resp. $\mathcal{L}^-$).

Theorem 2.2: $\mathcal{D}'$, $\mathcal{L}'^+$, and $\mathcal{L}'^-$ form a partition of π_r. The proof results from Propositions 2.2, 2.5, and 2.6. More precisely, inside $\mathcal{D}'$, $\mathcal{L}'^+$, and $\mathcal{L}'^-$ one separates the orbits with the help of the following propositions.

Proposition 2.7: Given π and π' belonging to $\mathcal{D}$, π of type $D(l, m_0)$ and π' of type $D(l', m'_0)$.
(1) When $l \neq l'$, then $O(\pi) \cap O(\pi') = \emptyset$.
(2) When $l = l'$, then $O(\pi') = O(\pi)$ if and only if $\pi' = \pi \circ A$, where A is one of the automorphisms i_μ or Ω_μ, $\mu \in \mathbf{C} - \{0\}$.

Proof: (1) is evident. (2) If $O(\pi) = O(\pi')$, then there exists A such that $\pi' = \pi \circ A$. Since $\pi(AY)$ has an eigenvalue, $AY = \pm Y$ (Proposition 2.3), and it is easy to see that A is necessarily of the form given. QED

Proposition 2.8: Given π and π' belonging to $\mathcal{L}^+$ (resp. $\mathcal{L}^-$), then $O(\pi) = O(\pi')$ if and only if $\pi = \pi \circ A$, where A is one of the automorphisms

$$A = e^{\mathrm{ad}\lambda F} \circ i_\mu, \quad \lambda \in \mathbf{C}, \quad \mu \in \mathbf{C} - \{0\}.$$

Proof: Let us suppose $O(\pi) = O(\pi')$, there exists an automorphism A such that $\pi' = \pi \circ A$. Then $\pi(AF)$ has the eigenvalue 0 and AF is nilpotent; thus $AF = \mu F$ (Corollary (2.4), $\mu e\ \mathbf{C} - \{0\}$, and it is easy to see that A is of the form given. QED

Since $\mathcal{A}'$ acts transitively on each orbit, it only remains to determine the stabilizers.

Proposition 2.9: Given π belonging to $\mathcal{D}$,
(1) if π is $D(l, m_0)$ with $m_0 \neq 0$ and $m_0 \neq 1/2$, the stabilizer of π is made up of the automorphisms i_μ, $\mu \in \mathbf{C} - \{0\}$.
(2) if π is $D(l, 0)$ or $D(l, 1/2)$, the stabilizer of π is the group of the automorphisms i_μ and Ω_μ, $\mu \in \mathbf{C} - \{0\}$.

Proof: The proof results from Proposition 2.7.

Proposition 2.10: If π belongs to $\mathcal{L}^+$ or $\mathcal{L}^-$, the stabilizer of π is made up of the automorphisms $e^{\mathrm{ad}\lambda F} \circ i_\mu$, $\lambda \in \mathbf{C}$, $\mu \in \mathbf{C} - \{0\}$.

Proof: The proof results from Propositions 2.6 and 2.8.

We can now carry out a classification.

(a) Let $\mathcal{C}$ be the set of pairs of complex numbers (l, m_0) satisfying: l admissible, $0 \leq \mathrm{Re}\, m_0 < 1$, $l - m_0$ and $l + m_0$ nonintegers; R the equivalence relation on $\mathcal{C}$ defined by:

$$(l, m_0) \quad R(l', m'_0)\, l = l' \text{ and } \begin{cases} \text{either } m'_0 = m_0 \\ \qquad \text{or} \\ m'_0 = -m_0 + 1 \end{cases} \Big\} \text{if Re } m_0 \neq 0$$

$$\text{or } m'_0 = \pm m_0 \text{ if Re } m_0 = 0.$$

According to what precedes, $\mathcal{C}/\mathrm{R}$ is in bijection with the set of orbits of $\mathcal{D}$ under $\mathcal{A}'$.

Theorem 2.3: The elements of $\mathcal{D}'$ are characterized by:
(1) an element $[l, m_0]$ of $\mathcal{C}/\mathrm{R}$;

(2) An element $\hat{A}$ of $\mathcal{A}'/\{i_\mu\}$ (where $\{i_\mu\}$ is the group of automorphisms i_μ, $\mu \in \mathbf{C} - \{0\}$) when $[l, m_0] \neq [l, 0]$ and $[l, 1/2]$; an element A of $\mathcal{A}'/\{i_\mu, \Omega_\mu\}$ (where $\{i_\mu, \Omega_\mu\}$ is the group of automorphisms i_μ and Ω_μ, $\mu \in \mathbf{C} - \{0\}$) otherwise.

Proof: Given π belonging to $\mathcal{D}'$. There exists one and only one orbit O such that $\pi \in O$; Thus we conclude (1).

Now, depending on the case:

If $[l, m_0] \neq [l, 0]$ and $[l, 1/2]$, one knows that the stabilizer of the two elements of $\mathcal{D}$ in O is $\{i_\mu\}$.

If $[l, m_0] = [l, 0]$ or $[l, 1/2]$, then the stabilizer of the

element of $\mathcal{O}$ in O is $\{i_\mu, \Omega_\mu\}$. QED

(b) Let us now treat the case of L'^+ and L'^-.

Theorem 2.4: The elements of L'^+ (resp. L'^-) are characterized by:
(1) A complex l satisfying the hypotheses of the series $\downarrow\dot{l}$ (resp. $\downarrow\bar{l}$) of Miller;
(2) An element A of $\mathcal{A}'/\{e^{\mathrm{ad}\lambda F}\circ i_\mu\}$, where $\{e^{\mathrm{ad}\lambda F}\circ i_\mu\}$ is the group of automorphisms of the form $e^{\mathrm{ad}\lambda F}\circ i_\mu$, $\lambda \in \mathbf{C}$, $\mu \in \mathbf{C}-\{0\}$.

Proof: Given π an element of L'^+ (for example), it is clear that $O(\pi)\cap L^+$ contains only one element (Proposition 2.6) characterized by l, of which the stabilizer is $\{e^{\mathrm{ad}\lambda F}\circ i_\mu\}$ (Proposition 2.10). QED

Remark: One can see that the classification of π_r by Theorems 2.3 and 2.4 is equivalent to the one we gave in Ref. 7.

C. Construction and classification of the irreducible representations of $sl(2)$ for which the representative of a nilpotent element (at least) has an eigenvalue.

We denote by π_n the set of the classes of such representations.

Proposition 2.11: Given π an element of π_n. We suppose that there exists a nilpotent $X \in sl(2)$ such that $\pi(X)$ has the eigenvalue 0. Then π belongs to π_r.

Proof: There exists an automorphism A of $sl(2)$ such that $AF=X$. Let $\pi'=\pi\circ A$ and let $\varphi\neq 0$ be a vector such that $\pi(X)\varphi=0$. According to Lemma 1.1, $\pi(Q)=qI$, $q\in\mathbf{C}$.

Thus, we have $q\varphi=\pi'(G)\pi'(F)\varphi+\pi'(Y)\varphi+\pi'(Y)^2\varphi$, $[-q+\pi'(Y)+\pi'(Y^2)]\varphi=0$, which proves that $\pi'(Y)$ has an eigenvalue. QED

As a result, we are interested in the elements of π_n in which the representative of a nilpotent element has an eigenvalue different from 0. Let us first construct a particular representation from which we shall obtain all the others.

Theorem 2.5: When in an irreducible representation π of $sl(2)$ $\pi(F)$ has the eigenvalue 1, there exists a basis f_n, $n\in\mathbf{N}$, of the representation space V such that:

$$\pi(Y)f_n=f_{n+1},\ \pi(F)f_n=[\pi(Y)-1]^n f_0,\ \pi(Q)f_n=qf_n, q\in\mathbf{C},$$
$$\pi(G)f_n=[\pi(Y)+1]^n(qf_0-f_1-f_2).$$

Conversely, if for a representation of $sl(2)$ there exists a basis f_n, $n\in\mathbf{N}$, of V on which the representation acts as above, then $\pi(F)$ has only the eigenvalue 1 which is of multiplicity 1; the representation is irreducible and characterized (up to equivalence) by q.

Proof: The direct part results immediately from the formulas of Lemma 1.3 on setting $f_n=\pi(Y)^n f_0$.

Conversely, calculation shows that $\pi(F)$ has only the eigenvalue 1 and that it is of multiplicity 1 (keeping in mind the fact that $\pi(Y)$ has no eigenvalue). Now let V_n be the subspace of V generated by $(f_0, f_1, \ldots, f_n)$, and let $X\neq 0$ be an element of V_n.

$\pi(F-1)X\in V_{n-1}$; thus, there exists an integer k such that $\pi(F-1)^k X=0$ and $\pi(F-1)^{k-1}X\neq 0$, That is, $\pi(F-1)^{k-1}X=\lambda f_0$, $\lambda\in\mathbf{C}-\{0\}$.

Now if W is an invariant subspace containing a vector X which is not zero, according to what has just been said $f_0\in W$, Thus $W=V$. QED

Let us direct the study of π_n to that of the set of the representations of Theorem 2.5, which we denote by $\mathcal{M}_1$:

Proposition 2.12: π belongs to $\pi_n-\pi_r$ if and only if there exists π' belonging to $\mathcal{M}_1$ such that $O(\pi)=O(\pi')$.

Proof: Let X be a nilpotent element such that $\pi(X)$ has an eigenvalue $\mu\neq 0$. There exists an automorphism A of $sl(2)$ such that $AF=(1/\mu)X$. Let π' be the representation $\pi\circ A$. π' belongs to $\mathcal{M}_1$ and $O(\pi)=O(\pi')$. QED

Proposition 2.13: When π and π' belong to $\mathcal{M}_1$ and $\pi\neq\pi'$, one has $O(\pi)\cap O(\pi')=\emptyset$.

Proof: The proof follows from Theorem 2.5.
A simple calculation proves the following proposition.

Proposition 2.14: When π belongs to $\mathcal{M}_1$, $\pi(X)$ has an eigenvalue if and only if X belongs to $\mathbf{C}F$.
Thus, one has

Proposition 2.15: Given π an element of $\mathcal{M}_1$, the stabilizer of π is $\{e^{\mathrm{ad}\lambda F}, \lambda\in\mathbf{C}\}$.

Proof: In fact, $\pi=\pi\circ A$ implies that $\pi(AF)$ has the eigenvalue 1, and hence that $AF=F$ according to Proposition 2.14. QED
Let us denote by $\mathcal{M}_1'$ the set $\pi_n-\pi_r$.

Theorem 2.6: The elements of $\mathcal{M}_1'$ are characterized by:
(1) $q\in\mathbf{C}$;
(2) An element of $\mathcal{A}'/\{e^{\mathrm{ad}\lambda F}\}$, where $\{e^{\mathrm{ad}\lambda F}\}$ is the group of automorphisms $e^{\mathrm{ad}\lambda F}$, $\lambda\in C$.

Proof: The proof is the same as for theorems 2.3 and 2.4, with the aid of Propositions 2.12, 2.13, and 2.15.

Remarks: (1) The classification is equivalent to the one we have given in [7].
(2) Paragraphs 2 and 3 give a construction and classification of the elements of $\pi_{sl(2)}$.

3. THE ENVELOPING ALGEBRA OF $sl(2)$

A. Decomposition of the adjoint representation

The term adjoint representation of $sl(2)$ in $\mathcal{U}$ is applied to the representation defined by:

$$\mathrm{ad}g(X)=[g,X],\ g\in sl(2),\ X\in\mathcal{U}.$$

Let $(\mathcal{U}_n)_{n\in N}$ be the canonical filtration of[2] $\mathcal{U}$. The subspaces $\mathcal{U}_n$ are stable for this representation. The restriction of the adjoint representation to $\mathcal{U}_n$ is completely reducible; hence the adjoint representation itself is completely reducible. Let us establish the exact form of the decomposition. We reason by induction and suppose that:

$$\mathcal{U}_{n-1}=\underbrace{[\oplus_{i_1} R(F^{n-1-2i_1}Q^{i_1}, G^{n-1-2i_1}Q^{i_1})]}_{H_{n-1}}$$

$$\oplus\Big[\underbrace{\oplus_{i_2} R(F^{n-2-2i_2}Q^{i_2}, G^{n-2-2i_2}Q^{i_2})}_{H_{n-2}}\Big]\oplus\cdots\oplus \underbrace{R(F,G)}_{H_1}\oplus\underbrace{C}_{H_0}$$

$$= H_{n-1}\oplus H_{n-2}\oplus\cdots\oplus H_1\oplus H_0,$$

with

$$0\le i_k\le (n-k)/2 \quad \text{if } (n-k) \text{ is even},$$

$$0\le i_k\le (n-k-1)/2 \quad \text{if } (n-k) \text{ is odd};$$

where $R(u,v)$ is the representation space of the finite-dimensional irreducible representation in which the vector of "maximal" weight is u and the vector of minimal weight is v, [for example, $R(F,G)=sl(2)$]. In addition, we assume that the non-null elements of $H_{n-k}=\oplus_{i_k} R(F^{n-k-2i_k}Q^{i_k}, G^{n-k-2i_k}Q^{i}_{k})$ are of degree $(n-k)$. The hypothesis of induction is true for ranks 0 and 1.

The only element (within a scalar multiple) of $\mathcal{U}_n$ which is of weight n is F^n; thus this element necessarily belongs to one of the irreducible representations which decompose the restricted adjoint to $\mathcal{U}_n$. Let $R(F^n, V)$ be the representation space; since the only vectors of $\mathcal{U}_n$ of weight-n are the scalar multiples of G^n, one has

$$R(F^n, V)=R(F^n, G^n) \text{(Theorem 2.1)}.$$

Since $Q\in Z$ and $\mathcal{U}$ is entire, one has

$$H'_n = QH_{n-2}=\oplus_{i_2} R(F^{n-2-2i_2}Q^{i_2+1}, G^{n-2-2i_2}Q^{i_2+1})$$
$$=\oplus_{i'_2} R(F^{n-2i'_2}Q^{i'_2}, G^{n-2i'_2}Q^{i'_2})$$

where

$$0<i'_2\le n/2 \quad \text{if } n \text{ is even},$$

$$0<i'_2\le (n-1)/2 \quad \text{if } n \text{ is odd}.$$

By the hypothesis of induction the non-null elements of H'_n are of degree n; thus $H'_n\cap \mathcal{U}_{n-1}=\{0\}$.

Moreover $(\mathcal{U}_{n-1}\oplus H'_n)\cap R(F^n, G^n)$ is an invariant subspace of $R(F^n, G^n)$ which contains neither F^n nor G^n and thus reduces to $\{0\}$.

The dimension of the direct sum

$$\mathcal{U}_{n-1}\oplus H'_n\oplus R(F^n, G^n)$$

is

$$[(n+1)(n+2)(n+3)]/6,$$

that is, dim $\mathcal{U}_n$. QED

Theorem 3.1: The decomposition of the adjoint representation of $sl(2)$ in $\mathcal{U}$ is given by

$$\mathcal{U}=\oplus_{n\ge 0} H_n,$$

where

$$H_0=C,\quad H_n=\oplus_{i_n} R(F^{n-2i_n}Q^{i_n}, G^{n-2i_n}Q^{i_n}),$$

with

$$0\le i_n\le n/2 \quad \text{if } n \text{ is even},$$

$$0\le i_n\le (n-1)/2 \quad \text{if } n \text{ is odd},$$

with $\mathcal{U}_n=\oplus_{0\le j\le n} H_j$, the nonnull elements of H_n being of degree n.

Corollary 3.1: When X belongs to $sl(2)$, the centralizer $\mathcal{C}(X)$ of X in $\mathcal{U}$ is $C[X,Q]$.

Corollary 3.2: When X_1 and X_2 belong to $sl(2)$ and are linearly independent,

$$\mathcal{C}(X_1)\cap\mathcal{C}(X_2)=Z.$$

Proof: If one of the two elements considered is regular (for example, X_1), one is led to the case of Y and X_2 [by automorphism of $sl(2)$]. Let π be an element of $\mathcal{D}$. Since X_2 is not a multiple of Y, $\pi(X_2)$ has no eigenvalue (Proposition 2.3). Let $\mu\in\mathcal{C}(Y)\cap\mathcal{C}(X_2)$. $\pi(\mu)$ is diagonal, but if $\mu\notin Z$, this implies that $\pi(X_2)$ has an eigenvalue, which is contradictory.

If X_1 and X_2 are nilpotent, one carries out an analogous demonstration with an element of $\mathcal{N}_1$. QED

The decomposition of the adjoint representation permits us to establish

Theorem 3.2: Let W be the subspace of $\mathcal{U}$ defined by $W=\oplus_{n=0}^{\infty} R(F^n, G^n)$. One has $\mathcal{U}=W\oplus(Q-\lambda)\mathcal{U}, \forall\lambda\in C$. In fact, the decomposition of *Theorem* 3.1 can be written

$$\mathcal{U}=\bigoplus_{p=0}^{\infty}(Q-\lambda)^p W.$$

Corollary: $\mathcal{U}$ is an infinite-dimensional free module on its center $Z=C[Q]$.

(This corollary has been demonstrated in a much more general framework by Kostant.)[8]

We are now able to determine the kernels of the finite-dimensional representations.

Theorem 3.3: Let π be an irreducible representation of type $D(2n)$, $2n\in N$; then

$$\text{Ker }\pi=\bigoplus_{k\ge 2n+1} R(F^k, G^k)\oplus(Q-\alpha_n)\mathcal{U}, \quad \text{where } \alpha_n=n(n+1).$$

Proof: Since F^k belongs to Kerπ when k is greater than or equal to $2n+1$, we have the inclusion

$$\mathcal{X}=\bigoplus_{k\ge 2n+1} R(F^k, G^k)\oplus(Q-\alpha_n)\mathcal{U}\subset \text{Ker}\pi.$$

Since the sum $\oplus_{k<2n+1} R(F^k, g^k)+\mathcal{X}=\mathcal{U}$ is direct, the dimension of $\mathcal{U}/\mathcal{X}$ is $(2n+1)^2$. Then, according to Burnside's theorem,[9] dim $\mathcal{U}/\text{ker}\pi=(2n+1)^2$. Thus we conclude the theorem. QED

B. Particular automorphisms of $\mathcal{U}$

Let us denote by $\mathcal{A}$ the group of automorphisms of $\mathcal{U}$. We shall identify $\mathcal{A}'$ with a subgroup of $\mathcal{A}$ (each element of $\mathcal{A}'$ being extended in a unique way to an element of $\mathcal{A}$ by the Poincaré-Birkhoff-Witt theorem).

Theorem 3.4: Every element A of $\mathcal{A}$ satisfies $AQ=Q$.

Proof: Since A belongs to $\mathcal{A}$, we have $Z=C[Q]=C[AQ]$. Thus, $AQ=\lambda Q+\mu\lambda$, $\mu\in C, \lambda\neq 0$. Now let π be a representation of $sl(2)$, of type $D(2n)$, π' the representation $\pi'=\pi\circ A$. Since $A\mathcal{U}=\mathcal{U}$, the representation π' is also irreducible and thus of type $D(2n)$. That is,

$$\pi'(Q)=n(n+1).$$

Now,

$$\pi'(Q)=\lambda\pi(Q)+\mu=\lambda n(n+1)+\mu.$$

This being true for all n, it follows that $\lambda=1$, $\mu=0$. QED

Corollary 3.3: When A belongs to $\mathcal{A}$, we have $\mathcal{U} = AW \oplus (Q-\lambda)\mathcal{U}, \forall \lambda \in \mathbf{C}$.

Corollary 3.4: When A belongs to $\mathcal{A}$, it induces an automorphism of the quotient algebra $B_\lambda = \mathcal{U}/(Q-\lambda)\mathcal{U}$ of $\mathcal{U}$ by the two-sided ideal $(Q-\lambda)\mathcal{U}$.

We shall now characterize some particular automorphisms of $\mathcal{U}$.

Theorem 3.5: The only elements of $\mathcal{A}$ such that $AY = Y$ (respectively $AY = -Y$) are the extensions to $\mathcal{U}$ of the automorphisms i_μ (respectively Ω_μ) of $sl(2)$.

Proof: Given A belonging to $\mathcal{A}$ such that $AY = Y$. Since $AQ = Q$ we have $AFAG = FG$. AF and AG belong to $[\mathcal{U},\mathcal{U}]$ and neither can be of degree 0; thus they are of degree 1 and have no component on Z in the decomposition $\mathcal{U} = Z \oplus [\mathcal{U},\mathcal{U}]$. Thus they belong to $sl(2)$. QED

Corollary 3.5: Given A belonging to $\mathcal{A}$. A belongs to $\mathcal{A}'$ if and only if there exists a regular element X of $sl(2)$ such that $AX \in sl(2)$.

Corollary 3.6: Let A and A' be two automorphisms of $\mathcal{U}$ satisfying $AY = A'Y$ and $AF = A'F$; then $A = A'$.

Moreover,

Theorem 3.6: The only elements A of $\mathcal{A}$ such that $AF = F$ are the automorphisms of the form $e^{\mathrm{ad}\mu}$, $\mu \in C(F)$.

The existence of such automorphisms is justified by the following technical lemma.

Lemma 3.1: Given u belonging to $C(F) = \mathbf{C}[F,Q]$, then ad u is a locally nilpotent derivation of $\mathcal{U}$. [i.e., $\forall V \in \mathcal{U} / \exists n$ such that $(\mathrm{ad}u)^n(V) = 0$].

Proof of the theorem: One has $[AY,F] = F$ or, equivalently, $[AY - Y, F] = 0$. Let $AY - Y = \Sigma_i \lambda_i(Q)F^i$, where $\lambda_i(Q) \in \mathbf{C}[Q][\lambda_0(Q) = 0]$. If we now set $A' = \exp(-\mathrm{ad}\Sigma_i(1/i)\lambda_i(Q)F^i)$, we obtain $AY = A'Y$, $AF = A'F$, and, according to Corollary 3.6, $A = A'$. QED

C. Abelian subalgebras of B_λ and $\mathcal{U}$.

We recall that B_λ denotes the algebra $B_\lambda = \mathcal{U}/(Q-\lambda)\mathcal{U}$. (One may consult Ref. 10 for a detailed study of these algebras.) As Dixmier[10] points out (without proof), the properties of the abelian subalgebras of B_λ are analogous to those of the abelian subalgebras of the Heisenberg algebra A_1. We give a proof of this assertion (using Ref. 11) in Ref. 12. The result is

Proposition 3.1:

(1) Let X be a non-scalar element of B_λ (resp. a non-central element of $\mathcal{U}$). Then $C(X)$ is commutative.

(2) Let $\mathcal{B}$ be a subalgebra of B_λ (resp. of $\mathcal{U}$). The following conditions are equivalent:

(a) $\mathcal{B}$ is maximal abelian.

(b) There exists an element X of $\mathcal{B}$ with $X \notin \mathbf{C}$ (resp. $X \notin Z$), such that $\mathcal{B} = C(X)$.

(c) $\mathcal{B}$ is different from $\mathbf{C}$ (resp. from Z), and for all y of $\mathcal{B}$ not belonging to $\mathbf{C}$ (resp. to Z) one has $\mathcal{B} = C(y)$.

(3) Let x and y be two elements of B_λ (resp. of $\mathcal{U}$) not belonging to $\mathbf{C}$ (resp. to Z). $[x,y] = 0$ if and only if $C(X) = C(y)$.

(4) In B_λ (resp. in $\mathcal{U}$) the intersection of two different maximal abelian subalgebras is $\mathbf{C}$ (resp. Z).

(5) Let $\mathcal{B}$ be a subalgebra of B_λ (resp. of $\mathcal{U}$) different from $\mathbf{C}$ (resp. from Z) and $C(\mathcal{B})$ be its centralizer in B_λ (resp. in $\mathcal{U}$).

(a) If $\mathcal{B}$ is not Abelian, $C(\mathcal{B}) = \mathbf{C}$ (resp. $C(\mathcal{B}) = Z$).

(b) If $\mathcal{B}$ is Abelian, $C(\mathcal{B})$ is maximal Abelian.

(6) Let $\mathcal{B}$ be a maximal Abelian subalgebra of B_λ (resp. of $\mathcal{U}$), y an element of B_λ (resp. of $\mathcal{U}$), and p a nonscalar element of $\mathbf{C}[X]$ (resp. of $Z[X]$). If $p(y)$ belongs to $\mathcal{B}$, then y belongs to $\mathcal{B}$.

4. REPRESENTATIONS

A. Preliminaries and notation

When π is an infinite-dimensional irreducible representation of $sl(2)$, there exists a complex number λ such that Ker $\pi = (Q-\lambda)\mathcal{U}$ (according to Lemma 1.1 and Ref. 13). Thus π induces on passing to the quotient a faithful irreducible representation of B_λ. Conversely, for every infinite-dimensional irreducible representation π_λ of B_λ, there exists an irreducible representation π of $\mathcal{U}$ such that Ker $\pi = (Q-\lambda)\mathcal{U}$ and π_λ is induced by π on passing to the quotient.

Let $\mathcal{A}_\lambda$ be the group of automorphisms of B_λ, Π (resp. Π_λ) the set of classes of infinite-dimensional irreducible representations of $\mathcal{U}$ (resp. of B_λ). $\mathcal{A}$ (resp. $\mathcal{A}_\lambda$) acts on Π (resp. on Π_λ): the image of π by A is $\pi' = \pi \circ A$.

Every automorphism A of $\mathcal{U}$ satisfies $AQ = Q$ (Theorem 2.2) and induces on passing to the quotient an automorphism of B_λ. Incidentally, by Ref. 10 every automorphism of B_λ comes on passing to the quotient from an automorphism (at least) of $\mathcal{U}$. Consequently, to study the action of $\mathcal{A}$ on Π, we study the action of $\mathcal{A}_\lambda$ on Π_λ.

Let us denote by $\mathcal{A}'_\lambda$ the group of the automorphisms of the canonical image of $sl(2)$ in B_λ, which we equally denote by $sl(2)$.

Let us denote by $\mathcal{D}_\lambda$ (resp. π^+_λ, π^-_λ) the set of the elements of Π_λ induced by the elements of $\mathcal{D}$ of type $D(l, m_0)$ (resp. the element of Π_λ induced by the representation $\overset{+}{\mathcal{U}}{}^l$, resp. induced by the representation $\overset{+}{\mathcal{U}}{}^{\bar{l}}$) satisfying $l(l+1) = \lambda$, and by π^0_λ the elements of Π_λ induced by the element of $\mathcal{M}$ such that the representative of the Casimir element is λ.

We denote by the same letter a class of representations and one of its elements, and by $\mathcal{O}(\pi)$ the orbit of the element π of Π_λ under the action of $\mathcal{A}_\lambda$.

Remark: Theorems 3.5 and 3.6 of Sec. 3 remain true for $\mathcal{A}_\lambda$ with analogous proofs.

B. Spectral theory

Given an element π of Π_λ and a vector φ of its representation space, we shall term the annihilator of φ (written Ann φ) the left ideal $\{X \in B_\lambda \mid \pi(X)\varphi = 0\}$.

Proposition 4.1: Let π be an element of $\mathcal{D}_\lambda$, φ an eigenvector of $\pi(Y)$ of eigenvalue ν_φ. Then Ann $\varphi = B_\lambda(Y - \nu_\varphi)$.

Proof: There exists a basis $\{f_n, n \in \mathbf{Z}\}$ of the space such that

$$\pi(Y)f_n = \lambda_n f_n, \quad \lambda_n = m_0 + n, \quad 0 \leq \mathrm{Re}\, m_0 < 1,$$
$$\pi(F)f_n = \alpha_n f_{n+1}, \quad \alpha_n, \beta_n \neq 0, \quad \forall n,$$
$$\pi(G)f_n = \beta_n f_{n-1},$$
$$\varphi = f_p, \quad \nu_\varphi = \lambda_p. \quad \text{(see Ref. 5).}$$

Let $X = \Sigma_n G^n p_n(Y) + \Sigma_k F^k p'_k(Y) \in \mathrm{Ann}\ \varphi, p_n(Y) \in \mathbf{C}[Y]$, $p'_k(Y) \in \mathbf{C}[Y]$. $\pi(X)\varphi = 0$ may be rewritten

$$\sum_n p_n(\lambda_p)\beta_p \cdots \beta_{p-n} f_n + \sum_k p'_k(\lambda_p)\alpha_p \cdots \alpha_{p+k} f_{p+k} = 0$$

It results from this expression that

$$p_n(\lambda_p) = p'_k(\lambda_p) = 0 \quad \forall n, k \qquad \text{QED}$$

Corollary: Let l be an admissible complex number such that $l(l+1) = \lambda$. Let ν be a complex number such that neither $l + \nu$ nor $l - \nu$ is an integer. Then the left ideal $B_\lambda(Y - \nu)$ is maximal.

Proposition 4.2: Let φ be an eigenvector of $\pi^+_\lambda(Y)$ [resp. of $\pi^-_\lambda(Y)$] of eigenvalue ν_φ. Then there exists an integer k such that

$$\mathrm{Ann}\ \varphi = B_\lambda(Y - \nu_\varphi) + B_\lambda F^{k+1}.$$

Proof: We carry out the proof for $\pi = \pi^+_\lambda$. There exists a basis $\{f_n, n \in \mathbf{N}\}$ of the representation space such that

$$\pi(Y)f_n = \lambda_n f_n \text{ with } \lambda_n = l - n.$$
$$\pi(F)f_n = \alpha_n f_{n-1} \text{ with } \begin{cases} \alpha_n \neq 0 & \text{if } n \neq 0, \\ \alpha_0 = 0, \end{cases}$$
$$\pi(G)f_n = \beta_n f_{n+1} \text{ with } \beta_n \in \mathbf{C} - \{0\},$$
$$\varphi = f_k, \quad \nu_\varphi = \lambda_k.$$

It suffices to show that $\mathrm{Ann}\ \varphi C B_\lambda(Y - \lambda_k) + B_\lambda F^{k+1}$. Let us first consider the case of an element X of $\mathrm{Ann}\ \varphi$ satisfying $d^0_F X = 0$.

Let $X = \Sigma_j \mu_j(Y)G^j$, $\mu_j(Y) \in \mathbf{C}[Y]$. The equality $\pi(X)(f_k) = 0$ leads to

$\pi[\mu_j(Y)]f_{k+j} = 0 \quad \forall j$. That is, there exists $u_j(Y) \in \mathbf{C}[Y]$ such that

$$\mu_j(Y) = u_j(Y)(Y - \lambda_{k+j}).$$

We thus obtain $X = (\Sigma_j u_j(Y)G^j)(Y - \lambda_k) \in B_\lambda(Y - \lambda_k)$.

Let us suppose that all the elements of $\mathrm{Ann}\ \varphi$ of which the degree in F is less than or equal to $(n-1)$ belong to $B_\lambda(Y - \lambda_k) + B_\lambda F^{k+1}$, and let $X \in \mathrm{Ann}\ \varphi$ satisfy $d^0_F X = n$; for example, $X = \Sigma_j \mu_j(Y)F^j + \ldots$ If $n \geq k+1$, $V = X - \mu_n(Y)F^n \in \mathrm{Ann}\ \varphi$ and satisfies $d^0_F V \leq n-1$; thus, $X \in B_\lambda(Y - \lambda_k) + B_\lambda F^{k+1}$. If not, $\pi(X)f_k = 0$ leads to $\alpha_k \cdots \alpha_{k-j+1}\mu_j(\lambda_{k-j}) = 0$, $\forall j$, where the α_t are all different from 0. Hence, $\mu_j(\lambda_{k-j}) = 0 \ \forall j$.

In particular, it follows that $\mu_n(Y)F^n \in B_\lambda(Y - \lambda_k)$, and that the element $V = X - \mu_n(Y)F^n \in \mathrm{Ann}\ \varphi$ and satisfies $d^0_F V \leq n-1$. QED

Corollary: Let ν be a complex number satisfying one of the following hypotheses:
(a) $\nu = -l - 1 - k$, $k \in N$;
(b) $\nu = l - k$, $k \in N$, when $l \neq (h-1)/2$, $h \in \mathbf{N}$;
then the left ideal $B_\lambda(Y - \nu) + B_\lambda F^{k+1}$ is maximal.

Proposition 4.3: Let φ be an eigenvector of $\pi^0_\lambda(F)$; then one has $\mathrm{Ann}\ \varphi = B_\lambda(F - 1)$.

Proof: Let us recall that the only eigenvalue of $\pi^0_\lambda(F)$ is 1 and that the associated eigenspace is $\mathbf{C}\varphi$ (Sec. 2). Let $X = \Sigma_i \lambda_i(Y)G^i + \Sigma_j \mu_j(Y)F^j \in \mathrm{Ann}\ \varphi, \lambda_i(Y)$ and $\mu_j(Y) \in \mathbf{C}[Y]$ with $\mu_0(Y) = 0$. Since $\pi^0_\lambda(X)(\varphi) = \pi^0_\lambda [\Sigma_i \lambda_i(Y)\alpha(Y+i-1)\cdots\alpha(Y) + \Sigma_j \mu_j(Y)](\varphi) = 0$ and since $\pi^0_\lambda(Y)$ has no eigenvalues, we have

$$\sum_i \lambda_i(Y)\alpha(Y+i-1)\cdots\alpha(Y) + \sum_j \mu_j(Y) = 0.$$

Let us reason by induction on the d^0_F of the elements of $\mathrm{Ann}\ \varphi$. Let us suppose $d^0_F X = 0$, or $X = \Sigma_i \lambda_i(Y)G^i$. One writes

$$X = \sum_i \lambda_i(Y)G^i(1 - F^{i+1})$$
$$+ \left[\sum_i \lambda_i(Y)\alpha(Y)\cdots\alpha(Y+i-1)\right] F \in B_\lambda(F-1).$$
$$0$$

Let us suppose the property is true for rank $(n-1)$; let $X \in \mathrm{Ann}\ \varphi$ be such that $d^0_F X = n$, $\mu_n(Y)$ is the coefficient of F^n in the development of X. Then the element $V = X - \mu_n(Y)(F-1)^n \in \mathrm{Ann}\ \varphi$ and satisfies $d^0_F V \leq n-1$. QED

Corollary: The left ideal $B_\lambda(F-1)$ is maximal.

Proposition 4.4: Let π be an element of $\mathcal{D}_\lambda$, φ a vector of the representation space which is not an eigenvector of $\pi(Y)$; Then $\mathrm{Ann}\ \varphi$ is not monogenic.

Proof: We make use of the notation of Proposition 4.1 and set $\varphi = \Sigma_k \lambda_k f_k$. The element $y = \Pi_k(Y - \lambda_k) \in \mathrm{Ann}\ \varphi$.

If $\mathrm{Ann}\ \varphi$ were monogenic there would exist u and X such that

$$\begin{cases} X \in \mathrm{Ann}\ \varphi, \\ y = uX. \end{cases}$$

On writing $y = uX$, we obtain

$$\left.\begin{array}{l} d^0_F u = 0 \\ \text{or} \\ d^0_F X = 0 \end{array}\right\} \text{ and } \left\{\begin{array}{l} d^0_G u = 0 \\ \text{or} \\ d^0_G X = 0. \end{array}\right.$$

If $d^0_F u = d^0_G u = 0, X$ is a scalar multiple of y, and the proposition is true since $B_\lambda y$ is not maximal.

Let us consider the case where $d^0_F u = 0$ and $d^0_G u \neq 0$; then we have $d^0_G X = 0$, which enables us to write

$$\left.\begin{array}{l} u = \sum_j \mu_j(Y)G^j \\ X = \sum_k \lambda_k(Y)F^k \end{array}\right\} \mu_j(Y), \quad \lambda_k(Y) \in \mathbf{C}[Y];$$
$$y = \sum_{j,k} \mu_j(Y)\lambda_k(Y+j)G^j F^k.$$

One then has two possibilities:

(1) There exists k such that $k \geq j \ \forall_j$. Then if k_0 (resp. j_0) denotes the largest k (resp. the smallest j),

$$G^{k_0}F^{j_0} = p(Y)G^{k_0 - j_0}, \quad p(Y) \in \mathbf{C}[Y].$$

Thus $k_0 = j_0$.
This leads to $\begin{cases} u = \mu(Y)G^{k_0}, \\ X = \lambda(Y)F^{k_0}. \end{cases}$

Since $X \in \text{Ann}\,\varphi$, $\lambda(Y) \in \text{Ann}\,F^{k_0}\varphi$. However, this requires that $d^0_Y \lambda(Y) \geq d^0_Y y$ and is contradictory to $y = uX$ in the case where $k_0 > 0$. Only the case where $k_0 = 0$ remains, and then a previous line of reasoning is applied.

(2) There exists some j such that $j \geq k\ \forall k$. One again uses analogous reasoning. Finally the case $d^0_F u \neq 0$, $d^0_G u = 0$, and $d^0_F X = 0$ is deduced from what precedes by the action of the automorphism Ω_1. QED

Proposition 4.5: Let $\varphi = \Sigma_{k=0}^{p} \rho_k f_k$ (notation of Proposition 4.2, $\rho_p \neq 0$, be a vector of the representation space π^+_λ(resp. π^-_λ). There exists an element v of $\mathbf{C}[F]$ such that

$$\text{Ann}\,\varphi = B_\lambda(e^{\text{ad}v}Y - l + p) + B_\lambda F^{p+1}$$

$$(\text{resp. Ann}\,\varphi = B_\lambda(e^{\text{ad}v}Y + l + p + 1) + B_\lambda F^{p+1}).$$

Proof: One may choose the basis $\{f_k, k \in \mathbf{N}\}$ in such a way that $\pi^+_\lambda(F) f_k = f_{k-1}$, $k > 0$. (It suffices to replace f_k by $(\alpha_1 \cdots \alpha_k)^{-1} f_k$, $k > 0$.)

We solve the equation

$$\pi^+_\lambda(Y + \mu_1 F + \cdots + \mu_p F^p)\,\varphi = (l - p)\,\varphi$$

by reducing it to a linear system. Let $\mu_1 \cdots \mu_p$ be the solutions and A the automorphism

$$A = \exp\{-\text{ad}[\mu_1 F + (\mu_2/2)F^2 + \ldots + (\mu_p/p)F^p]\}$$

The representation $\pi^+_\lambda \circ A$ is equivalent to π^+_λ, and φ is the $p^{\text{-th}}$ vector of the basis on which $\pi^+_\lambda \circ A(Y)$ is diagonal. Thus Proposition 4.2 is applicable. QED

To obtain the analogue of Propositions 4.4 and 4.5 in the case of π^0_λ, we prove the following lemmas.

Lemma 4.1: Let φ be an eigenvector of $\pi^0_\lambda(F)$, $\Psi = \Sigma_{i=0}^{p} \lambda_i \pi^0_\lambda(Y)^i \varphi$, where $\lambda_i \in \mathbf{C}, \lambda_p \neq 0$, $X = \Sigma_{j=0}^{q} \mu_j(Y) F^j$, where $\mu_j(Y) \in C[Y]$, such that $\pi^0_\lambda(X)\,\Psi = 0$. Then there exists u of the form $\Sigma_k \nu_k(Y) F^k$, where $\nu_k(Y) \in \mathbf{C}[Y]$, such that $X = u(F - 1)$.

Proof: $\pi^0_\lambda(X)\Psi = \pi^0_\lambda[\Sigma_j \mu_j(Y)\Sigma_i \lambda_i (Y - j)^i]\,\varphi = 0$, and since $\pi^0_\lambda(Y)$ has no eigenvalue, $\Sigma_j \mu_j(Y)\Sigma_i \lambda_i (Y - j)^i = 0$.

However, the term of highest degree in Y of such a polynomial is the same as the term of highest degree of the product $p_1 p_2$ where $p_1 = \Sigma_j \mu_j(Y)$ and $p_2 = \Sigma_i \lambda_i Y^i$. This then implies that $p_1 = 0$.

Hence $\pi^0_\lambda(X)\,\varphi = \pi^0_\lambda[p_1(Y)]\,\varphi = 0$

By Proposition 4.3, there exists u such that $X = u(F - 1)$. The hypothesis $d^0_G u > 0$ is contradictory, since $d^0_G uF < d^0_G u$. QED

Lemma 4.2: Let X and Ψ be as in Lemma 4.1. Then one has $d^0 X \geq p + 1$.

Proof: We reason by induction on the degree in $\pi^0_\lambda(Y)$ of the decomposition of Ψ. If Ψ is a scalar multiple of φ, the lemma is true by Proposition 4.5. Let us suppose the property to be true for all elements ξ of the representation space of the form

$$\xi = \sum_{i=0}^{k} \xi_i \pi^0_\lambda(Y)^i \varphi,\ \xi_k \neq 0,\ k \leq p,$$

and let

$$\Psi = \sum_{i=0}^{p+1} \lambda_i \pi^0_\lambda(Y)^i \varphi,\ \lambda_{p+1} \neq 0.$$

Let us make the hypothesis that there exists an element X of Ann Ψ of the form specified in Lemma 4.1 but of total degree equal to $p + 1$. According to Lemma 4.1 there exists u such that $X = u(F - 1)$ with $d^0_G u = 0$. It is clear that one necessarily has $d^0 u = p$.

Let us set $\xi = \pi^0_\lambda(F - 1)$. Since $\Psi \neq \lambda\varphi$, $\lambda \in \mathbf{C}$, ξ is different from 0 and u belongs to Ann ξ.

Now $\xi = \Sigma_{i=0}^{p+1} \lambda_i [\pi^0_\lambda(Y - i)^i - Y^i]\varphi$ is of the form $\Sigma_{i=0}^{k} \xi_i \pi^0_\lambda(Y)^i \varphi$, with $k = p$. Thus the hypothesis of induction gives us $d^0 u \geq p + 1$, which is contradictory.

We can now establish

Proposition 4.6: When Ψ is not a scalar multiple of φ, Ann Ψ is not monogenic.

Proof: Let $\Psi = \Sigma_{i=0}^{p} \lambda_i \pi^0_\lambda(Y)^i \varphi$, $\lambda_p \neq 0$. One verifies easily that $(F - 1)^{p+1}$ belongs to Ann Ψ. (However, when $k < p + 1$, $(F - 1)^k \notin \text{Ann}\,\Psi$ according to Lemma 4.2.)

If Ann Ψ were monogenic, there would exist u and X such that

$$\begin{cases} (F - 1)^{p+1} = uX, \\ X \in \text{Ann}\,\Psi \end{cases}$$

On writing the first equality, one sees that either $d^0_G X = 0$ or $d^0_G u = 0$. In the first case, Lemma 4.2 proves that $d^0 X \geq p + 1$. Thus the proposition is true, because $B_\lambda(F - 1)^{p+1}$ is not maximal. In the second case, we write

$$\left.\begin{aligned} u &= \sum_k \nu_k(Y) F^k \\ X &= \sum_i \lambda_i(Y) F^i + \sum_j \mu_j(Y) G^j \end{aligned}\right\} \quad \nu_k(Y),\ \lambda_i(Y),\ \mu_j(Y) \in \mathbf{C}[Y],$$

$$(F - 1)^{p+1} = \underbrace{\sum_{i,k} \lambda_i(Y)\nu_k(Y - 1)F^{k+i}}_{(1)} + \underbrace{\sum_{j,k} \mu_j(Y)\nu_k(Y + j)G^j F^k}_{(2)}.$$

Let $j_0 \neq 0$ (resp. k_0) be the largest of the j (resp. the smallest of the k) which occur. If we suppose that $j_0 > k_0$, we have $d^0_G(\mu_{j_0}(Y)\nu_{k_0}(Y + j_0)G^{j_0} F^{k_0}) > 0$, and this term cannot disappear. If one then supposes that $j_0 < k_0$, all the $G^j F^k$ which occur may be written as $p_{i_j}(Y)F^{k-j}$, $p_{i_j}(Y) \in \mathbf{C}[Y]$, $k - j > 0$. Since j_0 is strictly positive, k_0 is also. Thus (1) and (2) contain only strictly positive powers of F. This is in contradiction with the hypothesis $(F - 1)^{p+1} = (1) + (2)$.

There remains only the case where $j_0 = k_0$, but then the equation

$$\mu_{j_0}(Y)\nu_{j_0}(Y + j_0)G^{j_0} F^{j_0} = 1$$

should be satisfied. This is impossible, since $G^{j_0} F^{j_0}$ is a polynomial in Y of degree $2j_0$. Since the hypothesis that $j_0 \neq 0$ is contradictory is each case, one concludes that $j_0 = 0$, and returns to the first case. QED

C. Orbits

Lemma 4.3: Let π be an element of Π_λ, $\pi' = \pi \circ A$ an element of its orbit, and φ an element of the representation space of π and π'; then $\text{Ann}_{\pi'}\,\varphi = A^{-1}[\text{Ann}_\pi \varphi]$.

Theorem 4.1: When π belongs to $\mathcal{D}_\lambda$, one has

$$O(\pi)\cap O(\pi_\lambda^+)=O(\pi)\cap O(\pi_\lambda^-)=O(\pi)\cap O(\pi_\lambda^0)=O(\pi_\lambda^+)\cap O(\pi_\lambda^0)$$
$$=O(\pi_\lambda^-)\cap O(\pi_\lambda^0)=\emptyset.$$

Proof: None of the annihilators corresponding to π_λ^+ and π_λ^- is monogenic (Proposition 4.5). However, for π the annihilators of the basis vectors are monogenic (Proposition 4.1), and for π_λ^0 the annihilator of the initial vector is also (Proposition 4.3). Thus:

$$O(\pi)\cap O(\pi_\lambda^+)=O(\pi)\cap O(\pi_\lambda^-)=O(\pi_\lambda^+)\cap O(\pi_\lambda^0)=O(\pi_\lambda^-)\cap O(\pi_\lambda^0)$$
$$=\emptyset \text{ (Lemma 4.3).}$$

Let us suppose there exists an $A\in\mathcal{A}_\lambda$ such that $\pi_\lambda^0=\pi\circ A$. The only monogenic annihilator corresponding to π_λ^0 is $B_\lambda(F-1)$. Thus there exists a complex number ν such that

$$B_\lambda(AF-1)=B_\lambda(Y-\nu) \text{ (Lemma 4.3).}$$

Consequently, $AF=1+c(Y-\nu)$, $c\in\mathbf{C}-\{0\}$. This result is impossible, since AF is strictly nilpotent and $1+c(Y-\nu)$ is strictly semi-simple [10]. QED

Proposition 4.7: When π belongs to $\mathcal{D}$, the only automorphisms A such that $\pi\circ A$ belongs to $\mathcal{D}$ are

$$i_\mu,\ \mu\in\mathbf{C}-\{0\},$$

$$\Omega_\mu,\ \mu\in\mathbf{C}-\{0\},$$

Proof: Let us suppose that π is of type $D(l,m_0)$ and $\pi\circ A$ of type $D(l,m_0')$. Given Propositions 4.1 and 4.4, there exists $n\in\mathbf{Z}$ such that

$$B_\lambda(AY-m_0')=B_\lambda(Y-m_0+n).$$

Thus there exists $\nu\in C$ such that

$$AY=\nu Y+\nu(n-m_0)+m_0'.$$

Since Y and $AY\in[B_\lambda,B_\lambda]$, $AY=\nu Y$ by necessity [10], and since the spectrum of ad AY is $\mathbf{Z}$ (the same as that of ad$Y$), $\nu=\pm1$ by necessity. QED

Corollary 4.1: When π belongs to $\mathcal{D}$, the stabilizer of π is

$\{i_\mu,\ \mu\in\mathbf{C}-\{0\}\}$, when π is of type $D(l,m_0)$ with $m_0\neq0$ or $1/2$;

$\{i_\mu$ and $\Omega_\mu,\ \mu\in\mathbf{C}-\{0\}\}$, otherwise.

Corollary 4.2: $\pi\circ A$ belongs to $\Pi_{sl(2)}$ if and only if A belongs to $\mathcal{A}_\lambda'$.

Proof: If $\pi\circ A\in\Pi_{sl(2)}$, there exists a regular element z of $sl(2)$ such that $\pi\circ A(z)$ has an eigenvalue (Theorem 4.1 and Section 2). Let $A\in\mathcal{A}_\lambda'$ be such that $A'Y=z$, $\pi'=\pi\circ A\circ A'$, $\pi'(Y)=\pi\circ A\circ A'(Y)=\pi\circ A(z)$.

Now, by Theorem 4.1 π' belongs to $\mathcal{D}$, and by Proposition 2.12 we have necessarily that $A\circ A'=i_\mu$ or $A\circ A'=\Omega_\mu$, $\mu\in\mathbf{C}-\{0\}$.

Conversely, if $A\in\mathcal{A}_\lambda'$, $\pi\circ A(A^{-1}Y)=\pi(Y)$ has an eigenvalue, and $A^{-1}Y\in sl(2)$. Thus $\pi\circ A\in\Pi_{sl(2)}$. QED

Remark: Since $\mathcal{A}_\lambda'$ is strictly contained in $\mathcal{A}_\lambda$, one has actually constructed representations which don't belong to $\Pi_{sl(2)}$, that is, for which no representative of any element of $sl(2)$ has an eigenvalue.

Proposition 4.8: The stabilizer of π_λ^0 is $\{e^{\mathrm{ad}u}, u\in\mathbf{C}[F]\}$.

Proof: Considering Propositions 4.3 and 4.6 and Lemma 4.3, one writes

$$B_\lambda(AF-1)=B_\lambda(F-1).$$

Thus there exists $\xi\in\mathbf{C}$ such that

$$AF=\xi F-\xi+1,$$

and since AF and F belong to $[B_\lambda,B_\lambda]$, $AF=F$.

Corollary 4.3: $\pi_\lambda^0\circ A$ belongs to $\Pi_{sl(2)}$ if and only if A is of the form

$$e^{\mathrm{ad}u}\circ A',\ u\in\mathbf{C}[f],\ A'\in\mathcal{A}_\lambda'.$$

Proof: By arguments analogous to those of Theorem 2.6, $\pi\circ A\in\Pi_{sl(2)}$ if and only if there exists a nilpotent z in $sl(2)$ such that $\pi\circ A(z)$ has an eigenvalue μ different from 0. Let $A'\in\mathcal{A}_\lambda'$ be such that $A'F=(1/\mu)z$. The representation $\pi\circ A\circ A'$ is equivalent to π_λ^0. Thus $A\circ A'=e^{\mathrm{ad}u}$, $u\in\mathbf{C}[F]$.

The converse is evident.

Remark: As before, one in fact obtains representations where no representative of the elements of $sl(2)$ has an eigenvalue.

5. THE REPRESENTATIONS π_λ^n

Let us denote by Π_λ' the union of the orbits under $\mathcal{A}_\lambda$ of π_λ^0, π_λ^+, π_λ^-, and $\mathcal{D}_\lambda$. The problem which presents itself is to know whether $\Pi_\lambda'\subset\Pi_\lambda$ strictly or note.

A. Construction of π_λ^n

Let $\mathcal{J}_n$ be the left ideal $B_\lambda(Y^nF-1)$ of B_λ, n an integer >0. Since (Y^nF-1) has no inverse [10], there exists a maximal left ideal $\mathcal{J}_n$ which contains $\mathcal{J}_n$. Let π_λ^n be the irreducible representation defined on the quotient $V_n=B_\lambda/\mathcal{J}_n$ and φ the class of 1.

Lemma 5.1: $\pi_\lambda^n(Y)$ and $\pi_\lambda^n(F)$ have no eigenvalue.

Proof: If $\pi_\lambda^n(Y)$ had an eigenvalue, one would have $\pi_\lambda^n=\pi$, where π belongs to $\mathcal{D}$ or to $\{\pi_\lambda^+\}\cup\{\pi_\lambda^-\}$. This is impossible, since $\pi(Y^nF\}$ can have no eigenvalue ($\neq0$).

Lemma 5.2: φ is not an eigenvector of $\pi_\lambda^n(Y^mF)$, $0\leq m<n$.

Proof: In fact, if one had $\pi_\lambda^n(Y^mF)\varphi=\mu\varphi$, then it would follow that $\pi_\lambda^n(Y^nF)\varphi=\pi_\lambda^n(Y^{n-m})\pi_\lambda^n(Y^mF)\varphi=\mu\,\pi_\lambda^n(Y^{n-m})\varphi=\varphi$. This result is in contradiction with Lemma 5.1.

Theorem 5.1: $\{\pi_\lambda^n(Y^k)\varphi,\ \pi_\lambda^n(Y^mF^p)\varphi,\ k\in\mathbf{N},\ p\in\mathbf{N}-\{0\},\ 0\leq m<n\}$ is a basis of V_n.

Proof: By construction $V_n=\pi_\lambda^n(B_\lambda)\varphi$; thus $\{\pi_\lambda^n(Y^kF^p)\varphi,\ \pi_\lambda^n(Y^lG^r)\varphi,\ l,p,k\in\mathbf{N},\ r\in\mathbf{N}-\{0\}\}$ is generator.

Since $\pi_\lambda^n(G^r)\varphi=\pi_\lambda^n(G^rY^nF)\varphi=\pi_\lambda^n[(Y+r)^n\alpha(Y+r-1)G^{r-1}]\varphi$, an induction on r shows that $\{\pi_\lambda^n(Y^kF^p)\varphi\}$ generates V_n. Finally, for $k>n$, one has

$$Y^kF^p=Y^{k-n}F^{p-1}(Y+p-1)^nF$$

or $\pi_\lambda^n(Y^kF^p)\varphi=\pi_\lambda^n(Y^{k-n}F^{p-1})\varphi+\Sigma_{j=1}^n C_n^j(p-1)^j$

$\pi_\lambda^n[Y^{k-n}(Y-p+1)^{n-j}F^p]\varphi$, which expresses $\pi_\lambda^n(Y^kF^p)\varphi$ as

a linear combination of the $\pi_\lambda^n(Y^{k-j}F^p)\varphi$, $1 \leq j \leq n$, and the $\pi_\lambda^n(Y^{k-n}F^{p-1})\phi$. By induction on k, one deduces that $\pi_\lambda^n(Y^kF^p)\varphi$ can be expressed as a combination of the $\pi_\lambda^n(Y^p)\varphi$ and the $\pi_\lambda^n(Y^mF^r)\varphi$, $0 \leq m < n$, and thus that $\{\pi_\lambda^n(Y^p)\varphi, \pi_\lambda^n(Y^mF^p)\varphi, k \in \mathbb{N}, p \in \mathbb{N} - \{0\}, 0 \leq m < n\}$ are generators.

To show that it is free amounts to proving that if a linear combination of its elements of the form $\Sigma\lambda_k Y^k + \Sigma\mu_{m,p}Y^mF^p$ belongs to the annihilator of φ, then it is null.

Let us denote the largest p coming in by p_0. If $p_0 = 0$ the proposition is true by Lemma 5.1. Let us suppose it is true up to rank $p_0 - 1$, and let us consider a combination of this type with $\sup p = p_0$.

$$\pi_\lambda^n(G)\left[\sum_k \lambda_k \pi_\lambda^n(Y^k)\varphi + \sum_{m,p\neq 0} \mu_{m,p}\pi_\lambda^n(Y^mF^p)\varphi\right]$$

$$= \pi_\lambda^n\left[\sum_k \lambda_k(Y+1)^{k+n}\alpha(Y) + \sum_{m,p\neq 0}\mu_{m,p}(Y+1)^m\alpha(Y)F^{p-1}\right]\varphi = 0.$$

Thus, by the hypothesis of induction

$$\sum_k \lambda_k(Y+1)^{k+n}\alpha(Y) + \sum_{m,p\neq 0}\mu_{m,p}(Y+1)^m\alpha(Y)F^{p-1} = 0.$$

From this result one deduces $\mu_{m,p} = 0\ \forall m, \forall p > 1$. There then remains $[\Sigma_k\lambda_k(Y+1)^{k+n} + \Sigma_{m,1}\mu_{m,1}(Y+1)^m]\alpha(Y) = 0$. However, $k + n \geq n > m\ \forall k, \forall m$, and B_λ is entire. Thus $\mu_{m,1} = \lambda_k = 0\ \forall_m, \forall_k$. QED

B. Spectral theory in π_λ^n

Let $\pi_\lambda^{n\prime}$ be the representation (*a priori* not irreducible) defined on the quotient $B_\lambda/\mathcal{J}_n = V_n'$, φ' the class of 1.

Theorem 5.2: $\{\pi_\lambda^{n\prime}(Y^k)\varphi', \pi_\lambda^{n\prime}(Y^mF^p)\varphi', k \in \mathbb{N}, p \in \mathbb{N} - \{0\}, 0 \leq m < n\}$ is a basis of V_n'.

Proof: The first part of the proof of Theorem 5.1 is applicable word for word (since it does not bring in the irreducibility of π_λ^n), and proves that the set is a system of generators.

Let us consider $\Sigma\lambda_k\pi_\lambda^{n\prime}(Y^k)\varphi' + \Sigma_{m,p}\mu_{m,p}\pi_\lambda^{n\prime}(Y^mF^p)\varphi' = 0$. This expression is equivalent to

$\sum_k \lambda_k Y^k + \sum_{m,p}\mu_{m,p}Y^mF^p \in \mathcal{J}n$, which is not possible unless $\lambda_k = \mu_{m,p} = 0\ \forall k, m, p$ (since $\mathcal{J}n \subset \mathcal{J}n$). QED

Corollary 5.1: $\pi_\lambda^{n\prime}$ and π_λ^n are equivalent.

Corollary 5.2: The left ideal $B_\lambda(Y^nF - 1)$ is maximal ($\mathcal{J}n = \mathcal{J}n$).

We can now answer the question asked at the beginning of this section.

Theorem 5.3: For every $n > 0$, π_λ^n does not belong to Π_λ'.

Proof: Since Ann φ is monogenic, π_λ^n cannot belong to $\mathcal{O}(\pi_\lambda^*)$ or $\mathcal{O}(\pi_\lambda^-)$ (Proposition 4.3). If there existed an automorphism A of $\mathcal{A}_\lambda$ such that $\pi_\lambda^n = \pi' \circ A$, $\pi' \in \mathcal{D}_\lambda$, then by Propositions 4.1 and 4.4 there would exist complex numbers ν and $c \neq 0$ such that

$$Y^nF = 1 + c(A^{-1}Y - \nu).$$

However, $1 + C(A^{-1}Y - \nu)$ is strictly semi-simple, whereas Y^nF is not (Ref. 10, Lemma 5.2). Thus the hypothesis is contradictory.

Likewise, Y^nF not being strictly nilpotent, $\pi_\lambda^n \notin \mathcal{O}(\pi_\lambda^0)$.

QED

ACKNOWLEDGMENT

It is a pleasure to express our gratitude to our thesis advisor, Dr. M. Flato, for his patience and guidance at all stages of this paper.

[1]J. P. Serre, *Algèbres de Lie semi-simples complexes* (Benjamin, New York, 1966).
[2]N. Jacobson, *Lie Algebras* (Interscience, New York, 1966).
[3]J. Dixmier, An. Acad. Bras. Cienc. 35, 491 (1963).
[4]D. Quillen, Proc. Am. Math. Soc. 21, 271 (1969).
[5]W. Miller, Jr., Mem. Am. Math. Soc. No. 50 (1964).
[6]W. Miller, Jr., *Lie Theory and Special Functions* (Academic, New York, 1968).
[7]D. Arnal and G. Pinczon, C.R. Acad. Sci. (Paris) 272, 1369 (1971).
[8]B. Kostant, Am. J. Math. 85, 327 (1963).
[9]N. Bourbaki, *Elements de mathêmctiques. Livre II: Algèbre* (Hermann, Paris, 1958), Chap. 8.
[10]J. Dixmier, J. Algebra 24, 557 (1973).
[11]S.A. Amitsur, Pac. J. Math. 8, 1 (1958).
[12]D. Arnal and G. Pinczon, thesis (Faculté des Sciences, Dijon, 1972).
[13]Y. Nouaze and P. Gabriel, J. Algebra 6, 77, 79 (1967).

On nonlinear transformations in vector spaces. Colineation and conformal groups

Hans Tilgner

Fachbereich 19 Mathematik der Freien Universität, Königin-Luisestr. 24, 1 Berlin 33, Germany
(Received 27 March 1972)

The sets of special colineations and conformal transformations of a pseudo-orthogonal vector space can be given only a Lie groupoid structure, rather than a Lie group structure because of the nonvanishing denominators. With every such groupoid one can associate a unique Lie group, which, however, no longer consists of transformations. For one-parameter subgroups one can define infinitesimal transformations and a bilinear composition, called Lie bracket, which reduces for linear transformations to the commutator. In the special cases of colineations and conformal transformations on pseudo-orthogonal vector spaces of arbitrary finite dimension and signatures, covering homomorphisms onto matrix groups are given together with the corresponding Lie algebra isomorphisms.

I. INTRODUCTION

There have been several applications of nonlinear transformation groups in physics. Especially the group of conformal transformations of Minkowski space has been used in electrodynamics and particle physics for the description of massless particles. Colineations were used in physics only once, namely in a three-dimensional Euclidean space for the optical mapping by lenses. Since the colineation group of Minkowski space, which is shown below to be isomorphic to $Sl(5, \mathrm{R})$, contains the Poincaré group as a subgroup, it may serve as a generalization of the latter to describe internal symmetry. However its kinematical interpretation in Minkowski space seems to be difficult, since it has no such special geometrical aspects as the Poincaré group and the conformal group, except being the most general transformation group which transforms straight (world) lines into straight (world) lines. This geometrical property suggests its application in scattering theory in a way which parallels its use in geometrical optics. In a scattering process, the S matrix should be invariant under the full colineation group.

In the following a mathematical description of the full colineation and conformal groups of arbitrary finite dimensional pseudo-orthogonal vector spaces over the real numbers is given. It is shown that sets of rational transformations like special colineations and special conformal transformations form a groupoid only, rather than a group, since there are nonvanishing denominators. There is a unique group associated with every such groupoid, which however consists no longer of vector space transformations. By means of the directional derivative it is possible to define for every one-parameter subgroup a socalled infinitesimal transformation on the underlying vector space. For the two cases in question these infinitesimal transformations are polynomial; therefore there is no trouble with their domains of definition. The set of infinitesimal transformations can be made a vector space and even a Lie algebra, and the Lie bracket coincides with the commutator of two infinitesimal transformations only if they are linear.

For the construction of those Lie algebras, one-parameter subgroups of some classical matrix groups are described. The full colineation group of a pseudo-orthogonal vector space V is shown to be globally isomorphic to $Sl(\mathrm{R}\oplus V, \mathrm{R})$ for all possible signatures of the bilinear form on V. The full conformal group is locally isomorphic (in some special cases, depending on dimension and signature, even globally isomorphic) to the noncompact pseudo-orthogonal groups in $\mathrm{R}\oplus V\oplus \mathrm{R}$ with a suitable indefinite, nondegenerate, symmetric bilinear form. Both covering homomorphisms are given explicitly. The Lie algebras of infinitesimal transformations are isomorphic to $sl(\mathrm{R}\oplus V, \mathrm{R})$ resp. to the noncompact pseudo-orthogonal Lie algebras in $\mathrm{R}\oplus V\oplus \mathrm{R}$. The latter isomorphism establishes a bijection between the conformal Lie algebras in V and the noncompact pseudo-orthogonal Lie algebras in $\mathrm{R}\oplus V\oplus \mathrm{R}$.

II. THE GROUP OF BIRATIONAL TRANSFORMATIONS OF A VECTOR SPACE

Given a n-dimensional real vector space V with basis $p^1, \cdots, p^n$, a *rational* (*polynomial*) transformation T of V can be written

$$T(x)=\sum_{i=1}^{n}\Omega_i(\xi_1,\cdots,\xi_n)p^i, \tag{2.1}$$

where the coefficients Ω_i are rational (polynomial) functions of the $\xi_i\in\mathrm{R}$, and $x=\xi_1p^1+\cdots+\xi_np^n$. Writing the Ω_i's as reduced quotients of polynomials, the least-common multiple of their denominators is uniquely determined up to a constant factor and called *denominator* of T, $\mathrm{den}(T)$. In general T is defined only on an open subset $\mathrm{Dom}(T)$ (in the R^n topology of V) of V, called its *domain*, given by all $x\in V$ for which $\mathrm{Den}(T)\neq 0$. Two transformations S and T are said to be *composable* on $\mathrm{Dom}(ST)$ if

$$\mathrm{Dom}(ST)=\{x\in V/x\in\mathrm{Dom}(T)\text{ and }Tx\in\mathrm{Dom}(S)\} \tag{2.2}$$

is not empty, i.e. if $\mathrm{Im}(T)\cap\mathrm{Dom}(S)\neq\emptyset$. A rational transformation is said to be *birational* if there is another rational transformation T^{-1} such that T and T^{-1} as well as T^{-1} and T are composable and $T^{-1}T=\mathrm{id}_{\mathrm{Dom}(T)}$, $TT^{-1}=\mathrm{id}_{\mathrm{Im}(T)}$. The composition

$$(S,\mathrm{Dom}(S))\,(T,\mathrm{Dom}(T))=(ST,\mathrm{Dom}(ST)) \tag{2.3}$$

defines an associative multiplication on the pairs $(T,\mathrm{Dom}(T))$. Every element (id_A, A) with A an open subset of V and id_A the identical transformation on A, is an *identity*, i.e.,

$$\begin{aligned}&(T,\mathrm{Dom}(T))\,(\mathrm{id}_A,A)=(T,A\cap\mathrm{Dom}(T))\\&(\mathrm{id}_A,A)\,(T,\mathrm{Dom}(T))=(T,\mathrm{Dom}(T)\cap\mathrm{Im}(T)\cap A).\end{aligned} \tag{2.4}$$

For a birational transformation T, the domain of the inverse transformation T^{-1} is again an open subset $\mathrm{Dom}(T^{-1})=\mathrm{Im}(T)$ of V. The inverse of $(T, \mathrm{Dom}(T))$, T birational, is $(T^{-1}, \mathrm{Im}(T))$ since $(T, \mathrm{Dom}(T))\,(T^{-1}, \mathrm{Im}(T))=(\mathrm{id}, \mathrm{Im}(T))$ and $(T^{-1}, \mathrm{Im}(T))\,(T, \mathrm{Dom}(T))=(\mathrm{id}, \mathrm{Dom}(T))$ are identities. Thus the set of pairs $(T, \mathrm{Dom}(T))$, T birational, is a groupoid, but not a group (for the definition of groupoids see Ref. 1, p. 112 and Ref. 2). The relation

$$(T, \mathrm{Dom}(T))\sim(S, \mathrm{Dom}(S)) \text{ iff } T=S \text{ in } \mathrm{Dom}(T)\cap\mathrm{Dom}(S) \tag{2.5}$$

is an equivalence relation. $\mathrm{Dom}(T)\cap\mathrm{Dom}(S)$ is a nonempty open subset of V. The set of equivalence classes is a group which we call *group of birational transformations* of V, $\mathrm{Birat}(V)$. It is isomorphic to the group $P(V)$ defined in Ref. 3, p. 3 and Ref. 4, p. 354.

III. THE LIE ALGEBRA OF INFINITESIMAL TRANSFORMATIONS OF A LIE SUBGROUP OF BIRAT(V)

Given two vector spaces V_1, V_2, a mapping f from some open subspace $\mathfrak{D}\subset V_1$ into V_2 is called *differentiable* in $x\in\mathfrak{D}$, if there is a linear mapping $\partial f(x)/\partial x : V_1\to V_2$, called the *directional derivative* of f, such that for all $u\in V_1$

$$\frac{\partial f(x)}{\partial x}u=\lim_{\epsilon\to 0}\frac{1}{\epsilon}[f(x+\epsilon u)-f(x)].$$

Given a mapping θ from R into the set of birational transformations of V, such that the induced mapping into $\mathrm{Birat}(V)$ is a one-parameter subgroup, we call the transformation $\delta\theta$, defined by

$$(\delta\theta)(x)=\lim_{\mu\to 0}\frac{\partial\theta(\mu)(x)}{\partial\mu}\,1,\quad x\in\bigcap_{\mu<\epsilon}\mathrm{Dom}(\theta(\mu)), \tag{3.1}$$

the *infinitesimal transformation* of θ. In the special cases of the colineation and conformal groups, $\delta\theta$ is a polynomial transformation for all one-parameter subgroups (as shown below) and therefore $\mathrm{Dom}(\delta\theta)=V$. Given one-parameter subgroups θ, θ', ... such that the corresponding infinitesimal transformations are polynomial, the *sum* of their infinitesimal transformations $\delta\theta$, $\delta\theta'$, ... is defined by

$$(\alpha\,\delta\theta+\beta\,\delta\theta'+\cdots)(x)=\lim_{\mu\to 0}\frac{\partial\theta(\alpha\mu)\theta'(\beta\mu)\cdots(x)}{\partial\mu}\,1. \tag{3.2}$$

The set of (polynomial) transformations generated in this way by the infinitesimal transformations is a vector space (Ref. 5, p. 114) which is finite dimensional if the one-parameter subgroups are taken from a Lie subgroup of $\mathrm{Birat}(V)$. Taking all possible one-parameter subgroups from this group G, the dimension of this vector space $\mathrm{Lie}(G)$ becomes equal to that of G, since the exponential mapping of Lie groups defines a bijection between the one-parameter subgroups and the elements of the Lie algebra [the latter is defined as the vector space of left invariant vector fields on G; thus it is different from $\mathrm{Lie}(G)$]. The sum of two infinitesimal transformations defined in (3.2) coincides with the pointwise sum. This can be verified easily for the two cases to be treated below. The Lie bracket of two infinitesimal transformations $\delta\theta$ and $\delta\theta'$ is defined by

$$\{\delta\theta, \delta\theta'\}(x)=\lim_{\mu\to 0}\frac{\partial\theta(\sqrt{\mu})\theta'(\sqrt{\mu})\theta(\sqrt{\mu})^{-1}\theta'(\sqrt{\mu})^{-1}(x)}{\partial\mu}\,1. \tag{3.3}$$

The pair $[\mathrm{Lie}(G), \{\,,\,\}]$ is a Lie algebra which is isomorphic to the Lie algebra of G (Ref. 5, p. 141). In physics one often uses besides $\mathrm{Lie}(G)$ another concept of Lie algebra given by vector fields on V (usually called *generators*) V^θ, defined by

$$(V^\theta f)(x)=\lim_{\mu\to 0}\frac{\partial f(\theta(\mu)(x))}{\partial\mu}\,1 \tag{3.4}$$

for some real-valued function on V (Ref. 6, p. 33) (the Lie algebras defined in mathematics as left invariant vector fields on G seem to play no role in physics).

The exponential mapping of $\mathrm{Lie}(G)$ into G is defined by $\mathrm{Exp} : \delta\theta\mapsto\theta(1)$. It is known, that it coincides for linear transformations with the exponential series (Ref. 7, p. 101). From the Baker—Campbell—Hausdorf formula (Ref. 7, p. 96) follows, that in this case the Lie bracket coincides with the commutator of two infinitesimal transformations. This cannot be generalized to nonlinear transformations. The Lie algebra of infinitesimal transformations can be defined for linear Lie transformation groups G by

$$\mathrm{Lie}(G)=\{\delta\theta\in\mathrm{end}(V)/\exp(\xi\delta\theta)\in G, \xi\in\mathrm{R}\} \tag{3.5}$$

together with the commutator (Ref. 3, p. 94).

IV. ONE-PARAMETER SUBGROUPS AND BASIS-FREE COMMUTATION RELATIONS OF SOME CLASSICAL MATRIX GROUPS

(i) The general linear groups. Given a symmetric bilinear form τ on V, and a mapping $g(b\otimes a)\in gl(V, \mathrm{R})$ defined by

$$g(b\otimes a)\,x=\tau(a, x)b, \tag{4.1}$$

the linear transformations

$$G^{(\mu)}(b\otimes a)=\mathrm{id}_V+[\tau(a, b)]^{-1}(e^{\mu\tau(a,b)}-1)g(b\otimes a),$$
$$G^{(\mu)}(b\otimes a)=\mathrm{id}_V+\mu g(b\otimes a), \tag{4.2}$$

for $\tau(a, b)\neq 0$, resp. $\tau(a, b)=0$, define one-parameter subgroups $\theta_{Gl}: \mu\mapsto G^{(\mu)}(b\otimes a)$ of $Gl(V, \mathrm{R})$ because of

$$\delta\theta_{Gl}=g(b\otimes a),\quad \exp[\mu g(b\otimes a)]=G^{(\mu)}(b\otimes a). \tag{4.3}$$

The Lie bracket, which is the commutator, becomes

$$[g(b\otimes a), g(d\otimes c)]=\tau(a, d)g(b\otimes c)-\tau(b, c)g(d\otimes a). \tag{4.4}$$

If τ is nondegenerate, in which case we will write $\langle a, b\rangle$ with $a, b\in V$, $g(V\otimes V)$ generates $gl(V, \mathrm{R})$, i.e., every element of $gl(V, \mathrm{R})$ is a sum of $g(b\otimes a)$'s. The adjoint of $g(b\otimes a)$ with respect to τ is $g(a\otimes b)$. If the matrix I of $\langle\,,\,\rangle$ reduces to the identity, $g(b\otimes a)$ becomes the usual tensor product of the column of b and the row of a in the

chosen basis of V. The adjoint then reduces to the transposed. From (4.2) and (4.3) follows for $\tau(a,b)\neq 0$

$$\mathrm{id}_V + \alpha g(a\otimes b) = \exp[\tau(a,b)^{-1}\log[1+\alpha\tau(a,b)]g(a\otimes b)], \tag{4.5}$$

if $1+\alpha\tau(a,b)>0$, and for $\tau(a,b)=0$

$$\mathrm{id}_V + \alpha g(a\otimes b) = \exp[\alpha g(a\otimes b)]. \tag{4.6}$$

Using then that for all $n\times n$ matrices A

$$\det e^A = \exp(\mathrm{Spur}\, A) \tag{4.7}$$

and $\mathrm{Spur}\, g(a\otimes b)=\tau(a,b)$, one proves that

$$\det[\mathrm{id}_V + \alpha g(a\otimes b)] = 1+\alpha\tau(a,b),$$

for $1+\alpha\tau(a,b)\geq 0$. For $1+\alpha\tau(a,b)\neq 0$

$$[\mathrm{id}_V + \alpha g(b\otimes a)]^{-1} = \mathrm{id}_V - \{\alpha/[1+\alpha\tau(a,b)]\}\, g(b\otimes a). \tag{4.9}$$

(ii) The pseudo-orthogonal groups. The group of invertible transformations M of V with $\langle Mx, My\rangle = \langle x, y\rangle$ will be called $\mathrm{Aut}(V, \langle\,,\rangle)$. For positive-definite $\langle\,,\rangle$ we write $O(V,\mathrm{R})$. Its Lie algebra of infinitesimal transformations is written $\mathrm{der}(V,\langle\,,\rangle)$ resp. $so(V,\mathrm{R})$. $\mathrm{der}(V,\langle\,,\rangle)$ is the set of all linear transformations R with $\langle Rx, y\rangle + \langle x, Ry\rangle = 0$. The *discriminant* of $a, b\in V$ is

$$\mathrm{dic}(a,b) = \langle a,b\rangle^2 - \langle a,a\rangle\langle b,b\rangle. \tag{4.10}$$

The linear transformation $o(b\otimes a)$ of V defined by

$$o(b\otimes a)x = \langle a,x\rangle b - \langle b,x\rangle a \tag{4.11}$$

is in $\mathrm{der}(V,\langle\,,\rangle)$. For $\mathrm{dic}(a,b)\neq 0$ resp. $\mathrm{dic}(a,b)=0$, one-parameter subgroups $\theta_{\mathrm{Aut}(V,\langle\,,\rangle)}: \mu \to O^{(\mu)}(b\otimes a)$ of $\mathrm{Aut}(V,\langle\,,\rangle)$ are given by

$$\begin{aligned} O^{(\mu)}(b\otimes a) &= \mathrm{id}_V + \mathrm{dic}(a,b)^{-1}\{\cosh[\mu\sqrt{\mathrm{dic}(a,b)}]-1\}\, o(b\otimes a)^2 \\ &\quad + \mathrm{dic}(a,b)^{-1/2}\sinh\mu\sqrt{\mathrm{dic}(a,b)}\, o(b\otimes a), \\ O^{(\mu)}(b\otimes a) &= \mathrm{id}_V + \mu\, o(b\otimes a) + (\mu^2/2!)\, o(b\otimes a)^2, \end{aligned} \tag{4.12}$$

respectively. It is easy to establish the corresponding relations to (4.3)

$$\delta\theta_{\mathrm{Aut}(V,\langle\,,\rangle)} = o(b\otimes a), \quad \exp[\mu o(b\otimes a)] = O^{(\mu)}(b\otimes a). \tag{4.13}$$

The commutation relations of $\mathrm{der}(V,\langle\,,\rangle)$ become

$$\begin{aligned} [o(b\otimes a), o(d\otimes c)] &= \langle a,c\rangle\, o(d\otimes b) + \langle b,d\rangle\, o(c\otimes a) \\ &\quad - \langle a,d\rangle\, o(c\otimes b) - \langle b,c\rangle\, o(d\otimes a), \end{aligned} \tag{4.14}$$

which may be deduced from the Clifford algebra as well (Ref. 8, p. 232). For $\mathrm{dic}(a,b)<0$, which because of the Schwartz inequality is always the case for positive definite $\langle\,,\rangle$, (4.12) becomes

$$\begin{aligned} O^{(\mu)}(b\otimes a) &= \mathrm{id}_V + \mathrm{dic}(a,b)^{-1/2}\sin\mu\sqrt{-\mathrm{dic}(a,b)}\, o(b\otimes a) \\ &\quad + \mathrm{dic}(a,b)^{-1}\{\cos[\mu\sqrt{-\mathrm{dic}(a,b)}]-1\} \\ &\quad o(b\otimes a)^2. \end{aligned} \tag{4.15}$$

For linearly dependent a and b, $\mathrm{dic}(a,b)=0$. A discussion of the use of $O^{(\mu)}(b\otimes a)$ in the geometry of $SO(2,\mathrm{R})$, $SO(3,\mathrm{R})$ and the Lorentz group is given in (Ref. 9, Chaps. 6 and 7).

(iii) The symplectic groups. For completeness we give the one-parameter subgroups of the symplectic group $Sp(E,\sigma)$ of a symplectic vector space (E,σ), the skew bilinear form σ being nondegenerate. Defining

$$\begin{aligned} s(b\otimes a)x &= \sigma(a,x)b + \sigma(b,x)a, \\ t(b\otimes a)x &= \sigma(a,x)b - \sigma(b,x)a, \end{aligned} \tag{4.16}$$

the linear transformation $s(b\otimes a)$ is in the symplectic matrix Lie algebra $sp(E,\sigma)$ and

$$\begin{aligned} S^{(\mu)}(b\otimes a) &= \mathrm{id}_V + \sigma(a,b)^{-1}\sinh\mu\sigma(a,b)\, s(b\otimes a) \\ &\quad + [\cosh\mu\sigma(a,b)-1]\, t(b\otimes a) \\ S^{(\mu)}(b\otimes a) &= \mathrm{id}_V + \mu s(b\otimes a), \end{aligned} \tag{4.17}$$

for $\sigma(a,b)\neq 0$ resp. $\sigma(a,b)=0$, defines one-parameter subgroups $\theta_{Sp(E,\sigma)}: \mu \mapsto S^{(\mu)}(b\otimes a)$ of $Sp(E,\sigma)$. Since every element of $sp(E,\sigma)$ is a sum of $s(b\otimes a)$'s, the commutation relations of $sp(E,\sigma)$ may be summarized by

$$\begin{aligned} [s(b\otimes a),\, s(d\otimes c)] &= \sigma(a,d)s(c\otimes d) + \sigma(b,d)s(c\otimes a) \\ &\quad + \sigma(a,c)s(d\otimes b) + \sigma(b,c)s(d\otimes a), \end{aligned} \tag{4.18}$$

which may be deduced from the canonical Weyl algebra as well (Ref. 10, f. 120). It is easy to establish the corresponding relations to (4.3) and (4.13).

V. THE COLINEATION GROUP OF A PSEUDO-ORTHOGONAL VECTOR SPACE

In the following we denote the matrix of $\langle\,,\rangle$ by I. A *special colineation* C_a is the rational transformation

$$C_a(x) = x/(1+\langle a,x\rangle), \qquad a, x\in V. \tag{5.1}$$

Since $\langle\,,\rangle$ is nondegenerate it is easy to prove

$$\bigcap_{x\in V}\mathrm{Dom}\, C_x = \{0\}, \quad \bigcup_{x\in V}\mathrm{Dom}\, C_x = V. \tag{5.2}$$

Let $C(V,\langle\,,\rangle)$ denote the corresponding n-dimensional group of birational transformations. It is commutative. Let $T(V)$ denote the group of translations on V, i.e., the set of affine transformations

$$T_a(x) = x + a, \quad x, a\in V. \tag{5.3}$$

The group generated by $T(V)$, $Gl(V,\mathrm{R})$, and $C(V,\langle\,,\rangle)$ is called the *full colineation group* of $(V,\langle\,,\rangle)$, which we write $\mathrm{Col}(V,\langle\,,\rangle)$. For $u\in\mathrm{Dom}\, C_b$ one verifies

$$C_u T_b = T_{C_u(b)} C_{u+\langle u,b\rangle u} N(b\otimes u), \tag{5.4}$$

where

$$N(b\otimes u) = (1+\langle u,b\rangle)^{-1}\,[\mathrm{id}_V - (1+\langle u,b\rangle)^{-1} g(b\otimes u)] \tag{5.5}$$

is in $Gl(V,\mathrm{R})$ and

$$N(b\otimes u)^{-1} = (1+\langle u,b\rangle)\,[\mathrm{id}_V + g(b\otimes u)] \tag{5.6}$$

from (4.9). From (5.4) one verifies for $G, \tilde{G}\in Gl(V,\mathrm{R})$ and $Gc\in\mathrm{Dom}\, C_b$ the multiplication law

$$(T_a C_b G)(T_c C_d G) = T_{a+C_b(Gc)} C_{b+\langle b,Gc\rangle b + I^{-1}[N(Gc\circ b)G]^{-1}\, T\, Id}$$

$$N(Gc\otimes b)\,G\,\tilde{G}. \tag{5.7}$$

For $Gc\notin \mathrm{Dom}\,C_b$ such a relation is not valid. From (5.7)

$$(T_aC_bG)^{-1}=T_{G^{-1}c_{-b(-a)}}C_{-(I+(a,b)I)^{-1}GT_{N(a\otimes b)}T_{Ib}}\,G^{-1}N(a\otimes b) \tag{5.8}$$

for $a\in \mathrm{Dom}\,C_b$. (5.7) and (5.8) can be only deduced for invertible I. It is for that reason, that $\langle\ ,\ \rangle$ was taken to be nondegenerate.

VI. THE ISOMORPHISM BETWEEN COL $(V, <,>)$ AND SL $(\mathrm{R}\oplus V, \mathrm{R})$

Given $\mathbf{x}=\xi\oplus x\in \mathrm{R}\oplus V$, we denote by $\mathscr{B}$ the set of all $\mathbf{x}$ such that $\xi\neq 0$. Then the mapping

$$\Gamma: x\to x/\xi,\quad \Gamma:\mathscr{B}\to V \tag{6.1}$$

is surjective with the fiber over $y\in V$ $\Gamma^{-1}(y)=\{\mathbf{x}\in\mathscr{B}/x=\xi y\}$. Trivially $\Gamma(\mathbf{x})=\Gamma(\mathbf{y})$ implies $\mathbf{x}=(\xi/\eta)\mathbf{y}$. For $\mathbf{A}\in Gl(\mathrm{R}\oplus V,\mathrm{R})$ with $\mathbf{Ax}\in\mathscr{B}$ then $\Gamma(\mathbf{x})=\Gamma(\mathbf{y})$ implies $\Gamma(\mathbf{Ax})=\Gamma[(\xi/\eta)\mathbf{Ay}]$ and this is equal to $\Gamma(\mathbf{Ay})$ because the two arguments lie in the same fiber. Thus the group homomorphism $\Gamma:\mathbf{A}\to\Gamma(\mathbf{A})$, $\Gamma: Sl(\mathrm{R}\oplus V,\mathrm{R})\to \mathrm{Birat}(V)$, defined by

$$\Gamma(\mathbf{A})\,\Gamma(\mathbf{x})=\Gamma(\mathbf{Ax}) \tag{6.2}$$

is well defined. To show that Γ is surjective onto $\mathrm{Col}(V,\langle\ ,\ \rangle)$ one has to consider the Lie algebra of infinitesimal transformations of $Sl(\mathrm{R}\oplus V,\mathrm{R})$, i.e., the set of square $n+1$ matrices with vanishing trace. Every such matrix can be written

$$\begin{aligned}&\delta\mathbf{T}_a+\delta\mathbf{C}_b+\delta\mathbf{G}\\ &:=\begin{pmatrix}0&0\\ a&0\end{pmatrix}+\begin{pmatrix}0&b^TI\\0&0\end{pmatrix}\\ &\quad+\begin{pmatrix}-\mathrm{Spur}\delta G/(1+n)&0\\0&\delta G-[\mathrm{Spur}\delta G/(1+n)]\,\mathrm{id}_V\end{pmatrix},\end{aligned} \tag{6.3}$$

with $\delta G\in gl(V,\mathrm{R})$, $a\in V$ identified with its column in the chosen basis, a^T being the corresponding row. From this we get the one-parameter subgroups

$$e^{\mu\delta\mathbf{T}_a}=\begin{pmatrix}1&0\\ \mu a&\mathrm{id}_V\end{pmatrix}, \tag{6.4}$$

$$e^{\mu\delta\mathbf{C}_b}=\begin{pmatrix}1&\mu b^TI\\0&\mathrm{id}_V\end{pmatrix}, \tag{6.5}$$

$$e^{\mu\delta\mathbf{G}}=(\det e^{\mu\delta G})^{-1/1+n}\begin{pmatrix}1&0\\0&e^{\mu\delta G}\end{pmatrix}, \tag{6.6}$$

of $\mathrm{Sl}(\mathrm{R}\oplus V,\mathrm{R})$. They generate $\mathrm{Sl}(\mathrm{R}\oplus V,\mathrm{R})$. If we write $\exp(\delta\mathbf{T}_a)=:\mathbf{T}_a$, $\exp(\delta\mathbf{C}_b)=:\mathbf{C}_b$, and $\exp(\delta\mathbf{G})=:\mathbf{G}$ then it is easy to prove

$$\Gamma(\mathbf{T}_a)=T_a,\quad \Gamma(\mathbf{C}_b)=C_b,\quad \Gamma(\mathbf{G})=G, \tag{6.7}$$

which shows that Γ is surjective onto $\mathrm{Col}(V,\langle\ ,\ \rangle)$. It is easy to verify that the kernel of Γ is trivial. If we choose on $\mathrm{Col}(V,\langle\ ,\ \rangle)$ the finest topology such that Γ is continuous then $\mathrm{Col}(V,\langle\ ,\ \rangle)$ is isomorphic as a Lie group to $\mathrm{Sl}(\mathrm{R}\oplus V,\mathrm{R})$.

VII. THE LIE ALGEBRA OF INFINITESIMAL COLINEATIONS

A complete set of one-parameter subgroups of $\mathrm{Col}(V,\langle\ ,\ \rangle)$ is given by $\mu\to C_{\mu a}$ (special colineations), $\mu\to T_{\mu b}$ (translations), and $\mu\to G^{(\mu)}(b\otimes a)$ (general linear transformations). From (3.1) follows

$$\delta C_a(x)=-\langle a,x\rangle x, \tag{7.1}$$

$$\delta T_b(x)=b, \tag{7.2}$$

$$\delta Gl(V,\mathrm{R})=gl(V,\mathrm{R}), \tag{7.3}$$

where obviously δT_b is homogeneous of degree zero, $gl(V,\mathrm{R})$ homogeneous of degree one (linear), and δC_a homogeneous of degree two.

The Lie brackets of the infinitesimal colineations $\mathrm{Col}(V,\langle\ ,\ \rangle)=\mathrm{col}(V,\langle\ ,\rangle)$ can be calculated from (3.3), (4.2), (4.9), (5.4) and (5.7)

$$\{\delta C_a,\delta C_b\}=\{\delta T_a,\delta T_b\}=0, \tag{7.4}$$

$$\{\delta C_a,\delta T_b\}=-\delta[g(b\otimes a)+\langle a,b\rangle\,\mathrm{id}_V], \tag{7.5}$$

$$\{\delta C_a,\delta G\}=\delta C_{\Gamma^{-1}(\delta G)Ta}, \tag{7.6}$$

$$\{\delta T_a,\delta G\}=-\delta T_{\delta Ga}, \tag{7.7}$$

together with the commutation relations (4.4) for $gl(V,\mathrm{R})$. The matrices $\delta\mathbf{T}_a$, $\delta\mathbf{C}_b$, $\delta\mathbf{G}$ fulfill the same commutation relations with

$$-\delta[g(b\otimes a)+\langle a,b\rangle\,\mathrm{id}_V]=\begin{pmatrix}\langle a,b\rangle&0\\0&-g(b\otimes a)\end{pmatrix}. \tag{7.8}$$

If we introduce a bilinear form on the vector space $\mathrm{sl}(\mathrm{R}\oplus V,\mathrm{R})$ by $\langle A,B\rangle=\mathrm{Spur}\,AB$, then a verification shows

$$\begin{aligned}\tfrac{1}{2}[[\delta\mathbf{C}_a,\delta\mathbf{T}_x],\delta\mathbf{T}_x]&=-\langle a,x\rangle\delta\mathbf{T}_x\\ &=-\langle\delta\mathbf{C}_a,\delta\mathbf{T}_x\rangle\delta\mathbf{T}_x=\delta C_{\delta\mathbf{T}_a}(\delta\mathbf{T}_x),\end{aligned} \tag{7.9}$$

$$[\delta\mathbf{G},\delta\mathbf{T}_x]=\mathrm{ad}(\delta\mathbf{G})\delta\mathbf{T}_x=\delta\mathbf{T}_{\delta Ga}, \tag{7.10}$$

$$\mathrm{ad}(\delta\mathbf{T}_a)^0\delta\mathbf{T}_x=\delta\mathbf{T}_x=\delta T_{\delta\mathbf{T}_a}(\delta\mathbf{T}_x). \tag{7.11}$$

VIII. THE CONFORMAL GROUP OF A PSEUDO ORTHOGONAL VECTOR SPACE

A special class of linear transformations is given by

$$S(a\otimes b)=\mathrm{id}_V-2\langle a,b\rangle^{-1}\,g(a\otimes b),\quad S(a\otimes a)=:S_a, \tag{8.1}$$

with $S(a\otimes b)^2=\mathrm{id}_V$, $\det S(a\otimes b)=-1$ from (4.8); hence $S(a\otimes b)\in Gl(V,\mathrm{R})$. The $S_a\in\mathrm{Aut}(V,\langle\ ,\ \rangle)$ are called *reflections*. It is known that for positive definite $\langle\ ,\ \rangle$ every element of $\mathrm{Aut}(V,\langle\ ,\ \rangle)$ which is different from the identity can be written as a product of at most n such reflections. By inclusion of the *dilatations*, defined by $D_\lambda x=\lambda x$ for $0\neq\lambda\in\mathrm{R}$, we get the $\frac{1}{2}n(n-1)+1$ dimensional group $D\mathrm{Aut}(V,\langle\ ,\ \rangle)$ resp. $DO(V,\mathrm{R})$ if $\langle\ ,\ \rangle$ is positive definite, which has one, two, or four connectivity components depending on the signature of $\langle\ ,\ \rangle$ and the dimension of V. The connectivity of each component is simple for the trivial one-dimensional case, infinite if n_1 or n_2 in the signature (n_1,n_2) of $\langle\ ,\ \rangle$ equals two, and twofold otherwise.

The nonlinear *special conformal* transformations are given by

$$K_a(x)=(x+\langle x,x\rangle a)/(1+2\langle a,x\rangle+\langle a,a\rangle\langle x,x\rangle). \tag{8.2}$$

The denominator will be shortened to $\omega(a,x)$ in the following.

$$R(x)=x/\langle x,x\rangle \tag{8.3}$$

is called an *inversion*. For $x\in\{y\in V/y$ and $T_aR(y)\in \mathrm{Dom}\,R\}$ it is easy to see that (Ref. 11, p. 399)

$$K_a=RT_aR. \tag{8.4}$$

A geometrical description of $\mathrm{Dom}\,K_a$ is given in Ref. 12, p. 1089 for diagonal, pseudo-orthogonal $\langle\ ,\ \rangle$ in two dimensions. Let $K(V,\langle\ ,\ \rangle)$ denote the corresponding n-dimensional, commutative subgroup of $\mathrm{Birat}(V)$. Then the group $\mathrm{Kon}(V,\langle\ ,\ \rangle)$ generated by $D\mathrm{Aut}(V,\langle\ ,\ \rangle)$, $K(V,\langle\ ,\ \rangle)$ and the translations in V is called the full *conformal group* of $(V,\langle\ ,\ \rangle)$. It is straightforeward to verify for $u\in\mathrm{Dom}\,K_b$

$$T_bK_u=K_{K_b(u)}T_{b+2\langle b,u\rangle b-\langle b,b\rangle u}D_{\omega(b,u)}S_{K_b(u)}S_u \tag{8.5}$$

from which one deduces the multiplication law

$$\begin{aligned}&(K_aT_bM)(K_cT_d\tilde{M})\\&=K_{K_b(u)+a}T_{b+2\langle b,u\rangle b-\langle b,b\rangle u+D_{\omega(b,u)}S_{K_b(u)}S_uMd}D_{\omega(b,u)}S_{K_b(u)}\\&\times S_uM\tilde{M},\end{aligned} \tag{8.6}$$

where $M,\tilde{M}\in D\mathrm{Aut}(V,\langle\ ,\ \rangle)$, $M=\alpha_MM_0$, $0\neq\alpha_M\in\mathbb{R}$, $M_0\in\mathrm{Aut}(V,\langle\ ,\ \rangle)$ and $u=(\alpha_M)^{-1}M_0c$. The inverse of K_aT_bM is

$$K_{-\alpha_MM_0^{-1}K_b(a)}T_{-M^{-1}S_{K_b(a)}S_aD_{\omega(a,b)}K_a(b)}M^{-1}S_{K_b(a)}S_aD_{\omega(a,b)}. \tag{8.7}$$

Obviously $R\notin\mathrm{Kon}(V,\langle\ ,\ \rangle)$. From

$$\langle K_a(x),K_a(x)\rangle=\omega(a,x)^{-1}\langle x,x\rangle \tag{8.8}$$

follows that light cones are preserved by $K(V,\langle\ ,\ \rangle)$. For positive definite $\langle\ ,\ \rangle$ the angle between two vectors x and y arc $\cos[\langle x,y\rangle(\sqrt{\langle x,x\rangle\langle y,y\rangle})^{-1}]$ is not invariant under special conformal transformations. K_b is a solution of the differential equation

$$\left\langle\frac{\partial K_a(x)}{\partial x}u,v\right\rangle+\left\langle u,\frac{\partial K_a(x)}{\partial x}v\right\rangle=\omega(a,x)^{-2}\langle u,v\rangle \tag{8.9}$$

(Ref. 11, p. 398), and the same type of differential equation with suitable factor on the right-hand side is valid for the other transformations of $\mathrm{Kon}(V,\langle\ ,\ \rangle)$.

There is an immediate generalization of the results of this section to the case of a degenerate symmetric bilinear form τ with the signature (n_1,n_2,n_0), $n_1+n_2+n_0=n$. One can choose a basis of V such that the matrix of τ is $\mathrm{diag}(1,\ldots,1,1,-1,\ldots,-1,0,\ldots,0)$ with n_1 times 1, n_2 times -1 and n_0 times 0. If the *bilinear kernel* of τ is defined by $\mathrm{Bk}(\tau)=\{x\in V/\tau(x,y)=0$ for all $y\in V\}$, then $V=V_1\oplus V_2\oplus\mathrm{Bk}(\tau)$ with $\dim V_1=n_1$, $\dim V_2=n_2$ and $\dim\mathrm{Bk}(\tau)=n_0$. $\mathrm{Aut}(V,\tau)$ is a $\frac{1}{2}n(n-1)+\frac{1}{2}n_0(n_0+1)$ dimensional group, which in matrix form is

$$\begin{pmatrix}\mathrm{Aut}(V_1\oplus V_2,\tau|_{V_1\oplus V_2}) & 0\\ \mathscr{M} & Gl(\mathrm{Bk}(\tau))\end{pmatrix}, \tag{8.10}$$

$\mathscr{M}$ being the set of arbitrary rectangular $(n_1+n_2)\times n_0$ matrices. The only result which has to be changed in the case of degenerate τ is $\cap_{x\in V}\mathrm{Dom}\,K_x=\mathrm{Bk}(\tau)$. Proof: Trivially $\mathrm{Bk}(\tau)\subset\mathrm{Dom}\,K_x$ for all $x\in V$. On the other hand, given $y\in V_1\oplus V_2$ with $y\in\mathrm{Dom}\,K_a$, it is easy to find $b\in V$ with $y\notin\mathrm{Dom}\,K_b$; namely if $\tau(y,y)\neq0$ one has $y\notin\mathrm{Dom}\,K_{-\tau(y,y)^{-1}y}$, and if $\tau(y,y)=0$ it is always possible to find b with $1+2\tau(y,b)=0$, since the restriction of τ to $V_1\oplus V_2$ is nondegenerate.

IX. THE HOMOMORPHISM OF CONFORMAL AND PSEUDO-ORTHOGONAL GROUPS

Following (Ref. 13, p. 16) it is possible to determine the Lie group data of the full conformal group.

Given $\tilde{x}=\xi_0\oplus x\oplus\xi_{n+1}\in\mathbb{R}\oplus V\oplus\mathbb{R}=:\tilde{V}$, the $n+2$ dimensional vector space $\tilde{V}$ can be equipped with an indefinite bilinear from $\prec,\succ$, whose matrix is $\mathrm{diag}(1,I,-1)$. Also let be $\mathscr{D}=\{\tilde{x}\in\tilde{V}/\prec\tilde{x},\tilde{x}\succ=0$ and $\xi_{n+1}-\xi_0\neq0\}$. The mapping

$$\Gamma:\tilde{x}\to x/(\xi_{n+1}-\xi_0),\quad \Gamma:\mathscr{D}\to V \tag{9.1}$$

is surjective with the fiber over $y\in V$ $\Gamma^{-1}(y)=\{\tilde{x}\in\mathscr{D}/x=(\xi_{n+1}-\xi_0)y\}$. If x is any element of $\mathscr{D}$, then

$$0=\prec\tilde{x},\tilde{x}\succ=(\xi_{n+1}-\xi_0)[(\xi_{n+1}-\xi_0)\langle\Gamma(\tilde{x}),\Gamma(\tilde{x})\rangle-(\xi_{n+1}+\xi_0)]$$

implies that $\xi_{n+1}+\xi_0=(\xi_{n+1}-\xi_0)\langle\Gamma(\tilde{x}),\Gamma(\tilde{x})\rangle$. Hence for $\tilde{x},\tilde{y}\in\mathscr{D}$ with $\Gamma(\tilde{x})=\Gamma(\tilde{y})$

$$\xi_{n+1}+\xi_0=[(\xi_{n+1}-\xi_0)/(\eta_{n+1}-\eta_0)](\eta_{n+1}-\eta_0),$$

from which $\xi_{n+1}=(\xi_{n+1}-\xi_0)(\eta_{n+1}-\eta_0)^{-1}\eta_{n+1}$. The same relation holds for ξ_0. Therefore

$$\Gamma(\tilde{x})=\Gamma(\tilde{y})\ \text{iff}\ \tilde{x}=[(\xi_{n+1}-\xi_0)/(\eta_{n+1}-\eta_0)]\tilde{y}. \tag{*}$$

Given $\tilde{A}\in\mathrm{Aut}(\tilde{V},\prec,\succ)$ with $\tilde{A}\tilde{x}\in\mathscr{D}$, $\Gamma(\tilde{x})=\Gamma(\tilde{y})$ implies that $\Gamma(\tilde{A}\tilde{x})=\Gamma[(\xi_{n+1}-\xi_0)(\eta_{n+1}-\eta_0)^{-1}\tilde{A}\tilde{y}]$, and this is equal to $\Gamma(\tilde{A}\tilde{y})$ because the arguements lie in the same fiber. Thus the group homomorphism $\Gamma:\tilde{A}\to\Gamma(\tilde{A})$, $\Gamma:\mathrm{Aut}(V,\prec,\succ)\to\mathrm{Birat}(V)$, given by

$$\Gamma(\tilde{A})\,\Gamma(\tilde{x})=\Gamma(\tilde{A}\tilde{x}) \tag{9.2}$$

is well defined. To show that Γ is a surjective homomorphism on the connectivity component of the identity in the conformal group (this component will be denoted in the following by an index 0) one has to consider the Lie algebra of infinitesimal transformations of $\mathrm{Aut}(\tilde{V},\prec,\succ)$, which is $\mathrm{der}(V,\prec,\succ)$. Every element of $\mathrm{der}(V,\prec,\succ)$ has the form

$$\begin{aligned}&\delta\tilde{K}_a+\delta\tilde{T}_b+\delta\tilde{M}+\delta\tilde{D}_\gamma\\&:=\begin{pmatrix}0&-a^TI&0\\a&0&a\\0&a^TI&0\end{pmatrix}+\begin{pmatrix}0&b^TI&0\\-b&0&b\\0&b^TI&0\end{pmatrix}+\begin{pmatrix}0&0&0\\0&\delta M&0\\0&0&0\end{pmatrix}+\begin{pmatrix}0&0&\gamma\\0&0&0\\\gamma&0&0\end{pmatrix},\end{aligned} \tag{9.3}$$

where $\gamma\in\mathbb{R}$ and $\delta M\in\mathrm{der}(V,\langle\ ,\ \rangle)$, i.e., in the chosen basis $(\delta M)^TI+I\,\delta M=0$. The corresponding one-parameter subgroups are

$$\exp(\mu\delta\tilde{K}_a)=\begin{pmatrix}1-\frac{1}{2}\mu^2\langle a,a\rangle & -\mu a^TI & -\frac{1}{2}\mu^2\langle a,a\rangle\\ \mu a & \mathrm{id}_V & \mu a\\ \frac{1}{2}\mu^2\langle a,a\rangle & \mu a^TI & 1+\frac{1}{2}\mu^2\langle a,a\rangle\end{pmatrix}, \tag{9.4}$$

$$\exp(\mu\delta\tilde{T}_b)=\begin{pmatrix}1-\frac{1}{2}\mu^2\langle b,b\rangle & \mu b^TI & \frac{1}{2}\mu^2\langle b,b\rangle\\ -\mu b & \mathrm{id}_V & \mu b\\ -\frac{1}{2}\mu^2\langle b,b\rangle & \mu b^TI & 1+\frac{1}{2}\mu^2\langle b,b\rangle\end{pmatrix}, \tag{9.5}$$

(Ref. 14, p. 856),

$$\exp(\mu\delta\tilde{M})=\begin{pmatrix}1 & 0 & 0\\ 0 & \exp(\mu\delta M) & 0\\ 0 & 0 & 1\end{pmatrix}, \tag{9.6}$$

$$\exp(\mu\delta\tilde{D}_\gamma)=\begin{pmatrix}\cosh\mu\gamma & 0 & \sinh\mu\gamma\\ 0 & \mathrm{id}_V & 0\\ \sinh\mu\gamma & 0 & \cosh\mu\gamma\end{pmatrix}. \tag{9.7}$$

Every element of $\mathrm{Aut}_0(\tilde{V}, \lhd, \rhd)$ is a product of such elements. Using the relations of the beginning of this section, it is to prove, for $\exp(\delta\tilde{K}_a) =: \tilde{K}_a$, $\exp(\delta\tilde{T}_b) =: \tilde{T}_b$, $\exp(\delta\tilde{M}) =: \tilde{M}$, and $\exp(\delta\tilde{D}_\gamma) =: \tilde{D}_{\exp\gamma}$, that

$$\Gamma(\tilde{K}_a)=K_a,\quad \Gamma(\tilde{T}_b)=T_b,\quad \Gamma(\tilde{M})=M,\quad \Gamma(\tilde{D}_\gamma)=D_\gamma. \tag{9.8}$$

Thus Γ maps $\mathrm{Aut}_0(\tilde{V}, \lhd, \rhd)$ onto $\mathrm{Kon}_0(V, \langle\ ,\ \rangle)$, if the latter group is equipped with the finest topology such that Γ is continuous. From (*) it is easy to see that the kernel of Γ consists of all multiples of $\mathrm{id}_{\tilde{V}}$ in $\mathrm{Aut}_0(V, \lhd, \rhd)$, i.e., of $\pm\mathrm{id}_{\tilde{V}}$ or $\mathrm{id}_{\tilde{V}}$ only, depending whether $-\mathrm{id}_{\tilde{V}} \in \mathrm{Aut}_0(V, \lhd, \rhd)$ or not. This obviously depends on dimension and signature of $(V, \langle\ ,\ \rangle)$. In case n_1 and n_2 are both odd, it is thus proved

$$\{\mathrm{id}_V\} \to \{\pm\mathrm{id}_{\tilde{V}}\} \to \mathrm{Aut}_0(\tilde{V}, \lhd, \rhd) \xrightarrow{\Gamma} \mathrm{Kon}_0(V, \langle\ ,\ \rangle) \to \{\mathrm{id}_V\}. \tag{9.9}$$

This for instance is the case for V being the Minkowski space. In the other cases $\mathrm{Aut}_0(\tilde{V}, \lhd, \rhd)$ and $\mathrm{Kon}_0(V, \langle\ ,\ \rangle)$ are isomorphic. So for instance the proper orthochronous Lorentz group is isomorphic to the connectivity component of the identity of the conformal group of a two-dimensional Euclidean vector space.

The square $n+2$ matrix

$$\tilde{R} := \begin{pmatrix}-1 & 0 & 0\\ 0 & \mathrm{id}_V & 0\\ 0 & 0 & 1\end{pmatrix}$$

is in $\mathrm{Aut}(\tilde{V}, \lhd, \rhd)$ but not in $\mathrm{Aut}_0(V, \lhd, \rhd)$. Hence $\Gamma(\tilde{R}) = R$ shows that R is not in $\mathrm{Kon}_0(V, \langle\ ,\ \rangle)$ and that $\mathrm{Kon}_0(V, \langle\ ,\ \rangle) \cup R\,\mathrm{Kon}_0(V, \langle\ ,\ \rangle)$ is a $\frac{1}{2}(n+2)(n+1)$ dimensional Lie group of two connectivity components.

X. THE LIE ALGEBRA OF INFINITESIMAL CONFORMAL TRANSFORMATIONS

A complete set of one-parameter subgroups of $\mathrm{Kon}_0(V, \langle\ ,\ \rangle)$ is given by $\mu \mapsto K_{\mu b}$ (special conformal), $\mu \mapsto T_{\mu b}$ (translations), $\mu \mapsto D_{\exp\mu\gamma}$ (dilatations), and the pseudo-orthogonal (4.12). From (3.1) follows

$$\delta K_a(x) = \langle x, x\rangle a - 2\langle a, x\rangle x, \tag{10.1}$$

$$\delta T_b(x) = b, \tag{10.2}$$

$$\delta D_{\exp\lambda} = D_\lambda, \quad \lambda \in \mathbb{R} \text{ arbitrary} \tag{10.3}$$

$$\delta\mathrm{Aut}(V, \langle\ ,\ \rangle) = \mathrm{der}(V, \langle\ ,\ \rangle). \tag{10.4}$$

δT_b is a nonlinear transformation homogeneous of degree zero, δD_γ and $\mathrm{der}(V, \langle\ ,\ \rangle)$ are homogeneous of degree one, and δK_a is homogeneous of degree two. One verifies the differential equation

$$\left\langle \frac{\partial\delta K_a(x)}{\partial x}u, v\right\rangle + \left\langle u, \frac{\partial\delta K_a(x)}{\partial x}v\right\rangle = -4\langle x, a\rangle\langle u, v\rangle \tag{10.5}$$

and the same type of equation with different factor on the right-hand side, is valid for the other infinitesimal conformal transformations. From

$$\langle\delta K_a(x), \delta K_a(x)\rangle = \langle a, a\rangle\langle x, x\rangle^2 \tag{10.6}$$

follows, that δK_a preserves light cones (as the infinitesimal colineations do as well).

Lie brackets of $\delta\,\mathrm{Kon}(V, \langle\ ,\ \rangle) = \mathrm{kon}(V, \langle\ ,\ \rangle)$ can be calculated from (3.3), (8.5), and (8.6). An easy but tedious calculation gives for $\delta M \in \mathrm{der}(V, \langle\ ,\ \rangle)$,

$$\{\delta K_a, \delta K_b\} = \{\delta T_a, \delta T_b\} = \{D_\lambda, D_\mu\} = \{\delta M, D_\lambda\} = 0, \tag{10.7}$$

$$\{\delta K_a, \delta T_b\} = 2\,o(a\otimes b) - 2D_{\langle a,b\rangle}, \tag{10.8}$$

$$\{\delta K_a, D_\lambda\} = \delta K_{\lambda a}, \tag{10.9}$$

$$\{\delta K_a, \delta M\} = -\delta K_{\delta M a}, \tag{10.10}$$

$$\{\delta T_a, D_\lambda\} = -\delta T_{\lambda a}, \tag{10.11}$$

$$\{\delta T_a, \delta M\} = -\delta T_{\delta M a}, \tag{10.12}$$

and the commutation relations (4.14) of $\mathrm{der}(V, \langle\ ,\ \rangle)$. The matrices $\delta\tilde{K}_a$, $\delta\tilde{T}_b$, $\delta\tilde{M}$, and $\delta\tilde{D}_\gamma$ have the same commutation relations. Thus the class of conformal Lie algebras of a (not necessarily pseudo-) orthogonal vector space is isomorphic to the class of (proper) pseudo-orthogonal Lie algebras in a vector space with two additional dimensions.

If $\mathrm{der}(\tilde{V}, \lhd, \rhd)$ is equipped with a bilinear form $\langle A, B\rangle := \frac{1}{4}\,\mathrm{Spur}\,A\,B$, then

$$\begin{aligned}\tfrac{1}{2}[[\delta\tilde{K}_a, \delta\tilde{T}_x], \delta\tilde{T}_x] &= \langle x, x\rangle\delta\tilde{T}_a - 2\langle a, x\rangle\delta\tilde{T}_x\\ &= \langle\delta\tilde{K}_x, \delta\tilde{T}_x\rangle\,\delta\tilde{T}_a - 2\langle\delta\tilde{K}_a, \delta\tilde{T}_x\rangle\,\delta\tilde{T}_x\\ &= \delta K_{\delta\tilde{T}_a}(\delta\tilde{T}_x),\end{aligned} \tag{10.13}$$

$$[\delta\tilde{M}, \delta\tilde{T}_x] = \mathrm{ad}(\delta\tilde{M})\,\delta\tilde{T}_x = \delta\tilde{T}_{\delta M a}, \tag{10.14}$$

$$\mathrm{ad}(\delta\tilde{T}_a)^0\delta\tilde{T}_x = \delta\tilde{T}_x = \delta T_{\delta\tilde{T}_a}(\delta\tilde{T}_x). \tag{10.15}$$

It is clear that the decompositions (6.3) and (9.3) are symmetric in the sense of M. Koecher (Ref. 7, Chap. II, Sec. 5) with the corresponding -1 graduation $P_0 \oplus P_2 \oplus P_1$ (Ref. 7, p. 5). Equations (7.9), (7.10) (7.11) and (10.13) (10.14) (10.15) give the relation to the underlying binary Lie algebras $\mathrm{col}(V, \langle\ ,\ \rangle)$ and $\mathrm{kon}(V, \langle\ ,\ \rangle)$.

ACKNOWLEDGMENTS

The author is indepted to Professor H. D. Doebner for useful discussions, to H. J. Schmidt for discussions on Secs. 2 and 3, and to L. Castell for remarks on colineations. This work was supported in part by the

Deutsche Forschungsgemeinschaft and is a part of the authors thesis at the University of Marburg.

[1]F. Pham, Ann. Inst. Henri Poincare A **6,** 89 (1967).
[2]Ngo Van Que, Ann. Inst. Fourier **17,** 157 (1967).
[3]M. Koecher, *An Elementary Approach to Bounded Symmetric Domains* (Rice U. P., Houston, 1969).
[4]M. Koecher, Math. Z. **109,** 349 (1969).
[5]I. Kaplansky, *Lie Algebras and Locally Compact Groups* (Chicago U. P., Chicago, 1971).
[6]O. Loos, *Symmetric Spaces I: General Theory* (Benjamin, New York, 1969).
[7]S. Helgason, *Differential Geometry of Symmetric Spaces* (Academic, New York, 1962).
[8]N. Jacobson, *Lie Algebras* (Interscience, New York, 1961).
[9]H. Bacry, *Lecons sur la theorie des groupes et les symetries des particules elementaires* (Gordon and Breach, Paris, 1966).
[10]H. Tilgner, Ann. Inst. Henri Poincare A **13,** 129 (1970).
[11]H. A. Kastrup, Ann. Phys. (Leipz.) **9,** 389 (1962).
[12]J. Wess, Nuovo Cimento **18,** 1086 (1960).
[13]A. J. Clark, thesis (Princeton University, 1968).
[14]G. Pinsky, Nuovo Cimento A 51, 856 (1967).

Faddeev-like equations with multibody forces

M. A. Gregorio*† and D. R. Avalos*

International Centre for Theoretical Physics, Trieste, Italy
(Received 2 May 1972)

Equations for a system of four particles interacting through two-, three-, and four-body potentials are derived using a generalization of Faddeev's technique for the three-body problem. Two alternative approaches are condsidered and simple equations for iterative solutions are written down. The method is generalized to the N-body problem with multibody forces.

I. INTRODUCTION

The N-body problem in potential theory cannot be solved by the Lippmann—Schwinger (L—S) equation[1] whenever the interactions are k-particle potentials, with $k<N$. The reason is quite clear: the kernel of the L—S equation is $VG_0(E)$, where $G_0(E)$ is the resolvent, or Green function, of the free Hamiltonian, and if V is a sum of k-body potentials there will be $N-k$ spectator particles in each term, each one of them giving a delta function when taking matrix elements. Therefore the kernel is unbounded and the corresponding L—S integral equation is singular. This difficulty is still present even if we add to V an N-body interaction.

The simplest of this kind of problem is the three-body system with two-body interactions, solved by Faddeev ten years ago.[2] The method proposed by Faddeev consists in rearranging the Born series for the T-matrix, in effect splitting it into a sum of three separate pieces, each being the sum of a specific subset of terms. He was then able to obtain a system of coupled integral equations with connected kernel for these three parts of the T matrix. Mathematically, this means that he first solved the singular parts of the equation in closed form, recasting the L—S equation into a connected form. The inhomogeneous term and the kernel of the Faddeev equations depend upon the off-energy-shell two-body amplitudes, a property which has been used to propose a practical theory for calculation in three-body systems.[3] This method can be extended to more complicated problems, as has been done, for instance, for the four-body problem with two-body forces.[4]

It is worth mentioning that Faddeev's equations are by no means the only solution which has been proposed for the three-body problem. The interested reader should consult the Brandeis lectures by Amado[5] to find a clear analysis of Faddeev's solution as well as its comparison with several other works on the problem. We refer also to Hepp's lectures[6] for a rigorous proof of the existence of at least one set of Faddeev-like equations for every N-particle system.

Here we want to go even further with Faddeev's method and treat the four-body system with multibody interactions: that is, two-, three-, and four-body potentials.

The consideration of interactions more complicated than the two-body potential has usually been ignored since their effects, if any, would be small as compared with the effects of two-body forces. On top of this clearly lie the great complexities which arise when handling multibody forces in specific calculations of their effects. On the other hand, a formal solution of the N-body problem with multibody forces can easily be obtained: We will show it explicitly for the four-body system, and then give a brief outline of the case of N particles.

The scattering problem for a system interacting through the sum of two or more potentials can always be expressed in different ways corresponding to the possible forms in which the Born series for the T matrix can be rearranged. To select one expression for the final equations depends, in general, on what interactions are considered as small perturbations compared with the rest. We present two alternatives for our problem and give two different sets of equations to solve the four-body system with two-, three-, and four-body forces. In the first case, developed in Sec. 3A, the two- and three-body force effects are all summed up in every possible three-body subsystem; that is, the "mixed" three-body system is solved first, and the four-body equations depend on both that solution and the four-body potential. Then in Sec. 3B we present the alternative solution which corresponds to first solving "pure" four-body problems (that is, with only one kind of interaction) and writing two uncoupled integral equations for the interference effects between the two- and three-body forces. Again, the four-body potential appears explicitly in the final equation. An iterative solution of these equations can bring an approximate estimation of the multibody force effects. A possible generalization of this second alternative is outlined in Sec. 4.

To avoid unnecessary repetitions we preferred to quote some formulas and identities needed in Sec. 2 as well as the well-known Faddeev equations for the three-body problem with two- and three-body forces. Finally, Sec. 5 contains some discussions and our conclusions.

II. THE THREE-BODY EQUATIONS AND SOME FORMULAS

As stated in the introduction, in this section we recast some known results for the two- and three-body problem. To save time, we quote them directly as expressed in the four-body Hilbert space.

We begin by defining the total four-body Hamiltonian. It is given by

$$H=H_0+V+W+Z=H_0+U, \tag{2.1}$$

where H_0, the free Hamiltonian, is merely the kinetic energy term

$$H_0=\sum_{r=1}^{4}\frac{p_r^2}{2m_r} \tag{2.2}$$

and U is the potential term built up of three parts:

$$V=\sum_{r<s}V_{rs}, \quad r,s=1,2,3,4 \tag{2.3}$$

is the total two-body potential, V_{rs} being the interaction between particles r and s;

$$W=\sum_{p<q<r} V_{pqr}, \quad p,q,r=1,2,3,4 \tag{2.4}$$

is the total three-body interaction term, V_{pqr} being the three-body potential acting on particles p, q, and r simultaneously. Finally, Z represents a possible four-body potential, included here mainly to render our treatment complete.

We define the resolvents, or Green's functions[7]:

$$G(s)=(s-H)^{-1}, \tag{2.5a}$$

$$G_0(s)=(s-H_0)^{-1}, \tag{2.5b}$$

$$G_{ij}(s)=(s-H_0-V_{ij})^{-1}, \tag{2.5c}$$

$$G_{ij,kl}(s)=(s-H_0-V_{ij}-V_{kl})^{-1}, \tag{2.5d}$$

$$G^{(2)}_{ijk}(s)=(s-H_0-V_{ij}-V_{ik}-V_{jk})^{-1}, \tag{2.5e}$$

$$G^{(3)}_{ijk}(s)=(s-H_0-V_{ijk})^{-1}, \tag{2.5f}$$

$$G_{ijk}(s)=(s-H_0-V_{ij}-V_{ik}-V_{jk}-V_{ijk})^{-1}, \tag{2.5g}$$

$$G_v(s)=(s-H_0-v)^{-1}, \tag{2.5h}$$

where, in the last formula, v stands for either V, W, or Z or any combination of them. There are many resolvent identities among these Green's functions and we quote only those which will prove useful later:

$$G(s)=G_{ij}(s)+G_{ij}(s)[V_{ik}+V_{jk}+V_{il}+V_{jl}+V_{kl}+W+Z]\cdot G(s), \tag{2.6a}$$

$$G(s)=G_{ij,kl}(s)+G_{ij,kl}(s)[V_{ik}+V_{il}+V_{jk}+V_{jl}+W+Z]\cdot G(s), \tag{2.6b}$$

$$G(s)=G_{ijk}(s)+G_{ijk}(s)[V_{il}+V_{jl}+V_{jk}+V_{kl}+V_{ijl}+V_{jkl}+V_{ikl}+Z]\cdot G(s), \tag{2.6c}$$

$$G(s)=G^{(3)}_{ijk}(s)+G^{(3)}_{ijk}(s)[V+V_{ijl}+V_{ikl}+V_{jkl}+Z]\cdot G(s), \tag{2.6d}$$

$$G_{ijk}(s)=G_0(s)+G_0(s)[V_{ij}+V_{ik}+V_{jk}+V_{ijk}]\cdot G_{ijk}(s), \tag{2.6e}$$

$$G_{ij,kl}(s)=G_0(s)+G_0(s)[V_{ij}+V_{kl}]G_{ij,kl}(s), \tag{2.6f}$$

$$G(s)=G_v(s)+G_v(s)[U-v]G(s), \tag{2.6g}$$

where in (2.6g) v has the same meaning as in Eq. (2.5h).

We now proceed to enumerate several results of two- and three-body systems but recasting them in the four-body Hilbert space. The two-body T matrix for scattering of particles i and j is defined by

$$t_{ij}(s)=V_{ij}+V_{ij}G(s)V_{ij}. \tag{2.7}$$

It satisfies the L–S equation

$$t_{ij}(s)=V_{ij}+V_{ij}G_0(s)t_{ij}(s). \tag{2.8}$$

Furthermore

$$t_{ij}(s)G_0(s)=V_{ij}G(s). \tag{2.9}$$

Its matrix elements are given in terms of the two-body T matrix in two-particle Hilbert space, $\hat{t}_{ij}$, by

$$\langle \mathbf{p}_i,\mathbf{p}_j,\mathbf{p}_k,\mathbf{p}_l|\,t_{ij}(s)\,|\mathbf{p}'_i,\mathbf{p}'_j,\mathbf{p}'_k,\mathbf{p}'_l\rangle = \delta^{(3)}(\mathbf{p}_k-\mathbf{p}'_k)\times\delta^{(3)}(\mathbf{p}_l-\mathbf{p}'_l) \times\langle \mathbf{p}_i,\mathbf{p}_j|\hat{t}_{ij}[s-(p_k^2/2m_k)-(p_l^2/2m_l)]|\mathbf{p}'_i,\mathbf{p}'_j\rangle. \tag{2.10}$$

In a similar manner we can write in this four-particle space the Faddeev equations[2] for scattering of particles i, j, and k, i.e.,

$${}^2T^l_{ij}(s)=t_{ij}(s)+t_{ij}(s)G_0(s)[{}^2T^l_{ik}(s)+{}^2T^l_{jk}(s)]. \tag{2.11}$$

Note that the total three-body T-matrix is given by

$${}^2T^l(s)=\sum_{m<n}{}^2T^l_{mn}(s), \tag{2.12}$$

its matrix elements being functions of three-body space T-matrix, $\hat{T}$,

$$\langle \mathbf{p}_i,\mathbf{p}_j,\mathbf{p}_k,\mathbf{p}_l|\,{}^2T^l(s)\,|\mathbf{p}'_i,\mathbf{p}'_j,\mathbf{p}'_k,\mathbf{p}'_l\rangle = \delta^{(3)}(\mathbf{p}_l-\mathbf{p}'_l)\langle \mathbf{p}_i,\mathbf{p}_j,\mathbf{p}_k|\,\hat{T}[s-(p_l^2/2m_l)]\,|\mathbf{p}'_i,\mathbf{p}'_j,\mathbf{p}'_k\rangle. \tag{2.13}$$

In the preceding equations the superscript l refers to the spectator particle, while the superscript 2 means that two-body forces *only* are being taken into account. Of course, by definition

$${}^2T^l_{ij}(s)=V_{ij}+V_{ij}G^{(2)}_{ijk}(s)[V_{ij}+V_{ik}+V_{jk}]. \tag{2.14}$$

The three-body system, with three-body forces *only*, can be solved directly by the corresponding L–S equation because in this case the equation is connected. That is, by definition

$${}^3T^l(s)=V_{ijk}+V_{ijk}G^{(3)}_{ijk}(s)V_{ijk}, \tag{2.15}$$

where the notation is obvious. ${}^3T^l(s)$ satisfies the L–S equation

$${}^3T^l(s)=V_{ijk}+V_{ijk}G_0(s){}^3T^l(s). \tag{2.16}$$

The three-body problem with "mixed" forces, i.e., with two- and three-body potentials, can be solved in complete analogy to (2.11). To that end one defines

$$T^l_{ij}(s)=V_{ij}+V_{ij}G_{ijk}(s)[V_{ij}+V_{ik}+V_{jk}+V_{ijk}], \tag{2.17a}$$

$$T^l_{ijk}(s)=V_{ijk}+V_{ijk}G_{ijk}(s)[V_{ij}+V_{ik}+V_{jk}+V_{ijk}] \tag{2.17b}$$

and, following the same procedure that yields (2.11) from (2.14), one gets the following system of connected equations:

$$T^l_{ij}(s)=t_{ij}(s)+t_{ij}(s)G_0(s)[T^l_{jk}(s)+T^l_{ik}(s)+T^l_{ijk}(s)], \tag{2.18a}$$

$$T^l_{ijk}(s)=V_{ijk}+V_{ijk}G_0(s)[T^l_{ij}(s)+T^l_{jk}(s)+T^l_{ik}(s)+T^l_{ijk}(s)]. \tag{2.18b}$$

The total three-body T matrix is given by

$$T^l(s)=T^l_{ijk}(s)+\sum_{m<n}T^l_{mn}(s). \tag{2.19}$$

In Sec. 4 we will express this solution in another way.

Now, besides the two- and three-body equations needed for the four-body problem, if one wants connected equations, there is another problem to be solved first. This corresponds to the T matrix for particles (i,j) and (k,l) when only these pairs interact independently of each other; in other words, the four-body T matrix when only the potentials V_{ij} and V_{kl} are acting. This is defined by

$$X_{ij,kl}(s)=(V_{ij}+V_{kl})+(V_{ij}+V_{kl})G_{ij,kl}(s)(V_{ij}+V_{kl}). \tag{2.20}$$

It can be shown[4,5] that, making the decomposition

$$X_{ij,kl}(s)=X_{ij}(s)+X_{kl}(s) \tag{2.21}$$

with

$$X_{ij}(s)=V_{ij}+V_{ij}G_{ij,kl}(s)(V_{ij}+V_{kl}), \tag{2.22}$$

it is possible to get as solution for Eq. (2.20):

$$X_{ij}(s)=t_{ij}(s)+t_{ij}(s)G_0(s)X_{kl}(s) \tag{2.23}$$

[and a similar equation for $X_{kl}(s)$] entirely in terms of $t_{ij}(s)$ and $t_{kl}(s)$.

This completes the preliminary results that we shall need in the following sections. We emphasize that all the foregoing equations are already cast in the four-body Hilbert space, so that the relationships of the matrix elements of the T-matrix to those defined in the two- and three-body spaces are given by expressions similar to those in Eqs. (2.10) and (2.13).

III. THE FOUR-BODY PROBLEM

A. Four-body equations with two-, three-, and four-body forces

In this section we solve the general four-body problem, that is, with two-, three-, and four-body potentials, according to Faddeev's general procedure. That is to say, we set up connected equations for the T matrix in such a way that the input in these equations is given by the off-energy-shell two- and three-body T matrices, the equations for which we have already discussed in the preceding section.

We begin by defining the four-body T matrix by

$$\tau(s)=U+UG(s)U, \tag{3.1}$$

where U and $G(s)$ are given in Eqs. (2.1) and (2.5a), respectively. This satisfies the L—S equation

$$\tau(s)=U+UG_0(s)\tau(s). \tag{3.2}$$

Furthermore, it is possible to show that

$$\tau(s)G_0(s)=UG(s). \tag{3.3}$$

In order to obtain connected equations which would replace (3.2), we split the T matrix into 11 pieces:

$$\tau(s)=\sum_{i<j}\tau_{ij}(s)+\sum_{i<j<k}\tau_{ijk}(s)+\tau_Z(s), \tag{3.4}$$

with

$$\tau_{ij}(s)=V_{ij}+V_{ij}G(s)U, \tag{3.5a}$$

$$\tau_{ijk}(s)=V_{ijk}+V_{ijk}G(s)U, \tag{3.5b}$$

$$\tau_Z(s)=Z+ZG(s)U. \tag{3.5c}$$

Our aim is to write coupled integral equations for the τ_{ij}'s, τ_{ijk}'s, and τ_Z's. To this end we proceed as follows.

Introducing the identity (2.6a) in (3.5a), we obtain

$$\begin{aligned}\tau_{ij}(s)=&V_{ij}+V_{ij}G_{ij}(s)V_{ij}+V_{ij}G_{ij}(s)[V_{ik}+V_{jk}+V_{ijk}]\\&+V_{ij}G_{ij}(s)[V_{il}+V_{jl}+V_{ijl}]+V_{ij}G_{ij}(s)V_{kl}\\&+V_{ij}G_{ij}(s)[V_{ikl}+V_{jkl}]+V_{ij}G_{ij}(s)[V_{ik}+V_{il}\\&+V_{jl}+V_{kl}+V_{jk}+W+Z]G(s)U.\end{aligned} \tag{3.6}$$

replacing now the $G(s)$ of the last term by the resolvent identity (2.6b), the total $G(s)$ reappears, and now it has to be replaced by (2.6c) with indices (ijk). Finally $G(s)$ is again replaced by (2.6c) but now with indices (ijl). After this is done, we can regroup the terms keeping in mind the definition (2.7) and using the identity (2.9). In this way we get from (3.6):

$$\begin{aligned}\tau_{ij}(s)=&t_{ij}(s)\\&+t_{ij}(s)G_0(s)[V_{ik}+V_{jk}+V_{ijk}]\{1+G_{ijk}(s)[V_{ij}+V_{jk}\\&+V_{ik}+V_{ijk}]\}+t_{ij}(s)G_0(s)[V_{il}+V_{jl}+V_{ijl}]\\&\times\{1+G_{ijl}(s)[V_{ij}+V_{jl}+V_{il}+V_{ijl}]\}\\&+t_{ij}(s)G_0(s)V_{kl}\{1+G_{ij,kl}(s)[V_{ij}+V_{kl}]\}\\&+t_{ij}(s)G_0(s)\{[V_{ikl}+V_{ikl}G(s)U]+[V_{jkl}+V_{jkl}G(s)U]\\&+[Z+ZG(s)U]\}+t_{ij}(s)G_0(s)[V_{ik}+V_{jk}+V_{ijk}]G_{ijk}(s)\\&\times\{V_{il}+V_{il}G(s)U+V_{jl}+V_{jl}G(s)U+V_{kl}+V_{kl}G(s)U\\&+V_{ijl}+V_{ijl}G(s)U+V_{ikl}+V_{ikl}G(s)U+V_{jkl}+V_{jkl}G(s)U\\&+Z+ZG(s)U\}+t_{ij}(s)G_0(s)[V_{il}+V_{jl}+V_{ijl}]G_{ijl}(s)\\&\times\{V_{ik}+V_{ik}G(s)U+V_{jk}+V_{jk}G(s)U+V_{kl}+V_{kl}G(s)U\\&+V_{ijk}+V_{ijk}G(s)U+V_{ikl}+V_{ikl}G(s)U+V_{jkl}\\&+V_{jkl}G(s)U+Z+ZG(s)U\}+t_{ij}(s)G_0(s)V_{kl}G_{ij,kl}(s)\\&\times\{V_{ik}+V_{ik}G(s)U+V_{jk}+V_{jk}G(s)U+V_{il}+V_{il}G(s)U\\&+V_{jl}+V_{jl}G(s)U+W+WG(s)U\\&+Z+ZG(s)U\}.\end{aligned} \tag{3.7}$$

To simplify this equation we use several definitions and identities. First, the sum inside the curly brackets of the last four terms can be replaced by the corresponding operators τ_{lm}, τ_{pqr}, and τ_Z, defined in (3.5a)—(3.5c), respectively. The factor multiplying these curly brackets should be compared with the second, third, and fourth inhomogeneous terms because they can all be simultaneously transformed. Let us consider, for instance, the second term. Using the identity (2.6e) to transform the curly bracket into $G_{ijk}(s)G_0^{-1}(s)$, we can write the said term in the form

$$t_{ij}(s)G_0(s)[V_{ik}+V_{jk}+V_{ijk}]G_{ijk}(s)G_0^{-1}(s)$$

[note that with the exception of $G_0^{-1}(s)$, this is the same factor which appears as kernel in the sixth term]. But using Eqs. (2.17a), (2.17b) and (2.18a), (2.18b), we find that this term is equal to

$$T_{ij}^{lc}(s)=T_{ij}^{l}-t_{ij},$$

that is, the connected part of the ij component of the three-body T matrix for the mixed problem with particles ijk involved [recalling the observation made above, the corresponding kernel will be $T_{ij}^{lc}(s)G_0(s)$]. The same analysis can be made with the fourth inhomogeneous term and we find it to be equal to $X_{ij}^{c}(s)=X_{ij}(s)-t_{ij}$ [the kernel being $X_{ij}^{c}(s)G_0(s)$]; therefore, Eq. (3.7) can finally be recast in the form

$$\begin{aligned}\tau_{ij}(s)=&t_{ij}(s)+T_{ij}^{lc}(s)+T_{ij}^{kc}(s)+X_{ij}^{c}(s)\\&+t_{ij}(s)G_0(s)[\tau_{ikl}(s)+\tau_{jkl}(s)+\tau_Z(s)]\\&+T_{ij}^{lc}(s)G_0(s)[\tau_{il}(s)+\tau_{jl}(s)+\tau_{kl}(s)+\tau_{ijl}(s)\\&+\tau_{ikl}(s)+\tau_{jkl}(s)+\tau_Z(s)]\end{aligned}$$

$$+T^{hc}_{ij}(s)G_0(s)[\tau_{ik}(s)+\tau_{jk}(s)+\tau_{kl}(s)+\tau_{ijk}(s)$$
$$+\tau_{ikl}(s)+\tau_{jkl}(s)+\tau_Z(s)]$$
$$+X^c_{ij}(s)G_0(s)[\tau_{ik}(s)+\tau_{jk}(s)+\tau_{il}(s)+\tau_{jl}(s)+\tau_{ijk}(s)$$
$$+\tau_{ijl}(s)+\tau_{ikl}(s)+\tau_{jkl}(s)+\tau_Z(s)]. \quad (3.8a)$$

To get the equation for τ_{ijk}, we start from (3.5b). Using the identity (2.6d), one obtains

$$\tau_{ijk}(s)=V_{ijk}+V_{ijk}G^{(3)}_{ijk}(s)V_{ijk}+V_{ijk}G^{(3)}_{ijk}(s)[V+V_{ijl}+V_{jkl}$$
$$+V_{ikl}+Z]+V_{ijk}G^{(3)}_{ijk}(s)[V+V_{ijl}+V_{jkl}+V_{ikl}$$
$$+Z]G(s)U.$$

By now (2.15) and the definitions (3.5a), (3.5b), (3.5c), the preceding equation gives

$$\tau_{ijk}(s)={}^3T^I(s)+{}^3T^I(s)G_0(s) \quad (3.8b)$$
$$\times\sum_{l<m}^{4}\tau_{lm}(s)+\tau_{ijl}(s)+\tau_{ikl}(s)+\tau_{jkl}(s)+\tau_Z(s)\ .$$

Finally the equation for $\tau_Z(s)$ is, from (3.5c) and (3.3),

$$\tau_Z(s)=Z+ZG_0(s)\tau(s). \quad (3.8c)$$

The set (3.8) is the solution to our problem. It has connected kernels, as can be seen by iterating the equations, and it can be verified that the inhomogeneous terms contain the sum of all the disconnected parts of the original L–S equation. Furthermore, they contain as special cases the "pure" four-body problems, i.e., the problems with n-body potentials *only* ($n\leq 4$). For example, if $n=2$, then $W=Z=0$, $\tau_Z(s)=\tau_{ijk}(s)=0$ [for all (ijk)] and (3.8a) reduces to already known expressions given by several authors.[4,5] [In this case T^{Ic}_{ij} becomes ${}^2T^{Ic}_{ij}$, etc., as can be seen from (2.17).] If $n=3$, $V=Z=0$ and $\tau_Z(s)=\tau_{ijk}(s)=0$. [for all (ij)] and there remains only the set (3.8b). Finally, if $V=W=0$, the equations reduce to (3.8c), which is very easy to solve by standard methods.

Note, however, that in order to get connected *and* simple equations, it has been necessary to express them as functions of three-body "mixed" problems [in (3.8a) $T^{Ic}_{ij}(s)$ appears, and by (2.17) it depends on both V and W]. This, of course, can be a drawback if, having already solved the four-body system with pairwise interactions, one wants to estimate corrections arising from three-body forces. It is desirable then to have an alternative approach such that the corrections can be estimated *directly* at the four-body-system level. We now turn to such an approach.

B. Alternative solution

The set of Eqs. (3.8) are the appropriate four-body equations when the mixed three-body problem has already been solved. As an alternative procedure, we can first solve the four-body problem with two-body potential *only*, then the pure four-body problem with three-body forces and, finally, compute the interference of the two kinds of force in terms of the solutions to the pure cases. In this way we shall have a more suitable scheme for an approximate evaluation of multibody force effects.

As before, we define the total T matrix τ by

$$\tau(s)=(V+W+Z)+(V+W+Z)G(s)(V+W+Z), \quad (3.9)$$

but now we decompose τ as

$$\tau(s)=\tau_V(s)+\tau_W(s)+\tau_Z(s), \quad (3.10)$$

where

$$\tau_V(s)=V+VG(s)(V+W+Z), \quad (3.11)$$
$$\tau_W(s)=W+WG(s)(V+W+Z), \quad (3.12)$$
$$\tau_Z(s)=Z+ZG(s)(V+W+Z), \quad (3.13)$$

correspond to (3.4), (3.5a)–(3.5c), respectively. Since the four-body potential Z does not produce any disconnectedness because there are no spectator particles left with respect to it, we first solve the problem with potentials V and W, and then superimpose the four-body-interaction effects.

Putting $Z=0$ in the previous equations, we have

$$\bar{\tau}(s)=\bar{\tau}_V(s)+\bar{\tau}_W(s), \quad (3.14a)$$

where

$$\bar{\tau}_V(s)=V+V\bar{G}(s)(V+W), \quad (3.14b)$$
$$\bar{\tau}_W(s)=W+W\bar{G}(s)(V+W) \quad (3.14c)$$

(the bar over τ and G indicates that the four-body potential has been removed).

Using the resolvent identities (2.6g) for $U=V+W$ and $v=V$ in (3.14b), and then for $v=W$ in (3.14c), we have

$$\bar{\tau}_V(s)=(V+VG_V(s)V)+VG_V(s)[W+W\bar{G}(s)(V+W)],$$
$$\bar{\tau}_W(s)=(W+WG_W(s)W)+WG_W(s)[V+V\bar{G}(s)(V+W)],$$

where $G_V(s)$ and $G_W(s)$ are defined as in (2.5g).

The total four-body T matrix with two-body potentials V is precisely[8]

$$T_V(s)=V+VG_V(s)V, \quad (3.15a)$$

and the four-body T matrix with three-body potentials W is[8]

$$T_W(s)=W+WG_W(s)W. \quad (3.15b)$$

Therefore, putting (3.15a), (3.15b) into the two previous equations, using the usual identity $vG_v(s)=T_v(s)G_0(s)$ and recalling the definitions (3.14b), (3.14c) for $\bar{\tau}_V$ and $\bar{\tau}_W$, we finally obtain

$$\bar{\tau}_V(s)=T_V(s)+T_V(s)G_0(s)\bar{\tau}_W(s), \quad (3.16a)$$
$$\bar{\tau}_W(s)=T_W(s)+T_W(s)G_0(s)\bar{\tau}_V(s). \quad (3.16b)$$

The input of these equations is given by T_V and T_W which are the solutions of the four-body problem with *only one kind* of interaction. Iterating Eqs. (3.16a), (3.16b) we obtain

$$\bar{\tau}_V(s)=T_V(s)[1+G_0(s)T_W(s)]+[T_V(s)G_0(s)T_W(s)G_0(s)]$$
$$\times\bar{\tau}_V(s), \quad (3.16c)$$
$$\bar{\tau}_W(s)=T_W(s)[1+G_0(s)T_V(s)]+[T_W(s)G_0(s)T_V(s)G_0(s)]$$
$$\times\bar{\tau}_W(s). \quad (3.16d)$$

We see that in order to compute the effects of the interference of both kinds of potentials we must, in principle, solve a pair of uncoupled integral equations. However, the three-body-force effects can be approximately estimated, since the kernels of Eqs. (3.16c), (3.16d) are still connected if we replace T_W by its Born

term W. Moreover, when an accurate evaluation of the effects due to W is desired, Eqs. (3.16c), (3.16d) can be the starting point for an iterative solution.

To complete our discussion, we superimpose the four-body force Z. This is easily done if we recall that under the action of Z the pieces $\bar{\tau}_V$ and $\bar{\tau}_W$ will acquire an extra contribution, $\Delta\tau$, due to the interference of $(V+W)$ and Z, that is,

$$\tau_V(s) = \bar{\tau}_V(s) + \Delta\tau_V(s), \tag{3.17a}$$

$$\tau_W(s) = \bar{\tau}_W(s) + \Delta\tau_W(s), \tag{3.17b}$$

and τ_V, τ_W are defined by (3.11) and (3.12), respectively. From these equations, using the resolvent identities (2.6g) for $U=V+W+Z$ and $v=V$ in (3.11), and $v=W$ in (3.12), we obtain

$$\tau_V(s) = T_V(s) + T_V(s)G_0(s)\tau_W(s) + T_V(s)G_0(s)\tau_Z(s),$$

$$\tau_W(s) = T_W(s) + T_W(s)G_0(s)\tau_V(s) + T_W(s)G_0(s)\tau_Z(s),$$

where (3.13), (3.15a), and (3.15b) as well as the usual identity $vG_v(s) = T_V(s)G_0(s)$ have been used. Subtracting the set (3.16a), (3.16b) from the previous equations, we obtain

$$\Delta\tau_V(s) = T_V(s)G_0(s)\tau_Z(s) + T_V(s)G_0(s)\Delta\tau_W(s),$$

$$\Delta\tau_W(s) = T_W(s)G_0(s)\tau_Z(s) + T_W(s)G_0(s)\Delta\tau_V(s),$$

whose solutions are

$$\Delta\tau_V(s) = \bar{\tau}_V(s)G_0(s)\tau_Z(s),$$

$$\Delta\tau_W(s) = \bar{\tau}_W(s)G_0(s)\tau_Z(s),$$

because, replacing them in the above equations, we get (3.16a) and (3.16b). Substituting this in Eqs. (3.17a) and (3.17b) we obtain

$$\tau_V(s) = \bar{\tau}_V(s)[1 + G_0(s)\tau_Z(s)],$$

$$\tau_W(s) = \bar{\tau}_W(s)[1 + G_0(s)\tau_Z(s)].$$

Therefore, the total four-body T matrix τ, given in (3.10), is

$$\tau(s) = \bar{\tau}(s) + [1 + \bar{\tau}(s)G_0(s)]Z[1 + G_0(s)\tau(s)], \tag{3.18}$$

where the definitions (3.13), together with $G(s)(V+W+Z) = G_0(s)\tau(s)$, and definition (3.14a) have been used.

Again, Eq. (3.18) is suitable for estimating the effects of the four-body potential by an iterative procedure.

IV. GENERALIZATION FOR N-BODY SYSTEMS

The method used in Sec. 3B can be easily generalized for more complicated problems. We outline below such a generalization for the N-body system interacting through two-, three-, ..., up to N-body forces. Although it can be argued that this is merely an academic problem, it is quoted here for the sake for completeness.

If V_k is the k-body potential, which is the sum of $N!/[k!(N-k)!]$ terms, the N-body T matrix for V_k is defined by

$$t_k(s) = V_k + V_k G_k(s) V_k, \tag{4.1}$$

where $G_k(s) = [s - H_0 - V_k]^{-1}$. Later on we shall make some comments about the complexities which arise when one tries to sum the disconnected parts of Eq. (4.1).

The N-body T matrix for the sum of two-, three-, ..., λ-body potentials will be denoted by $\tau^{(\lambda)}(s)$. Our aim is to get a relation between $\tau^{(\lambda)}$ and $\tau^{(\lambda-1)}$ for $\lambda < N$ in such a way that, given $\tau^{(\lambda-1)}$ and t_λ, one is able to compute $\tau^{(\lambda)}$.

The case $\lambda = N$ is considered separately since an equation with a connected kernel can be written using V_N only.

By definition

$$\tau^{(\lambda)}(s) = \sum_{k=2}^{\lambda} V_k + \sum_{k=2}^{\lambda} V_k \; G^{(\lambda)}(s)\left(\sum_{k=2}^{\lambda} V_k\right) \tag{4.2}$$

and satisfies the L—S equation

$$\tau^{(\lambda)}(s) = \sum_{k=2}^{\lambda} V_k + \sum_{k=2}^{\lambda} V_k G_0(s)\tau^{(\lambda)}(s). \tag{4.3}$$

In Eqs. (4.2) and (4.3)

$$G^{(\lambda)}(s) = \left[s - H_0 - \sum_{k=2}^{\lambda} V_k\right]^{-1}$$

and

$$G_0(s) = [s - H_0]^{-1}, \quad H_0 = \sum_{r=1}^{N} p_r^2/(2m_r)$$

are the usual resolvents for the total and free Hamiltonians, respectively. Defining

$$\tau_k^{(\lambda)}(s) = V_k + V_k G^{(\lambda)}(s)\left(\sum_{k=2}^{\lambda} V_\lambda\right), \tag{4.4}$$

we see that

$$\tau^{(\lambda)}(s) = \sum_{k=2}^{\lambda} \tau_k^{(\lambda)}(s). \tag{4.5}$$

Using now the identity

$$G^{(\lambda)}(s) = G_k(s) + G_k(s)\left[\sum_{\substack{m=2 \\ m\neq k}}^{\lambda} V_m\right] G^{(\lambda)}(s)$$

in Eq. (4.4), we have

$$\tau_k^{(\lambda)}(s) = (V_k + V_k G_k(s) V_k) + V_k G_k(s)\left[\sum_{\substack{l=2 \\ l\neq k}}^{\lambda} V_l + \left(\sum_{\substack{l=2 \\ l\neq k}}^{\lambda} V_l\right) G^{(\lambda)}(s)\left(\sum_{l=2}^{\lambda} V_l\right)\right].$$

Using definitions (4.1) and (4.4), together with the L—S equation for $t_k(s)$ in the form $V_k G_k(s) = t_k(s)G_0(s)$, the last equation reduces to

$$\tau_k^{(\lambda)}(s) = t_k(s) + t_k(s)G_0(s)\left[\sum_{\substack{l=2 \\ l\neq k}}^{\lambda} \tau_l^{(\lambda)}(s)\right], \quad k = 2, 3, \ldots, \lambda. \tag{4.6}$$

Now, in order to relate $\tau^{(\lambda)}$ to $\tau^{(\lambda-1)}$, we write

$$\tau_k^{(\lambda)} = \tau_k^{(\lambda-1)} + \Delta\tau_k, \tag{4.7a}$$

where $\Delta\tau_k$ corresponds to the modification of each $\tau_k^{(\lambda-1)}$ due to the presence of the added λ-body potential. Therefore, using Eq. (4.6) for λ and $\lambda-1$, and substracting them, we obtain

$$\Delta\tau_k = t_k(s)G_0(s)\tau_\lambda^{(\lambda)}(s) + t_k(s)G_0(s)\left(\sum_{\substack{l=2 \\ l\neq k}}^{\lambda-1} \Delta\tau_l\right). \tag{4.7b}$$

Comparison of Eq. (4.7b) and Eq. (4.6) for $(\lambda-1)$ suggests the ansatz

$$\Delta\tau_k=\tau_k^{(\lambda-1)}(s)G_0(s)\tau_\lambda^{(\lambda)}(s), \tag{4.7c}$$

and it is easily verified that (4.7c) is a solution of Eq. (4.7b), because when $\Delta\tau_k$, as given above, is replaced in Eq. (4.7b) we get Eq. (4.6) for $\lambda-1$. Therefore, replacing (4.7c) in (4.7a) we obtain

$$\tau_k^{(\lambda)}(s)=\tau_k^{(\lambda-1)}(s)[1+G_0(s)\tau_\lambda^{(\lambda)}(s)], \quad k=2,3,\ldots,\lambda-1, \tag{4.8}$$

which can be summed from $k=2$ to $k=\lambda-1$, giving

$$\tau^{(\lambda)}(s)=\tau_\lambda^{(\lambda)}(s)+\tau^{(\lambda-1)}(s)[1+G_0(s)\tau_\lambda^{(\lambda)}(s)]. \tag{4.9}$$

Of course, from (4.6),

$$\tau_\lambda^{(\lambda)}(s)=t_\lambda(s)+t_\lambda(s)G_0(s)[\tau^{(\lambda)}(s)-\tau_\lambda^{(\lambda)}(s)]. \tag{4.10}$$

Inserting Eq. (4.10) into (4.9), we obtain

$$\tau^{(\lambda)}(s)=\tau^{(\lambda-1)}(s)[1+\tau^{(\lambda-1)}(s)G_0(s)]t_\lambda(s) \times[1+G_0(s)\tau^{(\lambda)}(s)-G_0(s)\tau_\lambda^{(\lambda)}(s)].$$

But at this stage we can introduce the definition (4.4) of $\tau_\lambda^{(\lambda)}$ in terms of V_λ, because, as is easily verified, the kernel of the resulting equation is connected, since any delta function coming from V_λ is absorbed by $t_\lambda(s)$. Therefore the final expression is

$$\begin{aligned}\tau^{(\lambda)}(s)=\tau^{(\lambda-1)}(s)&+[1+\tau^{(\lambda-1)}(s)G_0(s)]t_\lambda(s)[1-G_0(s)V_\lambda]\\&+[1+\tau^{(\lambda-1)}(s)G_0(s)]t_\lambda(s)[1-G_0(s)V_\lambda]\\&\times G_0(s)\tau^{(\lambda)}(s).\end{aligned} \tag{4.11}$$

Finally, the equation for $\tau^{(N)}(s)$ is easily obtained from (4.9). Then, if we put $\lambda=N$ and use the definition (4.4) for $\tau_N^{(N)}$, we obtain

$$\tau^{(N)}(s)=\tau^{(N-1)}(s)+[1+\tau^{(N-1)}(s)G_0(s)]V_N[1+G_0(s)\tau^{(N)}(s)]. \tag{4.12}$$

Equation (4.11), and its corresponding recurrence throughout the whole problem, requires knowledge of the solution of what we call "pure problems", that is, $t_k(s)$. We have already quoted the corresponding connected equations for $t_k(s)$ in the case $N=4$, $k=2$ and $k=3$. The general case is no more difficult than the four-body problem with two-body forces only. The method should, by now, be quite clear: the $t_k(s)$ matrix is split into $N!/[k!(N-k)!]$ terms, that is, into as many terms as are contained in the potential. Each term satisfies an equation similar to (3.5a) *or* (3.5b). In order to get connected equations one writes down resolvent identities in such number that, by applying them consecutively, one should be able to group terms that correspond to three different kinds of process:

(a) First of all, there will be terms that reproduce the T matrix for k-body interacting via a k-body potential, while the remaining N-k particles are free. [In our four-body problem, this corresponds to $t_{ij}(s)$.]

(b) There will be terms that, grouped together, will be identifiable as solutions for the T matrix for $(N-1)$-body systems with k-body forces *only*. (In our example, they are terms like T_{ij}^{Ic}, etc.)

(c) Finally, there will appear terms that represent the scattering of k particles interacting independently of the remaining $(N-k)$ particles which, in turn, interact among themselves. [This corresponds to the X_{ij}^c term in (3.8a).] If $N-k\geq 2k$, this process can be applied once more to get two groups, both of k particles, interacting independently of each other and of the rest of the $N-2k$ particles. In this manner it is ensured that each piece of the total T-matrix for the pure problem $[t_k(s)]$ will not be coupled to itself and that when the equations are iterated the kernels are connected, i.e., there are no spectator particles left in them.

V. CONCLUSIONS

We have demonstrated that the method proposed by Faddeev for the three-body problem can be applied to more complicated systems. This has already been shown to be the case for the four-body problem with pairwise interactions.[4,5]

The introduction of multiparticle interactions does not present essential difficulties, as we have shown in this article, and equations with connected kernels are derived for the four-body problem with two-, three-, and four-body interactions. These equations depend on the off-energy-shell T-matrices for the two-body problem and for the mixed (i.e., with two- and three-body forces) three-body problem as well as on the four-body potential. However, as indicated in Sec. 3B the three-body force effects can be taken into account at the three-body system level by an equation equivalent to (3.18) so that, instead of depending on the mixed three-body T matrix, the equations effectively depend on the off-energy-shell three-body T matrix with two-body forces only and also on the three-body potential.

Of course, as pointed out before, to attack the problem in this way one must first solve the mixed three-body problem, the effects of three-body forces being reflected in the four-body system through the three-body T-matrices used as input.

We have also treated the problem from another point of view, guided by the idea that in some cases it would be more interesting to have equations to incorporate multibody-force effects directly at the four-body-system level. In this way we get two uncoupled integral equations for the T-matrix. These equations depend on the solution of two four-body "pure" problem, that is, with only two- or three-body forces. The equations account for the interference effects between the two potentials, and can be used to obtain an approximate solution by iterating the equation. Moreover, in such equations either the two- or the three-body force T matrix can be replaced by its corresponding Born approximation, because in this case the kernel is still connected. The inclusion of the four-body potential is always accomplished by the use of an equation like (3.18), which is also suitable for an iterative solution.

Without emphasizing its intrinsic interest and merely for reasons of completeness, we generalized the last described approach to the N-body problem with two-, three-, ..., up to N-body forces. The procedure is to use the "pure" T-matrices (say the T matrix with k-body forces only) and relate the mixed T-matrix which contains effects of multi-body forces up to $(N-1)$-body potential with the mixed T matrix that contains multi-

body-force effects up to $(N-2)$-body potential. These equations contain the pure T matrices and the $(N-1)$-body potentials. They are again well suited for the search of iterative solutions. The N-body potential is incorporated through Eq. (4.12).

Our equations give an immediate idea of the complexities which are encountered when calculating multibody-force effects, due to the large number of coupled multidimensional integral equations which have to be solved. But we think that the solution proposed in Sec. 3B is quite simple if approximate evaluations are to be made. It could be that such equations are useful in evaluations of binding energy in nuclear matter[9] when multibody-force effects are introduced in four-body clusters.

ACKNOWLEDGMENTS

The authors would like to thank Professor Abdus Salam, the International Atomic Energy Agency, and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste.

*Fellow of the Consejo Nacional de Investigaciones Cientificas y Tecnicas of Argentina. Address: Universidad Nacional de La Plata, La Plata, Argentina.

†On leave of absence from Universidad Nacional de La Plata, Argentina.

[1]B. A. Lippmann and J. Schwinger, Phys. Rev. **79,** 469 (1950).

[2]L. D. Faddeev, Zh. Eksp. Teor. Fiz. **39,** 1459 (1960) [Sov. Phys.-JETP **12,** 1014 (1961)]; Dokl. Akad. Nauk SSSR **138,** 565 (1961); Dokl. Akad. Nauk SSSR **145,** 301 (1962) [Sov. Phys.-Dokl. **6,** 384 (1961)]; Sov. Phys.-Dokl. 1, 600 (1962)]. See also L. D. Faddeev *Mathematical Aspects of the Three-Body Problem in Quantum Scattering Theory* (Israel Program for Scientific Translation, Jerusalem, 1965).

[3]C. Lovelace, Phys. Rev. **135,** B1225 (1964). Also C. Lovelace in *Strong Interaction Physics*, edited by R. G. Moorhouse (Oliver and Boyd, London, 1964); E. O. Alt, P. Grassberger, and W. Sandhas, Nucl. Phys. B **2,** 167 (1967).

[4]V. A. Alessandrini, J. Math. Phys. **7,** 215 (1966); see also Ref. 5, Sec. 8 and references contained therein.

[5]R. D. Amado, 1967 Brandeis Summer Institute, published as *Elementary Particles Physics and Scattering Theory*, edited by M. Chretien and S. Schweber (Gordon and Breach, New York, 1970), Vol. 2.

[6]K. Hepp, in *Proceedings of the 1967 International Conference on Particles and Fields*, edited by C. R. Hagen, G. Guralnik, and V. A. Mathur, (Wiley-Interscience, New York, 1967), p. 244. We thank Professor L. Fonda for pointing out this reference to us.

[7]Here and in the rest of Secs. II and III (i, j, k, l) are fixed and stand for any permutation of (1, 2, 3, 4).

[8]T_V corresponds to the solution of Eqs. (3.8a) for the case $W = Z = 0$, that is, with all τ_{ijk}'s and τ_z equal to zero. T_W is the solution of (3.8b) for $V = Z = 0$, that is, with all τ_{ij}'s and τ_z equal to zero.

[9]B. H. J. McKellar and R. Rajaraman, Phys. Rev. C 3, 1877 (1971).

A model unified field theory

R. Mirman

Physics Department, Long Island University, Zeckendorf Campus, Brooklyn, New York 11201
(Received 2 March 1972; revised manuscript received 16 June 1972)

The geometry of space is probed by using a quantum mechanical wavefunction, which allows the introduction of the ordinary curvature tensor, and also provides room for a vector which is identified with the electromagnetic vector potential. The coordinates of space are taken as complex. Thus the electromagnetic field is geometrized and its presence is found to affect the gravitational field equations.

I. INTRODUCTION

That attempts to develop a unified field theory geometrizing both gravitiation and electromagnetism can succeed is rather unlikely, as there is no clear indication of an experimental need for such a theory nor any experimental results that have to be explained. Indeed, the very concept of success in this field is rather undefined.

Nevertheless model unified field theories are of some use in developing our understanding of the concepts involved and of the theory of gravitiation upon which they seek to build. It is for this reason that we present here such a model.

The present study seeks to see what understanding can be gained from a theory that is based on quantum mechanics. In order to probe the properties of space it is customary to introduce a vector (say the spin of a particle) and study how its direction changes under translations. Here the vector that we use as a probe is the wavefunction of a particle. Now under translations not only may the direction change, but the phase also. This allows us to introduce a vector which we can identify with the electromagnetic vector potential.

Thus what we are exploring here is the view that the electromagnetic field is a geometrical property involved in the description of the behavior of quantum mechanical objects in the space—time in which they exist. The classical manifestations of electromagnetism are the limit of quantum phenomena, which can be described phenomenologically on the classical level, but which need a quantum viewpoint in order to be understood geometrically.

Since the quantum mechanical wave function that we use to explore the properties of space is complex, we would expect that a full description of its properties would require that space itself is complex.[1] It is this concept that we wish to study here and to see in more detail why it is suggested by quantum phenomena and how it might be used to geometrize the electromagnetic field.

II. EXPERIMENTAL FOUNDATION FOR THE COMPLEX NATURE OF SPACE-TIME

We wish here to explore the experimental evidence which supports the relevance of the idea that space—time is complex under certain conditions which are as follows. First there must be (certain types) of fields present. In the absence of (all) fields space will be flat and the coordinates real. Second the imaginary parts of the coordinates are relevant only if the quantum nature of the experimental apparatus plays a part in the measuring process (although classically we could measure the related fields but not interpret them in terms of complex coordinates). Third the imaginary parts of the coordinates, like the curvature of space is meaningful only if we replace the fields by the geometry of space—time, that is if we assume that a source produces not a field but a different geometry. Some sources cause space to curve while others cause it to become complex (or to do both).

An example of how such a concept is implied by a geometrical description of quantum phenomena is given by the consideration of how we might measure an interval of time using quantum mechanical objects. We take a one-dimensional beam of particles, all identical and all moving with the same velocity, spread uniformly throughout space, and so described by a plane wave. We now put on a uniform potential V for a time t, assuming that it is done in such a way that the effects from the transition period can be neglected. After the potential has returned to 0 what effect has it had on the wavefunction?

The answer is that the phase has been changed by an amount Vt. This can be measured by splitting the beam in two before the potential is applied (to one half of the beam) and then after the potential returns to zero doing an interference experiment between the two halves. Then knowing V we can find the time t.

This is a one-dimensional situation where only the time coordinate plays a role. And in one real dimension there is no geometry. How then are we to explain the effect on the system using the geometrical properties of time? Only it would seem by assuming that the time coordinate is complex, which could give enough of a geometry to allow the possibility of such an explanation.

What is the corresponding situation in classical mechanics? Actually there is no corresponding one-dimensional situation in classical mechanics, for there we deal with point particles and we must have both space and time dimensions. However quantum mechanically the space dimension can be averaged out by assuming a plane wave so the particle is in a sense spread over all space. Thus only the time dimension is relevant.

Also there is never a force along the single space dimension so no change in its momentum or position, thus nothing to measure and therefore no geometry is needed.

Let us consider the experiment in somewhat more detail. We set up two uniformly distributed equal and opposite infinite sheets of charge parallel to the xy plane,

and some distance away an identical pair parallel to the first with charges reversed. The regions outside the pairs have zero potential while between them the potential has a constant nonzero value.

To do the above experiment we take the particle moving parallel to the x axis and outside the pairs of plates, and then very quickly (so the impulse is almost zero) move one pair past the particle and leave it there for a time, and then move the other pair past it.

There is no force along the x axis and so classically this motion is unaffected. The impulses along the z axis can be made as small as desired, and they can be made equal and opposite. Hence studying the motion of the classical particle will not provide information on the length of time that it spent inside the potential.

This case does not provide any meaningful classical physics. However suppose that the particle were moving at some angle to the sheets of charge. Then there would be forces in the direction of motion and the time it takes to reach a certain point would be different from the time taken by a particle which did not go through the potential. This time delay is related to the time spent inside the potential and so provides information about it.

Quantum mechanically the situation is different. For even if the particle is moving parallel to the sheets there is an effect on it due to its presence in the potential, the change of phase. Hence here the one dimensional case has a measureable physical consequence.

If we consider the particle moving along the z axis then its potential varies as a function of position. We have a square barrier potential which we study as the next example.

In quantum mechanics this can again be viewed as a one-dimensional problem. We take a beam of particles spread out through space and having no time dependence, so the time dimension does not enter the problem, and measure its momentum outside the barrier, the change of intensity in passing through the barrier, and the change of phase. These three numbers are related but two of them are independent. Thus to the point at the end of the barrier we must assign two numbers, but there is only one dimension.

Of course, this is not surprising for the two numbers are the width and height of the barrier. But the point is that we cannot use the concept of height of the barrier because we want to replace the field by geometry, and here the geometry is that of a single dimension. Clearly a complex coordinate is again implied.

Classically we would take two particles, measure their (identical) momentum before one enters the barrier and then the time difference between the particle which went through the barrier and the one that did not. This is then repeated at a different energy. The results can again be expressed in terms of two numbers, but here there are two dimensions since the time was brought in explicitly.

How would we do a two-dimensional problem in quantum mechanics? Of course, we might do it by considering wavepackets making the situation analogous to the classical case. However, this would hide the information we wish to gain.

Let us rather consider a beam which is split into four parts. One part goes through a barrier and in addition has a time-dependent potential imposed on it, the second which has only a time-dependent potential, the third goes only through a space-dependent potential, and a fourth which goes through no potential. Then we perform interference experiments on the different subbeams in pairs, and in addition measure the intensities before and after they go through the potentials.

This will give four independent numbers, which are the two potentials and their two durations. Classically however, since the particles which went through the time-dependent potential were unaffected, we get only two numbers from the time delays at different energies which give the two parameters of the space-dependent potential.

So quantum mechanically we have two dimensions but four numbers, which implies that the x and t coordinates are complex with independent phases. In addition, if we attempted to describe the geometry by the curvature of space we would have one number, the curvature scalar, which is not enough.

Although we are assuming that the coordinates are complex, it should be understood that the full complex plane is not being used, for a given field. Rather the imaginary part of the coordinate is a function of the real part (the function depending on the field present).

Thus we assign to a coordinate axis a complex plane, and then replace the coordinate axis by some curve in that plane. If the field is zero the curve becomes the real axis, and a nonzero field pushes the curve up into the plane.

According to the field equations that we postulate below an electric charge results in both complex coordinates and curved space–time, while a gravitational field produces only curved space–time. How does this com pare with our *gedanken experimente*? While we do not analyze the problem in detail to see whether this situation is in accord with what should be expected from experimental results, the equations are not in disagreement with our limited analysis.

We see from the previous discussion on how to realize a square barrier that our analysis holds for an electric field and we should expect that it would produce complex coordinates as our field equations require.

However for the gravitational field, which has only one sign of the charge, we can not do the above experiments. So from our considerations we have no reason to require that the gravitational field produce a complex space–time.

Thus while it is by no means clear that all quantum experiments involving the electromagnetic field can be described in terms of complex coordinates (or even that no classical ones imply such coordinates), our heuristic considerations do suggest further study of this possibility and of the theories which are implied by it.

III. TENSORS

We consider a four-dimensional space of complex numbers z_i and functions defined over it $f(z, z^*)$, where the * denotes complex conjugation. Vectors are defined as transforming according to

$$\overline{V}^\mu = \frac{\partial \overline{z}^\mu}{\partial z^\nu} V^\nu + \frac{\partial \overline{z}^\mu}{\partial z^{*\nu}} V^{*\nu}, \quad \text{contravariant} \tag{III. 1}$$

$$\overline{W}_\mu = \frac{\partial z^\nu}{\partial \overline{z}^\mu} W_\nu + \frac{\partial z^{*\nu}}{\partial \overline{z}^\mu} W^*_\nu, \quad \text{covariant} \tag{III. 2}$$

or as

$$\overline{V}^\mu = \frac{\partial \overline{z}^\mu}{\partial z^\nu} V^\nu, \tag{III. 3}$$

$$\overline{W}_\mu = \frac{\partial z^\nu}{\partial \overline{z}^\mu} W_\nu, \tag{III. 4}$$

where our summation convention is that the indices run from 1 to 4 if the complex term is explicitly indicated [as in Eq. (III. 1, 2)] and from 1 to 8 if it is not indicated [as in Eq. (III. 3, 4)]. Tensors are defined in the usual way as transforming as the products of vectors. Clearly all the usual laws of tensor algebra hold here.

It is convenient (in fact necessary, when we use tensors of second or higher rank) to denote conjugates by underlining their indices. Thus $S^{*\nu} = S^{\underline{\nu}}$, and $T^{\mu\underline{\nu}}$ cannot be written with an asterisk. Also we use $a^\mu_{\underline{\nu}} = \partial \overline{z}^\mu / \partial z^\nu$ and $a^\mu_\nu = \partial \overline{z}^\rho / \partial z^{*\nu}$.

In defining the metric tensor we have to decide whether it is to be real or complex. As we saw above a phase can be defined independently for both the x and the t coordinates thus making it reasonable to take the coordinates as complex. Further, as can be especially seen by considering cases in which there is only one coordinate, it also seems reasonable to consider the distance as complex. Hence taking the metric tensor as real would be in accord with our analysis.

To take it as complex would imply that there are cases in which we would have to consider complex coordinates and real distances (or vice versa). This does not seem unreasonable but no examples are immediately evident. Hence, for reasons of simplicity alone, we limit ourselves to real metrics.

Thus the distance function is

$$ds^2 = g_{\mu\nu} dz^\mu dz^\nu + g_{\mu\underline{\nu}} dz^\mu dz^{\underline{\nu}} + g_{\underline{\mu}\nu} dz^{\underline{\mu}} dz^\nu + g_{\underline{\mu}\underline{\nu}} dz^{\underline{\mu}} dz^{\underline{\nu}}. \tag{III. 5}$$

We see below that the convariant derivative of the metric tensor is zero. Raising and lowering of the indices are operations defined in the usual way and commute with covariant differentiation.

Contraction of indices is defined in the standard manner with the sum going over real and imaginary parts. Thus, for example,

$$A^i B_i = A^1 B_1 + A^2 B_2 + A^{*1} B^*_1 + A^{*2} B^*_2, \tag{III. 6}$$

in two-dimensional space, and from the above transformation law this is a scalar. So contraction reduces the rank of a tensor by 2.

To establish the notation we prove that the contracted product of a second rank tensor and a vector is a vector. Let $S^\mu = T^{\mu\nu} V_\nu + T^{\mu\underline{\nu}} V_{\underline{\nu}}$, then

$$\begin{aligned} \overline{S}^\rho &= \overline{T}^{\rho\sigma} \overline{V}_\sigma + \overline{T}^{\rho\underline{\sigma}} \overline{V}_{\underline{\sigma}} \\ &= (a^\rho_\nu a^\sigma_\omega T^{\nu\omega} + a^\rho_{\underline{\nu}} a^\sigma_\omega T^{\underline{\nu}\omega} + a^\rho_\nu a^\sigma_{\underline{\omega}} T^{\nu\underline{\omega}} + a^\rho_{\underline{\nu}} a^\sigma_{\underline{\omega}} T^{\underline{\nu}\underline{\omega}}) \\ &\quad \times (a^{-1\xi}_\sigma V_\xi + a^{-1\underline{\xi}}_\sigma V_{\underline{\xi}}) + (\sigma \rightharpoonup \underline{\sigma}) \\ &= a^\rho_\nu (T^{\nu\omega} V_\omega + T^{\nu\underline{\omega}} V_{\underline{\omega}}) + a^\rho_{\underline{\nu}} (T^{\underline{\nu}\omega} V_\omega + T^{\underline{\nu}\underline{\omega}} V_{\underline{\omega}}) \\ &= a^\rho_\nu S^\nu + a^\rho_{\underline{\nu}} S^{\underline{\nu}}. \end{aligned} \tag{III. 7}$$

To prove the reverse assume that both S and V are vectors and we wish to show that T is a second-rank tensor. Then

$$\overline{S}^\rho = a^\rho_{\underline{\nu}} S^\nu + a^\rho_\nu S^{\underline{\nu}} = a^\rho_\nu (T^{\nu\omega} V_\omega + T^{\nu\underline{\omega}} V_{\underline{\omega}}) + a^\rho_{\underline{\nu}} (T^{\underline{\nu}\omega} V_\omega + T^{\underline{\nu}\underline{\omega}} V_{\underline{\omega}}) \tag{III. 8}$$

$$= \overline{T}^{\rho\xi} \overline{V}_\xi + \overline{T}^{\rho\underline{\xi}} \overline{V}_{\underline{\xi}} = a^\rho_\nu [T^{\nu\omega} (a^\xi_\omega \overline{V}_\xi + a^{\underline{\xi}}_\omega \overline{V}_{\underline{\xi}} + T^{\nu\underline{\omega}} (a^\xi_{\underline{\omega}} \overline{V}_\xi + a^{\underline{\xi}}_{\underline{\omega}} \overline{V}_{\underline{\xi}})] + (\nu \rightharpoonup \underline{\nu}),$$

and since the V's are arbitrary we can set V and V^* zero in turn to get

$$\overline{T}^{\rho\sigma} = a^\rho_\omega a^\sigma_\nu T^{\omega\nu} + a^\rho_{\underline{\nu}} a^\sigma_\omega T^{\underline{\nu}\omega} + a^\rho_\nu a^\sigma_{\underline{\omega}} T^{\nu\underline{\omega}} + a^\rho_{\underline{\nu}} a^\sigma_{\underline{\omega}} T^{\underline{\nu}\underline{\omega}} \tag{III. 9}$$

which is the required result.

The generalization to higher rank tensors follows the same pattern.

If the bar under an index is not written explicitly it is to be understood, so we may write $A_\mu B^\mu$ for $A_\mu B^\mu + A_{\underline{\mu}} B^{\underline{\mu}}$.

IV. CONNECTIONS

In defining the connections the basic idea is that the displacement of a vector changes not only its direction but also its phase. Thus as in Riemannian space, we assume that when we move a vector along a closed curve the final vector is not in general parallel to the initial one. And in addition we assume also that the phase of the final vector is different, in general, from that of the initial one.

Thus consider the wave function of a particle moved through a one-dimensional potential (a square barrier, say). We take specifically a time-independent beam of particles represented by a plane wave. Then there is a phase difference between the beam going through the potential compared to one which did not go through whose value is given in elementary quantum mechanics texts. Suppose now that we have a mirror far downstream from the potential and reflect the particles back to their original position by a path outside the barrier. Then in the closed circuit there is a total change of phase equal to the phase difference between the beams which did and did not go through the barrier.

Of course this analysis holds whether the wave function is a scalar, vector, or higher rank tensor. However it is usually most convenient to express concepts and formulas in terms of vectors.

To study how to take into account the change of phase we consider a scalar wave function $\phi(x) \exp[i\alpha(x)x]$ displaced an infinitesimal distance dx. It then becomes

$$\phi(x + dx) \exp i\{[\alpha(x + dx)](x + dx) + \chi\, dx\} \tag{IV. 1}$$

$$= \phi(x+dx)[\exp i\alpha(x)x][1+i(\alpha_{,x}x+\alpha+\chi)dx],$$

where the comma denotes differentiation.

We therefore postulate that the transplantation law for a vector is

$$d(\phi^\mu e^{i\alpha}) = dx^\rho \phi^\kappa e^{i\alpha}(\Gamma^\mu_{\rho\kappa} + i\delta^\mu_\kappa \chi_\rho), \tag{IV. 2}$$

and this equation defines the connections.

When a beam of particles passes through a potential not only is there a phase change but an intensity change also, as some particles are reflected.

Or consider a collection of atoms in some excited state and immerse the system in a potential for a period of time short (compared to the half-life) for free atoms. Then the number of atoms in the excited state after the potential will be less than before. These types of behavior can be represented phenomenologically by a complex phase, which here implies a complex connection.

While we postulate that the real part of the connection depends directly on the potential, the complex part is related to it in a very indirect way, and the explicit form depends on the internal structure of the atoms. To find this complex part the Schroedinger equation (or the appropriate generalization) for the atoms (or other system that we are considering) must be solved. Hence this equation must be added to the field equations in order to determine the behavior of the system.

We note that for a scalar

$$d\phi = i\phi\, dx\, \chi, \tag{IV. 3}$$

$$d\phi^* = -i\phi^*\, dx\, \chi^*. \tag{IV. 4}$$

For a tensor the vector transplantation law is generalized in the usual way except that we have to decide whether the phase depends on the number of indices. Suppose that we move a spinning particle through an electric field (slowly, since the connection refers to a virtual transplantation) then the change in energy is independent of the spin. In a magnetic field this need not be so. In addition as we see below the connection for the electromagnetic field depends on the charge density and so explicitly on the wave function, and thus plausibly on the spin. The details of this relationship are not investigated here.

Hence it is not clear from any simple analysis where in our formalism we should put the dependence on the number of indices. We therefore make the simplest assumption and take the χ term of the transplanation law to be independent of the number of indices. We then have, for example,

$$d(\phi^{\mu\nu}e^{i\alpha}) = dx^\rho e^{i\alpha}(\phi^{\omega\nu}\Gamma^\mu_{\rho\omega} + \phi^{\mu\omega}\Gamma^\nu_{\rho\omega} + i\chi_\rho \phi^{\mu\nu}e^{i\alpha}). \tag{IV. 5}$$

We define the covariant derivative of the wave function $\psi_\mu = \phi_\mu e^{i\alpha}$ to be

$$\begin{aligned}\partial_\eta(\phi^\mu e^{i\alpha}) &= e^{i\alpha}(\phi^\mu_{,\eta} - \Gamma^\mu_{\rho\eta}\phi^\rho) + \phi^\mu e^{i\alpha} i(\alpha_{,\eta} - \chi_\eta) \\ &= \psi^\mu_{,\eta} - \Gamma^\mu_{\rho\eta}\psi^\rho - i\chi_\eta\psi^\mu \equiv \psi^\mu_{;\eta},\end{aligned} \tag{IV. 6}$$

and similarly (with the usual change of sign for the Γ term) for the covariant vector.

The connection χ_η is taken as equal to $f(\psi)(\zeta_\eta + i\theta_\eta)$, where f is some function of the wavefunction ψ, determined by the nature of the particular situation. This says essentially that the energy [which we would expect to be related to χ from Eq. (IV. 1) by taking $dx = dt$] depends on the potential times the charge density.

We now consider the commutator of the cross partials

$$\begin{aligned}&\psi^\mu_{;\eta\xi} - \psi^\mu_{;\xi\eta} = \psi^\rho(-\Gamma^\mu_{\rho\eta,\xi} + \Gamma^\mu_{\rho\xi,\eta} - \Gamma^\mu_{\kappa\eta}\Gamma^\kappa_{\rho\xi} + \Gamma^\mu_{\kappa\xi}\Gamma^\kappa_{\rho\eta}) \\ &- \psi^\mu i(\chi_{\eta,\xi} - \chi_{\xi,\eta}) = R^\mu_{\eta\xi\rho}\psi^\rho - iV_{\eta\xi}f\psi^\mu + iW_{\eta\xi}\psi^\mu \\ &+ f(\theta_{\eta,\xi} - \theta_{\xi,\eta})\psi^\mu - (f_{,\xi}\theta_\eta - f_{,\eta}\theta_\xi\psi^\mu\end{aligned} \tag{IV. 7}$$

where

$$R^\mu_{\eta\xi\rho} = -\Gamma^\mu_{\rho\eta,\xi} + \Gamma^\mu_{\rho\xi,\eta} - \Gamma^\mu_{\kappa\eta}\Gamma^\kappa_{\rho\xi} + \Gamma^\mu_{\kappa\xi}\Gamma^\kappa_{\rho\eta}, \tag{IV. 8}$$

$$V_{\eta\xi} = \zeta_{\eta,\xi} - \zeta_{\xi,\eta}, \tag{IV. 9}$$

$$W_{\eta\xi} = \zeta_\eta f_{,\xi} - \zeta_\xi f_{,\eta}, \tag{IV. 10}$$

with R the normal curvature tensor, and we assume that χ for $\psi^\mu_{,\eta}$, $\Gamma\psi$ and for $\chi\psi$ is the same as for ψ.

We identify the electromagnetic vector potential A with ζ so

$$A_\eta = \zeta_\eta \tag{IV. 11}$$

and the electromagnetic field tensor F with V so

$$F_{\eta\xi} = V_{\eta\xi}. \tag{IV. 12}$$

Note that we can perform a gauge transformation to give a zero value of ζ only if V is zero. Thus the vanishing of V (assuming that W vanishes simultaneously) is the condition which allows us to introduce real coordinates. Of course, even if V vanishes, we can always formally introduce complex coordinates.

If $f = 0$ then in the usual way the vanishing of R is the requirement that we be able to introduce a set of flat coordinates. Again, even if R vanishes we can always introduce coordinates whose metric is non-Minowskian.

However if f is not zero then we cannot introduce flat coordinates even if R vanishes. The electromagnetic field causes space to curve as well as the gravitational. However its effect is rather strange coming through the system-dependent θ's, which are not directly related to the field variables but rather express the effect of the field on certain properties of the system, for example its lifetime. It is not clear whether this result is significant or simply an indication that the phenomenological way that we introduced the θ's is incorrect.

Notice that the commutator of the derivatives depends explicitly on the wavefunction, and so is different for different wavefunctions. Also, of course, θ depends on the wavefunction. This is contrast to the gravitational case where it is a property of space, independent of the particular function that it is applied to. However this should not be unexpected for in a gravitational field the behavior of particles is independent of their properties. This is not the case for the electromagnetic field. Perhaps the appearance of the wavefunction is related to this point.

It is not clear whether we should require the theory to be gauge invariant. The pure electromagnetic part should be, of course, but since there is no experimental

information on how the presence of an electromagnetic field modifies the gravitational field there is no reason to require invariance in this part of the theory. It is quite possible that the potential is determined by the field equations and cannot be changed by a gauge transformation.

The next step is to find a relation between the connections and the metric tensor. This is done in the usual manner by requiring the invariance of the scalar product along a curve (with parameter s). Thus

$$\frac{d}{ds}(g_{\mu\kappa}\xi^{*\mu}\eta^{\kappa}) = g_{\mu\kappa,\rho}\frac{dx^{\rho}}{ds}\xi^{*\mu}\eta^{\kappa} + g_{\mu\kappa}\frac{d\xi^{*\mu}}{ds}\eta^{\kappa} + g_{\mu\kappa}\xi^{*\mu}\frac{d\eta^{\kappa}}{ds} = 0, \tag{IV.13}$$

which gives

$$g_{\mu\kappa,\rho} + g_{\omega\kappa}\Gamma^{\omega}_{\mu\rho} + g_{\mu\omega}\Gamma^{\omega}_{\kappa\rho} - ig_{\mu\kappa}\chi^{*}_{\rho} + ig_{\mu\kappa}\chi_{\rho} = 0, \tag{IV.14}$$

and in the normal way rotating indices cyclically, adding and subtracting we get the required relation

$$\Gamma^{\omega}_{\kappa\mu} = -\frac{1}{2}g^{\omega\rho}(g_{\rho\mu,\kappa} + g_{\kappa\rho,\mu} - g_{\mu\kappa,\rho}) - \delta^{\omega}_{\mu}\frac{1}{2}i(-\chi^{*}_{\kappa} + \chi_{\kappa}) - \delta^{\omega}_{\kappa}\frac{i}{2}(-\chi^{*}_{\rho} + \chi_{\rho}) + g_{\mu\kappa}\frac{1}{2}i(-\chi^{*\omega} + \chi^{\omega}), \tag{IV.15}$$

so assuming f is real,

$$\Gamma^{\omega}_{\kappa\mu} = -\left\{{\omega \atop \kappa\mu}\right\} + \delta^{\omega}_{\mu}f\theta_{\kappa} + \delta^{\omega}_{\kappa}f\theta_{\mu} - g_{\mu\kappa}f\theta^{\omega}, \tag{IV.16}$$

where $\{\}$ is the Christoffel symbol.

Equation (IV. 13) requires some comment in view of our requirement that the phase term in the connection be independent of the number of indices which would seem to be in contradiction with our differentiation of each term.

However here what is important is not that we have a product of two vectors, but that we have a product of a term and a complex conjugate (both assumed to have the same phase connection). So, quantum mechanically if we had such a product it would represent a particle of zero charge which would consist of two oppositely charged particles, and which would have no phase change in going through a potential, but which might be reflected or be caused to decay because of the interaction of the internal particles and the field. Thus the connection would depend only on the imaginary part of the connection, as we have it. This is the reason for requiring this scalar product to be invariant.

Equation (IV. 14) shows that the covariant derivative of the metric tensor is zero, as can be seen by writing this derivative out and comparing equations, and remembering that the charge function f is zero for the metric tensor g so this term drops out.

It might be noted that some of the equations are superficially similar to the corresponding equations in Weyl's theory.[2] However the physical rationale and the mathematical structure are completely different. For example in our case of connection, Eq. (IV. 16), depends not on the electromagnetic potential (which is ξ) but on the factor $f\theta$.

V. THE CURVATURE TENSOR

The curvature tensor obeys the symmetry equations

$$R^{\mu}_{\eta\xi\rho} = -R^{\mu}_{\xi\eta\rho} \tag{V.1}$$

and

$$R^{\mu}_{\eta\xi\rho} + R^{\mu}_{\xi\rho\eta} + R^{\mu}_{\rho\eta\xi} = 0, \tag{V.2}$$

as can be seen directly from its expression in terms of the connection.

From these symmetries we can derive the Bianchi identities. We denote $(\chi_{\eta;\xi} - \chi_{\xi;\eta})$ by $T_{\eta\xi}$, and by direct computation we see that $T_{\xi\eta;\phi}$ has the symmetry given by Eq. (V. 2), using Eq. (V. 2), and the symmetries of R.

Also a second-rank tensor obeys

$$\Omega^{\mu}_{\phi;\eta\xi} - \Omega^{\mu}_{\phi;\xi\eta} = R^{\mu}_{\eta\xi\rho}\Omega^{\rho}_{\phi} - R^{\rho}_{\eta\xi\phi}\Omega^{\mu}_{\rho} - iT_{\eta\xi}\Omega^{\mu}_{\phi}. \tag{V.3}$$

Now differentiating Eq. (IV. 7) we get

$$\psi^{\mu}_{;\eta\xi\phi} - \psi^{\mu}_{;\xi\eta\phi} = R^{\mu}_{\eta\xi\rho;\phi}\psi^{\rho} + R^{\mu}_{\eta\xi\rho}\psi^{\rho}_{;\phi} - iT_{\eta\xi}\psi^{\mu}_{;\phi} - iT_{\eta\xi;\phi}\psi^{\mu}. \tag{V.4}$$

Rotating the indices η, ξ, ϕ cyclically, adding, using the symmetry of T to eliminate the $T_{\eta\xi}$ term, and substituting Eq. (V. 3) into the left-hand side so it cancels the $R_{\eta\xi\rho}$ and $T_{\eta\xi}$ terms, we get ψ^{ρ} times a coefficient, equal to zero. But since ψ^{ρ} is arbitrary (within a class, all of which have the same f) the coefficient must be zero, and this is the Bianchi identity.

$$R^{\mu}_{\eta\xi\rho;\phi} + R^{\mu}_{\xi\phi\rho;\eta} + R^{\mu}_{\phi\eta\rho;\xi} = 0. \tag{V.5}$$

From this it follows that the Ricci tensor is divergenceless, so

$$G^{\rho\sigma}_{;\sigma} = (R^{\rho\sigma} - \frac{1}{2}g^{\rho\sigma}R)_{;\sigma} = 0. \tag{V.6}$$

We have an expression above for the curvature tensor in terms of the connection Γ. It is more useful to write it in terms of the metric tensor and the θ's. To do this we use Eq. (IV. 16).

We then get that the curvature tensor is

$$\begin{aligned} R^{\mu}_{\xi\eta\rho} &= \Gamma^{\mu}_{\rho\eta,\xi} - \Gamma^{\mu}_{\rho\xi,\eta} - \Gamma^{\mu}_{\kappa\eta}\Gamma^{\kappa}_{\rho\xi} + \Gamma^{\mu}_{\kappa\xi}\Gamma^{\kappa}_{\rho\eta} \\ &= -\left[\left\{{\mu \atop \rho\eta}\right\}_{,\xi} - \left\{{\mu \atop \rho\xi}\right\}_{,\eta} - \left\{{\mu \atop \kappa\eta}\right\}\left\{{\kappa \atop \rho\xi}\right\} + \left\{{\mu \atop \kappa\xi}\right\}\left\{{\kappa \atop \rho\eta}\right\}\right] \\ &\quad + f^{2}[-\delta^{\mu}_{\eta}\theta_{\xi}\theta_{\rho} + \delta^{\mu}_{\xi}\theta_{\eta}\theta_{\rho} + (\delta^{\mu}_{\eta}g_{\rho\xi} - \delta^{\mu}_{\xi}g_{\eta\rho})\theta^{\kappa}\theta_{\kappa} \\ &\quad - g_{\rho\xi}\theta^{\mu}\theta_{\eta} + g_{\rho\eta}\theta_{\xi}\theta^{\mu}] + f[\delta^{\mu}_{\rho}(\theta_{\eta,\xi} - \theta_{\xi,\eta}) + \delta^{\mu}_{\eta}\theta_{\rho,\xi} \\ &\quad - \delta^{\mu}_{\xi}\theta_{\rho,\eta} - g_{\eta\rho}\theta^{\mu}_{,\xi} + g_{\rho\xi}\theta^{\mu}_{,\eta} - 2g_{\rho\eta,\xi}\theta^{\mu} + 2g_{\rho\xi,\eta}\theta^{\mu} + g_{\eta\rho}\left\{{\mu \atop \kappa\xi}\right\}\theta^{\kappa} \\ &\quad - g_{\xi\rho}\left\{{\mu \atop \kappa\eta}\right\}\theta^{\kappa} + \delta^{\mu}_{\eta}\theta_{\kappa}\left\{{\kappa \atop \rho\xi}\right\} - \delta^{\mu}_{\xi}\theta_{\kappa}\left\{{\kappa \atop \rho\eta}\right\}] + [\delta^{\mu}_{\rho}(f_{,\xi}\theta_{\eta} - f_{,\eta}\theta_{\xi}) \\ &\quad + (\delta^{\mu}_{\eta}f_{,\xi} - \delta^{\mu}_{\xi}f_{,\eta})\theta_{\rho} + (g_{\rho\xi}f_{,\eta} - g_{\rho\eta}f_{,\xi})\theta^{\mu}]. \end{aligned} \tag{V.7}$$

We next need the value of $\delta^{\mu}_{\mu} \equiv \delta^{\mu}_{\mu} + \delta^{\underline{\mu}}_{\underline{\mu}} = 4 + 4$ or 8.

The contracted curvature tensor is

$$\begin{aligned} R_{\eta\rho} &= R^{\mu}_{\mu\eta\rho} \\ &= -\left[\left\{{\mu \atop \rho\eta}\right\}_{,\mu} - \left\{{\mu \atop \rho\mu}\right\}_{,\eta} - \left\{{\mu \atop \kappa\eta}\right\}\left\{{\kappa \atop \rho\mu}\right\} + \left\{{\mu \atop \kappa\mu}\right\}\left\{{\kappa \atop \rho\eta}\right\}\right] \\ &\quad + f[\theta_{\eta,\rho} - 8\theta_{\rho,\eta} - g_{\eta\rho}\theta^{\mu}_{,\mu} + g_{\eta\rho}\left\{{\mu \atop \kappa\mu}\right\}\theta^{\kappa} - 2g_{\rho\eta,\mu}\theta^{\mu} \\ &\quad + 2g_{\rho\mu,\eta}\theta^{\mu} + g_{\mu\rho}\theta^{\mu}_{,\eta} - g_{\rho\mu}\left\{{\mu \atop \kappa\eta}\right\}\theta^{\kappa} - 7\theta_{\kappa}\left\{{\kappa \atop \rho\eta}\right\}] + 6f^{2}[-g_{\rho\eta}\theta^{\tau}\theta_{\tau} \end{aligned}$$

$$+\theta_\rho\theta_\eta]+[f_{,\rho}\theta_\eta-7f_{,\eta}\theta_\rho-g_{\rho\eta}f_{,\mu}\theta^\mu], \tag{V.8}$$

and the curvature scalar equals

$$R=R_c+f[-7g^{\nu\eta}\theta_{\nu,\eta}-5\theta^\rho g^{\nu\eta}(g_{\eta\rho,\nu}-g_{\nu\eta,\rho})-7\theta^\rho{}_{,\rho}]$$
$$-42f^2\theta_\rho\theta^\rho-14f_{,\rho}\theta^\rho. \tag{V.9}$$

The subscript c means the Christoffel symbols have been substituted for the connections Γ.

Note that the curvature a particle sees depends in part on its own state vector.

Unfortunately we cannot introduce, at this point, enough field equations to determine completely all the quantities appearing in the theory. To do this, we would probably have to add more quantum mechanical formalism.

However we can assume some relations. These are Maxwell's equations

$$V^{\mu\nu}_{;\nu}=j^\mu, \tag{V.10}$$

and the Einstein field equations

$$G^{\eta\rho}=\kappa T^{\eta\rho}, \tag{V.11}$$

where j is the current density and T the energy-momentum tensor.

The Einstein field equations are modified by the presence of the electromagnetic field and depend specifically on f, the charge function for the particle moving in the gravitational field. Thus these equations imply that the gravitational field that a particle feels depends on its charge, and possibly its charge distribution. Whether this disagrees with experiment is not clear as the theory is not developed enough to predict the value for f, or θ, which also affects the field.

[1]The possibility of complex space–time has been explored in other ways by A. Das, J. Math. Phys. 7, 45 (1966); J. Math. Phys. 7, 52 (1966); J. Math. Phys. 7, 61 (1966), and E. H. Brown, J. Math. Phys. 7, 417 (1966).

[2]R. Adler, M. Bazin, and M. Schiffer, *Introduction to General Relativity* (McGraw-Hill, New York, 1965), Sec. 13-2,3.

Noncompact Lie-algebraic approach to the unitary representations of $\widetilde{SU}(1,1)$: Role of the confluent hypergeometric equation

W. Montgomery and L. O'Raifeartaigh

Dublin Institute for Advanced Studies, Dublin 4, Ireland

(Received 3 February 1972; revised manuscript received 22 June 1972)

The continuous unitary irreducible representations of the covering group $\widetilde{SU}(1,1)$ of $SU(1,1)$ are studied using the self-adjoint representations of the Lie algebra in the case that one of the noncompact generators is diagonal. The study can be carried out for all the representations simultaneously and is shown to reduce to a study of the self-adjointness of the compact element of the Lie algebra, which in this basis turns out to be the confluent hypergeometric operator. Several basic results, such as the classification of the representations, and a formula for the transformation coefficients from the compact to the noncompact basis which is valid for *all* representations, emerge quite simply.

INTRODUCTION

The continuous unitary irreducible representations (CUIR'S) of $SU(1,1)$ and its covering group $\widetilde{SU}(1,1)$ can be studied either through the finite group[1-5] or through the self-adjoint representations of the Lie algebra [5] The latter method is based on Nelson's result[6] that a Lie algebra will exponentiate to a CUIR of the corresponding simply connected group if and only if an operator consisting of the sum of the squares of the generators is self-adjoint. Most of the standard treatments use the finite group, and those treatments which use the Lie algebra are mostly confined to the case in which the compact generator is diagonal. However, the general procedure for the case in which a noncompact generator is diagonal has been sketched in Refs. 7 and carried through in some detail for $SU(1,1)$ in Refs. 8. In this paper we carry through the general procedure in a different manner to Ref. 8. Our method is such that all the CUIR's of $\widetilde{SU}(1,1)$ can be treated simultaneously, and that many of the results found by other methods, including the classification of CUIR's, can be obtained quite simply. More specifically, we show that

(1) The generators of Lie algebra can be expressed as differential operators, which are unique up to a unitary transformation, and are the same (apart from the value of one parameter) for all CUIR's of $\widetilde{SU}(1,1)$. Two of the independent generators are trivial, while the third, the compact generator J_3, is the confluent hypergeometric operator,[9] a result that explains at a very elementary level the occurrence[2,3] of the confluent hypergeometric (Whittaker) functions in the study of the finite group representations.

(2) The self-adjoint condition for the generators reduces to a condition of orthogonality for the eigenfunctions of the compact generator J_3, and the orthogonality condition leads at once to the standard classification of the CUIR's of $\widetilde{SU}(1,1)$ and to a formula for the transformation coefficients from the compact to the noncompact basis which holds for all CUIR's. The classification parameters emerge, respectively, as the value of the Casimir operator for the Lie algebra, and the parameter characterizing the family of self-adjoint extensions[10] of J_3 in the case when it is not naturally self-adjoint. The formula for the transformation coefficients agrees (up to a phase) with the separate results for the principal and discrete series obtained in Ref. 2

We conclude the note by illustrating the difference between our results and those of the other authors[8] who use a noncompact basis for the Lie algebra by exhibiting for the principal series the unitary transformation which connects the two results.

2. LIE ALGEBRA, CLASSIFICATION AND TRANSFORMATION COEFFICIENTS

In the conventional basis, the Lie algebra of $\widetilde{SU}(1,1)$ is

$$[J_a, J_b] = i\epsilon_{abc}\lambda_c J_c, \quad a,b,c = 1,2,3 \; \lambda_1 = \lambda_2 = -\lambda_3 = 1. \tag{2.1}$$

and the Casimir operator, which is a real number in any CUIR of $\widetilde{SU}(1,1)$, may be written as

$$\mu^2 - \frac{1}{4} = J_3^2 - J_1^2 - J_2^2. \tag{2.2}$$

Let $J_1 + J_3$ be the noncompact element which we wish to have diagonal. From (2.1) it is clear that J_2 and $J_1 + J_3$ form the two-dimensional subalgebra

$$[J_2, J_1 + J_3] = i(J_1 + J_3), \tag{2.3}$$

and it is easy to verify[11] that the only nontrivial self-adjoint representations of this algebra are unitarily equivalent to the representations

$$J_1 + J_3 = \epsilon, \quad J_2 = i\epsilon\frac{d}{d\epsilon}, \quad \int \frac{d\epsilon}{|\epsilon|}\,|\psi(\epsilon)|^2 < \infty, \tag{2.4}$$

on $(-\infty\, 0]$ and $[0\, \infty)$. To complete the basis for the Lie algebra (2.1) we add the compact generator J_3, and using (2.1), (2.2), and (2.4) obtain for it the differential form

$$J_3 = -\tfrac{1}{2}\left(\epsilon\frac{d^2}{d\epsilon^2} - \epsilon + (\tfrac{1}{4} - \mu^2)\epsilon^{-1}\right). \tag{2.5}$$

From (2.2) we see that the Nelson operator can be chosen to be $2J_3^2 + \frac{1}{4} - \mu^2$. It then follows from Nelson's result that the necessary and sufficient condition for the differential operators (2.4) and (2.5) to generate a CUIR of $\widetilde{SU}(1,1)$ is that J_3 in (2.5) be self-adjoint. But since, in every CUIR of $\widetilde{SU}(1,1)$, the spectrum of J_3 can easily be shown from (2.1) to be discrete[12] (indeed integer-spaced and simple) the self-adjointness of J_3 is equivalent to the orthogonality of its eigenvectors. Thus the analyses of the CUIR's of $\widetilde{SU}(1,1)$ reduces to an analysis

of the eigenvector system of the differential operator J_3 belonging to real discrete eigenvalues k.

To analyze this eigenvector system we let $\psi_k(\epsilon)$ be the eigenvector belonging to the real eigenvalue k. Then from (2.5) $\psi_k(\epsilon)$ is a square integrable solution of the equation $J_3\psi_k(\epsilon)=k\psi_k(\epsilon)$ which, on making the trivial substitution $t=2\epsilon$ reduces to the confluent hypergeometric equation[9]

$$\left[\frac{d^2}{dt^2}+\left(-\frac{1}{4}+\frac{k}{t}+\frac{\frac{1}{4}-\mu^2}{t^2}\right)\right]\phi_k(t)=0, \quad \phi_k(t)=\psi_k(\epsilon). \tag{2.6}$$

Thus we obtain at once a direct relationship between the Lie algebra (2.1) and the confluent hypergeometric equation and we can use the known properties of the solutions of that equation to analyze the eigenvector system of J_3. In fact, the two linearly independent solutions of (3.2) are the Whittaker functions $W_{\pm k\mu}(\pm t)$, and we shall actually need only their asymptotic properties[9]

$$W_{k\mu}(t)=t^k e^{-t/2}[1+O(t^{-1})], \quad t\to\infty, \tag{2.7}$$

$$W_{k\mu}(t)=\left(\frac{\sqrt{2\mu}}{\sqrt{\frac{1}{2}+\mu-k}}\,t^{1/2-\mu}+\frac{\sqrt{-2\mu}}{\sqrt{\frac{1}{2}-\mu-k}}\,t^{1/2+\mu}\right)[1+O(t)], \quad t\to 0. \tag{2.7}$$

From (2.6) we have by partial integration

$$(k-k')\int_0^\infty \frac{dt}{t}\,W_{k\mu}(t)W_{k'\mu}(t) = [{}_0^\infty W_{k'\mu}(t)\dot{W}_{k\mu}(t)-\dot{W}_{k\mu}(t)W_{k'\mu}(t)], \quad \dot{W}\equiv\frac{d}{dt}W, \tag{2.8}$$

Hence from (3.3) and (3.4) we see that a solution is square integrable and we have the possibility of generating a CUIR of $\widetilde{SU}(1,1)$, if $\mu^2<\frac{1}{4}$ for all real k or $\mu^2\geq\frac{1}{4}$ and $\pm k=|\mu|+\frac{1}{2}+n$, $n=0,1,2,\ldots$. Furthermore, in those cases we have[13]

$$\int_0^\infty \frac{dt}{t}W_{k\mu}(t)W_{k'\mu}(t)=\frac{\pi}{(k-k')\sin 2\pi\mu} \times\left(\frac{1}{\sqrt{\frac{1}{2}+\mu-k}}\,\frac{1}{\sqrt{\frac{1}{2}-\mu-k'}}-\frac{1}{\sqrt{\frac{1}{2}-\mu-k}}\,\frac{1}{\sqrt{\frac{1}{2}+\mu-k'}}\right), \tag{2.9}$$

whence

$$\int_{-\infty}^0 \frac{dt}{|t|}\,\frac{W_{-k\mu}(-tW_{-k'\mu}(-t)}{\sqrt{\frac{1}{2}+\mu+k}\,\sqrt{\frac{1}{2}-\mu+k'}}+\int_0^\infty \frac{dt}{t}\,\frac{W_{k\mu}(t)W_{k'\mu}(t)}{\sqrt{\frac{1}{2}+\mu+k}\,\sqrt{\frac{1}{2}-\mu+k'}} =\frac{\sin\pi(k-k')}{\pi(k-k')}. \tag{2.10}$$

It follows by inspection of (2.9) and (2.10) that in general we obtain an orthogonal set of eigenvectors for J_3 if and only if the spectrum of J_3 is integer-spaced and we take the full t axis $(-\infty,\infty)$. The only exception occurs when $\mu^2\geq 0$, $\pm k=|\mu|+\frac{1}{2}+n$, in which case the spectrum is still integer-spaced, but we have orthogonality on the half-axes. Even that case, however, is included in the formula (2.10) since for $\pm k=|\mu|+\frac{1}{2}+n$ one of the terms on the left-hand side of (2.10) vanishes identically.

The above results allow us to classify the CUIR's of $\widetilde{SU}(1,1)$ immediately, and we obtain the conventional result, namely,

(1) $\mu^2\geq 0$, $\pm k=|\mu|+\frac{1}{2}+n$, $n=0,1,2,\ldots$ (discrete series),

(2) $\frac{1}{4}>\mu^2>0$, $0,1>h\geq 0$, $k=h+n$ (except $h=\mu+\frac{1}{2}$) (exeptional series),

(3) $\mu^2\leq 0$, $0,1>h\geq 0$, $k=h+n$ (exept $\mu=0$, $h=\frac{1}{2}$) (principal series).

The exceptions for the series (2) and (3) are simply to avoid overlapping the discrete series. The CUIR's for $SU(1,1)$ are obtained by imposing the extra restriction $2k=$ integer.

A bonus which we obtain from (3.11) is the set of transformation coefficients $(k|\epsilon)$ from the compact to the noncompact basis, since these are nothing but the normalized eigenfunctions $\psi_k(\epsilon)=\phi_k(t)$. For the principal series μ is pure imaginary, so we can normalize directly from (2.10). Since $W_{k\mu}$ is even in μ, it is real for all three series of representations, and hence for the principal series we obtain

$$\int_0^\infty \frac{dt}{t}[\omega_{-k}^*(t)\omega_{-k'}(t)+\omega_k^*(t)\omega_{k'}(t)]=\delta_{kk'} \tag{2.11}$$

for $k-k'=$ integer, where

$$\omega_k(t)=(\sqrt{\tfrac{1}{2}-\mu+k})^{-1}W_{k\mu}(t), \quad \mu+\mu^*=0. \tag{2.12}$$

For the other two series μ is real and we do not obtain the complex inner product so simply. What we do is in that case to introduce the function

$$N(k,k')=\sqrt{\tfrac{1}{2}+\mu+k}\,\sqrt{\tfrac{1}{2}-\mu+k'}\,(\sqrt{\tfrac{1}{2}+\mu+k'})^{-1}(\sqrt{\tfrac{1}{2}-\mu+k})^{-1}, \tag{2.13}$$

and multiply equation (2.10) across by $N^{1/2}(k,k')$. Since $N(k,k')=N(-k,-k')$ for $k-k'$ integer, we obtain (2.11), but with now

$$\omega_k(t)=(\sqrt{\tfrac{1}{2}-\mu+k})^{-1/2}(\sqrt{\tfrac{1}{2}+\mu+k})^{-1/2}W_{k\mu}(t), \quad \mu=\mu^*. \tag{2.14}$$

Equations (2.11), (2.12), and (2.15) can be combined to give a formula for the transformation coefficients which is valid for *all* CUIRs of $\widetilde{SU}(1,1)$, namely

$$(k|\epsilon)=\begin{bmatrix}\omega_k(2\epsilon), & \epsilon>0\\ \omega_{-k}(-2\epsilon), & \epsilon<0\end{bmatrix}=\begin{bmatrix}\omega_k(2\epsilon)\\ \omega_{-k}(2\epsilon)\end{bmatrix}, \quad \epsilon>0,$$

$$\omega_k(t)=(\sqrt{\tfrac{1}{2}-\mu+k})^{-1/2}(\sqrt{\tfrac{1}{2}+\mu^*+k})^{-1/2}W_{k\mu}(t). \tag{2.15}$$

For the principal series the factor preceding $W_{k\mu}(t)$ in (2.15) simplifies slightly to $(\sqrt{\frac{1}{2}-\mu+k})^{-1}$, while for the discrete series one of the functions $\omega_{\pm k}(2\epsilon)$ vanishes identically, and the other simplifies to

$$\omega_{\pm k}(t)=(-1)^n(\sqrt{2|\mu|+n+1})^{1/2}(\sqrt{n+1})^{-1/2}(\sqrt{2|\mu|+1})^{-1}$$

$$\times e^{-t/2}t^{|\mu|+1/2}L^{2|\mu|}_{2|\mu|+n}(t), \qquad L^{2|\mu|}_{2|\mu|+n}(0)=1, \tag{2.16}$$

where $L^{2|\mu|}_{2|\mu|+n}(t)$ is the associated Laguerre polynomial.

Let us now compare our transformation coefficients with those of other authors. For the exceptional series, no other orthonormalized coefficients seem to be available, but for the discrete series our results agree exactly with those of Refs. 2 and 8, and for the principal series our results agree with those of Ref. 2 up to the phases $e^{i\theta_k}=\sqrt{\frac{1}{2}-\mu\mp h}\,(\sqrt{\frac{1}{2}+\mu\mp h})^{-1}$ for $\omega_{\pm k}(t)$ [the $\omega_{\pm k}(t)$ are relatively real in this reference].

3. COMPARISON WITH REF. 8

So far as we know the only authors to use a similar approach to ours are those of Ref. 8, and, except in the case of the discrete series, their actual procedure differs markedly. We here illustrate the difference by exhibiting the unitary transformation which connected the two approaches in the case of the principal series. We start with the generators

$$J_1+J_3=\epsilon\otimes\tau=e^x\otimes\tau \text{ and } G_1+G_3=e^x\left(i\frac{d}{dx}+\frac{1}{2}+\sigma\right)\otimes\tau,$$

$$J_2=i\epsilon\frac{d}{d\epsilon}\otimes 1=i\frac{d}{dx}\otimes 1, \quad G_2=i\frac{d}{dx}\otimes 1,$$

$$\left.\begin{aligned} J_1-J_3&=\left(\epsilon\frac{d^2}{d\epsilon^2}+(\tfrac{1}{4}+\sigma^2)\epsilon^{-1}\right)\otimes\tau \\ &=e^{-x}\left[\left(\frac{d}{dx}-\frac{1}{2}\right)^2+\sigma^2\right]\otimes\tau \end{aligned}\right\},$$

$$G_1-G_3=e^{-x}\left(-i\frac{d}{dx}+\frac{1}{2}+\sigma\right)\otimes\tau, \tag{3.1}$$

respectively, where $\mu=i\sigma$ (σ real), $\epsilon\geq 0$, and $\tau=\left(\begin{smallmatrix}1&0\\0&-1\end{smallmatrix}\right)$. The range $\epsilon\leq 0$, $\tau=\pm 1$ is equivalent to the range $(-\infty<\epsilon<\infty)$ which is necessary for self-adjointness, but the introduction of τ simplifies the notation and allows the introduction of the variables $x=\log\epsilon$. Bothe J_a and G_a satisfy the commutation relations (2.1) and the Casimir relation (2.2), and the essential difference between them is that whereas J_1+J_3 and J_1-J_3 are zero and second-order differential operators, respectively, G_1+G_3 and G_1-G_3 are both first order. G_3 is then a first-order differential operator, with the simple orthonormal eigenfunction system

$$\theta_k(x)=\frac{1}{(2\pi)^{1/2}}\left(\frac{1-ie^x}{1+ie^x}\right)^k\frac{1}{(\cosh x)^{1/2-i\sigma}}, \tag{3.2}$$

and this is the chief advantage of the G_a. The diadvantage of the G_a is that they are Hermitian only in the case of the principal series (though the authors of Ref. 8 actually use the G_a for the exceptional series also, compensating the lack of Hermiticity by the introduction of an indefinite metric).

We now seek a unitary operator U which will connect the J_a and G_a. Since any such U must commute with $J_2=G_2$, it is convenient to make $J_2=G_2$ diagonal, i.e., to take the Fourier transform of (3.1). We then have (3.1) with

$$i\frac{d}{dx}=y, \quad e^x=e^{-i(d/dy)}, \tag{3.3}$$

where the domain of $e^{-i(d/dy)}$ is the set of functions analytic in the strip $0\leq \mathrm{Im}y\leq 1$ and square integrable on the lines $\mathrm{Im}y=\text{const}$ in the strip, and the required transformation is

$$U(y)=\frac{\sqrt{2z}}{(2\pi)^{1/2}}\begin{bmatrix} e^{izz} & e^{-izz} \\ e^{-izz} & e^{izz}\end{bmatrix}, \quad z=\tfrac{1}{4}-\tfrac{1}{2}i(y+\sigma). \tag{3.4}$$

It is easy to verify that $U(y)$ is unitary and satisfies

$$J_a=U^{-1}(y)\,G_aU(y). \tag{3.5}$$

ACKNOWLEDGMENTS

One of the authors (O'R) wishes to thank Dr. J. M. Golden and Dr. P. Yodzis for discussions concerning the special functions, and to thank the referee for bringing Refs. 2 and 3 to their attention.

[1]V. Bargmann, Ann. Math. 48, 568 (1947).
[2]P. Sally, Mem. Am. Math. Soc. 69, (1967).
[3]N. Vilenkin and M. Sleinikova, Mat. Sb. (N.S.) 81, 185 (1970) [Math. USSR Sb. 10, 173 (1970)].
[4]N. Nilenkin, *Group Representations*, Am. Math. Soc. Translations (Providence, 1968); W. Ruhl, *The Lorentz Group and Harmonic Analyses* (Benjamin, New York, 1970), and references therein.
[5]L. Pukansky, Math. Annalen 156, 96 (1964); A.O. Barut and C. Fronsdal, Proc. Roy. Soc. (London) 287A, 532 (1965); A.O. Barut and E.C. Phillips, Commun. Math. Phys. 8, 52 (1968).
[6]E. Nelson, Annals. Math. 70, 572 (1959).
[7]W. Miller, Mem. Am. Math. Soc. 50, 1 (1964); Siam J. Math. Analyses 2, 307 (1971) (see page 326); *Lie Theory and Special Functions* (Academic, New York, 1968).
[8]N. Mukunda, J. Math. Phys. 8, 2210 (1967); J.G. Kuriyan, N. Mukunda, and E.C.G. Sudarshan, J. Math. Phys. 9, 2100 (1968).
[9]E.T. Whittaker and G.N. Watson, *Modern Analyses* (Cambridge U.P., London, 1952), pp. 337, 343.
[10]N.I. Akhiezer and J.M. Glazman, *Theory of Linear Operators in Hilbert Space II* (Ungar, New York, 1963), p. 201.
[11]M.A. Naimark, *Normed Rings* (Noordhoff, Groningen, 1959), p. 381.
[12]S. Komy and L. O'Raifeartaigh, J. Math. Phys. 9, 738 (1968).
[13]Our result agrees with that given in the latest edition of H. Buchholtz, *The Confluent Hypergeometric Function* (Springer, New York, 1969), p. 116, but differs by a factor π from that given in the Bateman Manuscript, *Tables of Integral Transforms II* (McGraw-Hill, New York, 1954), pp. 409, 430 and in *Tables of Integrals, Series and Products*; I.S. Gradshteyn and I.M. Ryzmik (Academic, New York, 1965), p. 858, both of which quote an earlier edition of Buchholtz.

On some mathematical aspects of deterministic classical electrodynamics

B. W. Stuck

Bell Laboratories, Murray Hill, New Jersey 07974
(Received 14 November 1973)

A new commutative diagram summarizing some of the mathematical structure of deterministic classical electrodynamics is presented. The diagram clearly delineates the fundamentally different roles played by the space–time differentiable manifold (*vis à vis* exterior calculus) and by matter or vacuum (*vis à vis* the constitutive relations for dielectric permittivity and magnetic permeability) in electrodynamics. Two different elliptic operators, called here the Laplace–Beltrami and Laplace–Poisson operators, arise naturally from this formulation. Some properties of eigenfunctions of elliptic operators with compact support are briefly reviewed with regard to potential application in numerical analysis of practical problems in electrodynamics. The action of the so-called inhomogeneous Lorentz group on electrodynamical functions is described. Several scalar inner products which remain invariant under the action of this group are seen to arise naturally from the mathematical structure discussed here. By using some of these invariant quantities, a new variational approach to deterministic classical electrodynamics is then developed. First, a new Lagrangian function is presented and used to derive the Euler–Lagrange equations for electrodynamics. Second, a series of new Hamiltonian functions are presented and used to derive the Hamiltonian equations for electrodynamics. All results are illustrated by a detailed examination of the electrodynamical structure of a model for an inhomogeneous nonisotropic medium.

I. INTRODUCTION

Although an extensive literature exists on mathematical aspects of deterministic classical electrodynamics, there is apparently no clear rigorous exposition on the relationship of exterior calculus and differential forms to Maxwell's equations. This is somewhat surprising, because exterior calculus would hopefully clarify some of the mathematical structure underlying electrodynamics, while offering a different (more formal but less physical) view of the structure than that of conventional or classical vector calculus. Such work hopefully would continue the interaction of mathematics and physics, which has been so fruitful in the past. Finally, the tremendous technological importance of electrodynamics lends added interest to such work.

This paper attempts to fill some of this gap in the literature. The scope is limited to a particular class of models for an inhomogeneous nonisotropic medium. Within this framework, a number of novel and well-known results are obtained more easily and naturally than by methods based on conventional vector calculus.

The goal here is to unify and simplify certain mathematical aspects of electromagnetism. An example of a successful attempt along similar lines can be found in modern communication and control theory, which have been greatly unified through the concept of state and state variable techniques. It is hoped the approach discussed here will find application in other branches of physics, just as state variables have found wide application (e.g., in electrical network theory and in control system theory). This hope must be tempered by the following observation: Many practical problems can be adequately modeled by a set of first order ordinary differential equations, where the state space is a finite-dimensional vector space. The analogous state space for a distributed parameter system, such as is discussed here, is a finite-dimensional differentiable manifold and the vector fields associated with the manifold. The dimensionality of the state space for lumped parameter systems can be anything in practice; the underlying manifold for distributed parameter systems might be only four-dimensional, three spatial and one temporal, in practice. This suggests that future work should be directed toward a better understanding of the peculiarities of the four-dimensional case, as well as toward generalizations in higher dimensions.

Although interesting in their own right, the results presented here are interesting from a purely pedagogical point of view as well. One need only known the operations or rules of exterior algebra, as well as how to compute the total differential of a function; then the calculation of gradient, curl, and divergence become routine formal manipulations, but unfortunately often devoid of physical insight into the nature of the calculation. The conventional or classical vector calculus approach, with its line integrals and pillboxes, compliments this method by offering great physical insight into the nature of the calculation, but often at the expense of algebraic complexity in computing the correct answer. Both approaches have their merits and disadvantages, offering different views on the same situation.

The initial motivation for this work is found in Flanders.[1] While it was felt his approach was basically sound, it seemed sketchy at points and could be considerably more detailed. Another impetus is found in Dyson,[2] who has observed that the foundations of exterior calculus were laid by Grassmann in the mid-nineteenth century, but the tools he developed were discarded when the mathematical structure of electrodynamics was considered, in favor of tools developed to describe the structure of Lie groups and Lie algebras.

An appendix is included sketching and illustrating the basic concepts of multilinear algebra, differentiable manifolds, and exterior calculus. The reader familiar with these topics can proceed directly to the main body of the paper; otherwise, this detour is advised.

The second section presents a commutative diagram

which summarizes the electrodynamical equations for an inhomogeneous nonisotropic medium; this is analogous to a block diagram or signal flow graph in control and communication theory.

The third section discusses two different differential operators which arise from this formulation of electrodynamics; previous work has tended to ignore or obscure this point. Various properties of the spectrum and eigenfunctions of these operators are briefly reviewed, for the case where the operators are elliptic and compactly supported.

The fourth section dwells on a group of coordinate transformations which preserve the structure of the equations of electrodynamics. Several well-known and new scalar inner products are seen to arise naturally from the approach discussed here.

The fifth section develops an alternate calculus-of-variations approach to the mathematical structure of electrodynamics. A new Lagrangian function is discussed, and all of the electrodynamical equations are derived from it. A new series of Hamiltonian functions are derived from Legendre transformations on the Lagrangian, and all of the electrodynamical equations are rederived.

All these results are illustrated by examining again and again a model for an inhomogeneous nonisotropic medium.

II. A COMMUTATIVE DIAGRAM

Throughout this section, X is an oriented Riemannian differentiable manifold called space–time.[3] T_X^* denotes the cotangent bundle associated with X, and $\Lambda(T_X^*) = \sum_{k=0}^{4} \Lambda^k(T_X^*)$ the associated exterior algebra. This section is broken into two parts: first, a diagram is presented which summarizes Maxwell's equations for a particular class of inhomogeneous nonisotropic media (in effect, the equations can be read off with the aid of this diagram); second, an example is presented to illustrate more clearly this result.

Theorem (Classical electrodynamics—Maxwell's equations): If $M: \Lambda^2(T_X^*) \to \Lambda^2(T_X^*)$ is a smooth function of $\Lambda^2(T_X^*)$, and is invertible at every point of the manifold X, then the diagram shown below commutes

$$\begin{array}{ccccccccc} \Lambda^0 & \xrightarrow{d} & \Lambda^1 & \xrightarrow{d} & \Lambda^2 & \xrightarrow{d} & \Lambda^3 & \xrightarrow{d} & \Lambda^4 \\ \Big\downarrow{\scriptstyle d^2\circ M\circ d^2} & & \Big\downarrow{\scriptstyle d\circ M\circ d} & & \Big\downarrow{\scriptstyle M} & & \Big\uparrow{\scriptstyle M^{-1}} & & \Big\uparrow{\scriptstyle d^2\circ M^{-1}\circ d^2} \\ \Lambda^4 & \xleftarrow{d} & \Lambda^3 & \xleftarrow{d} & \Lambda^2 & \xleftarrow{d} & \Lambda^1 & \xleftarrow{d} & \Lambda^0 \end{array}$$

where d is the exterior derivative.

Proof: The proof proceeds in three steps: (i) all operations shown above must be well defined on all charts of the manifold, i.e., locally; (ii) all operations must be capable of being pieced together smoothly on overlapping charts; (iii) the diagram must commute, i.e., be independent of path. Since the exterior derivative and the linear transformation M are well defined, all operations shown in the diagram are valid on each chart of the manifold. On overlapping coordinate charts, the transition functions associated with these charts can be used to smoothly piece together the operators M, M^{-1}, $d\circ M\circ d$, $d\circ M^{-1}\circ d$, $d^2\circ M\circ d^2$, $d^2\circ M^{-1}\circ d^2$. Finally, the diagram commutes, because of the two preceding steps.

Remark: Since $d^2 = 0$, the maps from $\Lambda^0 \to \Lambda^4$ are trivial.

Example: Choose rectangular orthonormal basis vectors $\{dx, dy, dz, icdt\}$ for X, and orientation $+dx \wedge dy \wedge dz \wedge icdt$. ($c$ is the speed of light.) The physical nature of each differential form is well known:

(A) Λ^0—g_e, g_m—electric, magnetic gauge.

(B) Λ^1—(A_{ex}, A_{ey}, A_{ez}), (A_{mx}, A_{my}, A_{mz})—electric, magnetic vector potential; φ_e, φ_m—electrical, magnetic scalar potential.

(C) Λ^2—(D_x, D_y, D_z), (B_x, B_y, B_z)—electric displacement, magnetic flux; (E_x, E_y, E_z), (H_x, H_y, H_z)—electric, magnetic fields.

(D) Λ^3—(J_{mx}, J_{my}, J_{mz}), (J_{ex}, J_{ey}, J_{ez})—magnetic, electric current densities; ρ_m, ρ_e—magnetic, electric charge densities.

(E) Λ^4—s_m, s_e—magnetic, electric source.

The question arises of how to associate which differential form with which electromagnetic function. The choice adopted here offers a certain amount of physical appeal, and is self consistent and complete with respect to exterior calculus.

Since X is a four-dimensional differentiable manifold, the differential forms may be interpreted intuitively as follows:

(1) Λ^0—gauge—scalar functions of space–time.

(2) Λ^1—potentials—directed line elements or 1-volumes in space–time.

(3) Λ^2—fields—directed areas or 2-volumes in space–time, in part directed along purely spatial directions ($dy \wedge dz, dz \wedge dx, dx \wedge dy$) and in part directed along a mixture of space–time directions ($dx \wedge icdt, dy \wedge icdt, dz \wedge icdt$).

(4) Λ^3—current densities—directed shells or 3-volumes in space–time in part directed along a purely spatial direction ($dx \wedge dy \wedge dz$) and in part directed along a mixture of space–time directions ($dy \wedge dz \wedge icdt, dz \wedge dx \wedge icdt, dx \wedge dy \wedge icdt$).

(5) Λ^4—sources—directed volumes or 4-volumes in space–time.

It is interesting to give a physical interpretation to the commutative diagram, much as in control and communication theory problems one gives a physical interpretation to a block diagram. Suppose, for example, a 1-form or potential is known at every point in X. The exterior derivative of this potential specifies a 2-form or field at every point; applying the constitutive relations plus the exterior derivative to the field specifies a 3-form or current density, which in turn feeds back to modify the potentials, and so on.

In this choice of coordinates, Maxwell's equations can be written using exterior calculus as:

(1) $d(g_e + ig_m) = \frac{\partial}{\partial x}(g_e + ig_m)\,dx + \frac{\partial}{\partial y}(g_e + ig_m)\,dy$

$$+\frac{\partial}{\partial z}(g_e+ig_m)\,dz+\frac{\partial}{ic\,\partial t}(g_e+ig_m)icdt$$
$$=c(A_{ex}+iA_{mx})\,dx+c(A_{ey}+iA_{my})\,dy$$
$$+c(A_{ez}+iA_{mz})\,dz+(\varphi_e+i\varphi_m)icdt,$$

(2) $d[c(A_{ex}+iA_{mx})\,dx+c(A_{ey}+iA_{my})\,dy+c(A_{ez}+iA_{mz})\,dz$
$+(\varphi_e+i\varphi_m)icdt]$
$$=\left(\frac{\partial}{\partial y}c(A_{ez}+iA_{mz})-\frac{\partial}{\partial z}c(A_{ey}+iA_{my})\right)dy\wedge dz$$
$$+\left(\frac{\partial}{\partial z}c(A_{ex}+iA_{mx})-\frac{\partial}{\partial x}c(A_{ez}+iA_{mz})\right)dz\wedge dx$$
$$+\left(\frac{\partial}{\partial x}c(A_{ey}+iA_{my})-\frac{\partial}{\partial y}c(A_{ex}+iA_{mx})\right)dx\wedge dy$$
$$+\left(\frac{\partial}{\partial x}(\varphi_e+i\varphi_m)-\frac{\partial}{ic\,\partial t}c(A_{ex}+iA_{mx})\right)dx\wedge icdt$$
$$+\left(\frac{\partial}{\partial y}(\varphi_e+i\varphi_m)-\frac{\partial}{ic\,\partial t}c(A_{ey}+iA_{my})\right)dy\wedge icdt$$
$$+\left(\frac{\partial}{\partial z}(\varphi_e+i\varphi_m)-\frac{\partial}{ic\,\partial t}c(A_{ez}+iA_{mz})\right)dz\wedge icdt$$
$$=c(D_x+iB_x)\,dy\wedge dz+c(D_y+iB_y)\,dz\wedge dx$$
$$+c(D_z+iB_z)\,dx\wedge dy+(E_x+iH_x)\,dx\wedge icdt$$
$$+(E_y+iH_y)\,dy\wedge icdt+(E_z+iH_z)\,dz\wedge icdt,$$

(3) $d[c(D_x+iB_x)\,dy\wedge dz+c(D_y+iB_y)\,dz\wedge dx$
$+c(D_z+iB_z)\,dx\wedge dy+(E_x+iH_x)\,dx\wedge icdt$
$+(E_y+iH_y)\,dy\wedge icdt+(E_z+iH_z)\,dz\wedge icdt]$
$$=\left(\frac{\partial}{\partial y}(E_z+iH_z)-\frac{\partial}{\partial z}(E_y+iH_y)+\frac{\partial}{ic\,\partial t}c(D_x+iB_x)\right)$$
$$\times dy\wedge dz\wedge icdt+\left(\frac{\partial}{\partial z}(E_x+iH_x)-\frac{\partial}{\partial x}(E_z+iH_z)\right.$$
$$\left.+\frac{\partial}{ic\,\partial t}c(D_y+iB_y)\right)dz\wedge dx\wedge icdt+\left(\frac{\partial}{\partial x}(E_y+iH_y)\right.$$
$$\left.-\frac{\partial}{\partial y}(E_x+iH_x)+\frac{\partial}{ic\,\partial t}c(D_z+iB_z)\right)dx\wedge dy\wedge icdt$$
$$+\left(\frac{\partial}{\partial x}c(D_x+iB_x)+\frac{\partial}{\partial y}c(D_y+iB_y)+\frac{\partial}{\partial z}c(D_z+iB_z)\right.$$
$$\left.\times dx\wedge dy\wedge dz\right.$$
$$=(J_{mx}+iJ_{ex})\,dy\wedge dz\wedge icdt+(J_{my}+iJ_{ey})\,dz\wedge dx\wedge icdt$$
$$+(J_{mz}+iJ_{ez})\,dx\wedge dy\wedge icdt+c(\rho_e+i\rho_m)\,dx\wedge dy\wedge dz,$$

(4) $d[(J_{mx}+iJ_{ex})\,dy\wedge dz\wedge icdt+(J_{my}+iJ_{ey})\,dz\wedge dx\wedge icdt$
$+(J_{mz}+iJ_{ez})\,dx\wedge dy\wedge icdt+c(\rho_e+i\rho_m)\,dx\wedge dy\wedge dz]$
$$=\left(\frac{\partial}{\partial x}(J_{mx}+iJ_{ex})+\frac{\partial}{\partial y}(J_{my}+iJ_{ey})+\frac{\partial}{\partial z}(J_{mz}+iJ_{ez})\right.$$
$$\left.-\frac{\partial}{ic\,\partial t}c(\rho_e+i\rho_m)\right)dx\wedge dy\wedge dz\wedge icdt$$
$$=(s_m+is_e)\,dx\wedge dy\wedge dz\wedge icdt.$$

The classical electrodynamics equations are found by equating real and imaginary parts of (1)—(4). The sign on φ_e and ρ_m must be reversed to conform to that standard in physics.[4] It is assumed here the transformation M can be written in matrix form for an inhomogeneous nonisotropic medium as

$$\begin{bmatrix} cD_x\ dy\wedge dz \\ cD_y\ dz\wedge dx \\ cD_z\ dx\wedge dy \\ E_x\ dx\wedge icdt \\ E_y\ dy\wedge icdt \\ E_z\ dz\wedge icdt \end{bmatrix} = \{p\circ\tilde{M}_E\circ *\circ+(1-p)\circ *\circ\tilde{M}_E\circ\} \begin{bmatrix} cD_x\ dy\wedge dz \\ cD_y\ dz\wedge dx \\ cD_z\ dx\wedge dy \\ E_x\ dx\wedge icdt \\ E_y\ dy\wedge icdt \\ E_z\ dz\wedge icdt \end{bmatrix},$$

$$\begin{bmatrix} cB_x\ dy\wedge dz \\ cB_y\ dz\wedge dx \\ cB_z\ dx\wedge dy \\ H_x\ dx\wedge icdt \\ H_y\ dy\wedge icdt \\ H_z\ dz\wedge icdt \end{bmatrix} = \{p\circ\tilde{M}_M\circ *\circ+(1-p)\circ *\circ\tilde{M}_M\circ\} \begin{bmatrix} cB_x\ dy\wedge dz \\ cB_y\ dz\wedge dx \\ cB_z\ dx\wedge dy \\ H_x\ dx\wedge icdt \\ H_y\ dy\wedge icdt \\ H_z\ dz\wedge icdt \end{bmatrix},$$

where $0\leq p\leq 1$. In this example, M is assumed to be a convex combination of the star operator, $\tilde{M}_E$ and $\tilde{M}_M$, where

$$\tilde{M}_E=\begin{bmatrix} 0 & c\epsilon \\ c^{-1}\epsilon^{-1} & 0 \end{bmatrix},\quad \tilde{M}_M=\begin{bmatrix} 0 & c\mu \\ c^{-1}\mu^{-1} & 0 \end{bmatrix},$$

where μ, ϵ are 3×3 matrices, 0 is the all zero 3×3 matrix. μ is called magnetic permeability, while ϵ is dielectric permittivity; the units are meter·kilogram·second.

In other treatments[5,6] a different set of units are often used: In these units the dielectric permittivity ϵ and magnetic permeability μ are rescaled, and it is frequently states that (in these units) $E_x=D_x$, $H_x=B_x$, and so forth. Strictly speaking, these equalities are quite ill-defined because the electric field (E_x,E_y,E_z) and electric displacement (D_x,D_y,D_z) lie in orthogonal subspaces of Λ^2, as does the magnetic field (H_x,H_y,H_z) and magnetic flux (B_x,B_y,B_z). To emphasize this often ignored fact, meter·kilogram·second units have been adopted.

In order to model the inhomogeneity of the medium, matrix elements in ϵ and μ are smooth functions of x,y,z and ict. To account for the anisotropy of the medium, ϵ and μ are assumed not to be similar to scalar multiples of the identity matrix.

Clearly, this choice of assumed constitutive relations for $\tilde{M}$ is not the only one that can model an inhomogeneous nonisotropic medium: The only essential assumption is that $\tilde{M}$ must be invertible on its support, X. The example here was chosen as illustrative of linear constitutive relationships; it can be generalized in any number of ways. For example, the next section shows $\Lambda^2(X)$ can be considered as a Hilbert space, the space of all functions in $L^2(X)$; $\tilde{M}$ may now be defined as an invertible operator defined on Hilbert space. Other generalizations are possible.

III. ELLIPTIC OPERATORS

Two differential operators are seen to arise naturally from this formulation of electrodynamics, the Laplace—Beltrami operator and the Laplace—Poisson operator.

The Laplace—Beltrami operator $\Delta=d\delta+\delta d$, where

$\delta = *\circ d \circ *$, is elliptic (Warner, Ref. 7, pp. 250–251), and

$$\Delta: \ \Lambda^K \to \Lambda^K, \quad K = 0, 1, 2, 3, 4.$$

Since Δ depends only on the underlying manifold X, a picturesque description of Δ is that it is totally geometric or topological in nature. If X is a compact manifold, then the Hodge decomposition theorem shows that any differential form $u_p \in \Lambda^p$ ($p = 0, 1, 2, 3, 4$) can be written as the sum of an exact form, a coexact form, and a harmonic form which lies in the finite-dimensional kernel of Δ,

$$u_p = u_p{}^H \oplus du_{p-1}{}^E \oplus \delta u_{p+1}{}^C, \quad p = 0, 1, 2, 3, 4,$$

where the superscripts H, E, C denote harmonic, exact, and coexact, respectively (Ref. 7, p. 223).

The Laplace–Poisson operator $d \circ M \circ d = d \circ M^{-1} \circ d$ depends partly on the underlying manifold X (via the exterior derivative d) and partly on the physics (embodied in $\tilde{M}$); this operator may be considered as partly geometric or topological and partly physical. In the special case which is of great practical interest where the Laplace–Poisson operator can be shown to be elliptic (e.g., constant permittivity ϵ and permeability μ, a homogeneous nonisotropic media) a great deal more can be ascertained. If X is compact, then any p-form may be written as the sum of a p-form lying in the finite-dimensional kernel of the operator, plus a term in the orthogonal complement of this vector space.

Since the Laplace–Beltrami operator is always elliptic, while the Laplace–Poisson operator is often elliptic, a brief review of some of the properties of the eigenfunctions and eigenvalues of elliptic operators is included. Let E be an elliptic operator whose support is on a compact manifold X; then it is well known that (Ref. 7, pp. 254–256)

(1) nontrivial eigenvalues and eigenfunctions of E exist,

(2) there are an infinite number of eigenfunctions,

(3) all eigenvalues are nonpositive,

(4) the eigenfunctions are complete in $L^2(X)$,

(5) any function in $L^2(X)$ can be uniformly approximated by a sequence of these eigenfunctions, on X,

(6) the eigenvalues have no finite accumulation point,

(7) the eigenspaces associated with each eigenvalue are finite-dimensional.

Example: From the Hodge decomposition theorem,

(A) $g_e + ig_m = f_0^H \oplus \delta f_1^C$,

(B) $c(A_{ex} + iA_{mx})\,dx + c(A_{ey} + iA_{my})\,dy + c(A_{ez} + iA_{mz})\,dz + (\varphi_e + i\varphi_m)icdt = f_1^H \oplus \delta f_0^E \oplus \delta f_2^C$,

(C) $c(D_x + iB_x)\,dy \wedge dz + c(D_y + iB_y)\,dz \wedge dx + c(D_z + iB_z)\,dx \wedge dy + (E_x + iH_x)\,dx \wedge icdt + (E_y + iH_y)\,dy \wedge icdt + (E_z + iH_z)\,dz \wedge icdt = f_2^H \oplus df_1^E \oplus \delta f_3^C$,

(D) $(J_{mx} + iJ_{ex})\,dy \wedge dz \wedge icdt + (J_{my} + iJ_{ey})\,dy \wedge dx \wedge icdt + (J_{mz} + iJ_{ez})\,dx \wedge dy \wedge icdt + c(\rho_e + i\rho_m)\,dx \wedge dy \wedge dz = f_3^H \oplus df_2^E \oplus \delta f_4^C$,

(E) $(s_m + is_e)\,dx \wedge dy \wedge dz \wedge icdt = f_4^H \oplus df_3^E$,

where $f_K^H, f_K^E, f_K^C \in \Lambda^K$ ($K = 0, 1, 2, 3, 4$), and the superscripts H, E, C denote harmonic, exact, and coexact, respectively. If X is compact and simply connected, it can be shown that (Ref. 7, p. 158 and pp. 226–229)

$$f_0^H = \text{const},$$

$$f_1^H = 0, \quad f_2^H = 0, \quad f_3^H = 0,$$

$$f_4^H = (\text{const})\,dx \wedge dy \wedge dz \wedge icdt,$$

corresponding physically to a source-free region of space–time. The terms δf_K^C ($K = 1, 2, 3, 4$) and df_i^E ($i = 0, 1, 2, 3$) can be expressed as linear combinations of eigenfunctions of the Laplace–Beltrami operator. The exact and coexact forms are also called *Hertz vectors*.[4]

The Laplace–Poisson operator, since it is a different operator from the Laplace–Beltrami operator, will in general have different eigenfunctions. Note that any 3-form can be expressed as an infinite linear combination of these eigenfunctions denoted $\{\tilde{u}_k^3\}$, $k = 1, 2, \cdots$. Using the exterior derivative d, its adjoint δ, plus the Hodge star operator $*$, the following statements hold (recall the underlying manifold is four-dimensional, so $d\tilde{u}_k^3 = 0$) on a compact manifold:

(i) Any 0-form may be written as an infinite linear combination of

$$\{*\circ d \circ \tilde{u}_k^3\}, \quad k = 1, 2, \cdots,$$

(ii) Any 1-form may be written as an infinite linear combination of

$$\{*\tilde{u}_k^3\}, \quad k = 1, 2, \cdots,$$

(iii) Any 2-form may be written as an infinite linear combination of

$$\{\delta\tilde{u}_k^3\}, \quad k = 1, 2, \cdots,$$

(iv) Any 4-form may be written as an infinite linear combination of

$$\{d\tilde{u}_k^3\}, \quad k = 1, 2, \cdots.$$

This finding may have practical application. In semiconductor device work, or in transmission of electromagnetic energy, Maxwell's equations plus real boundary conditions are often analytically intractable, and a numerical approximation to the true solution must often be used. One type of numerical approximation is to expand all functions as a sum of a finite number of orthonormal functions, and to truncate the sum when an error criteria is sufficiently small. The approach presented here makes it possible to choose from two different sets of orthonormal functions; under some circumstances, one set may be preferable to the other.

IV. SOME GROUP THEORETIC ASPECTS

In certain situations, a great deal of insight is gained by a change of coordinates. This section is concerned with a class of coordinate transformations which form a group, and quantities which remain invariant under this class of transformations.

Consider the semidirect product of the Lie group $SO(4)$ with an affine group T, $G=SO(4)\tilde{\times}T$ ($\tilde{\times}$ denotes semidirect product); G is called the inhomogeneous Lorentz group. One parameter subgroups of T correspond physically to translations of the origin of the space–time coordinate frame. One parameter subgroups of $SO(4)$ correspond physically to rotation about an axis or motion along an axis. It is straightforward to show G acts transitively on X: given $x\in X$, $g\in G$, then $gx\in X$. Since $G: G\times X\to X$, G is well defined on scalar functions $f\in\Lambda^0$, and this action is denoted $L_0, L_0: L_0\times\Lambda^0\to\Lambda^0$.

Since T_X and T_X^*, the tangent and cotangent bundles of X, are isomorphic to the direct product of X with itself, G has a well-defined transitive action on T_X and T_X^*. Since T_X can be identified with $\Lambda^1(T_X)$, while T_X^* can be identified with $\Lambda^1(T_X^*)$, G acts in a well-defined manner on $\Lambda^1(T_X^*)$, denoted L_1, $L_1: L_1\times\Lambda^1\to\Lambda^1$.

It is now necessary to extend the action of G to Λ^2, Λ^3, and Λ^4. To illustrate how this is accomplished, consider an orthonormal set of basis vectors $\{e_1, e_2, e_3, e_4\}$ for Λ^1 (the extension to a general basis is straightforward). ge_k is the action of g on e_k ($k=1,2,3,4$) for some $g\in G$; $\{e_k'=ge_k\}$ is a set of orthonormal basis vectors for $\Lambda^{1\prime}$. Since $\{e_i\wedge e_j \mid i=1,2,3,\ j=2,3,4\}$ is a basis for Λ^2, $\{ge_i\wedge ge_j \mid i=1,2,3;\ j=2,3,4\}$ is a basis for $\Lambda^{2\prime}$, and $L_2: L_2\times\Lambda^2\to\Lambda^2$ is the well-defined action of G on $\Lambda^2(T_X^*)$. Similarly, $\{ge_i\wedge ge_j\wedge ge_k,\ i=1,2,\ j=2,3,\ k=3,4\}$ is a basis for $\Lambda^{3\prime}$, and $\{ge_1\wedge ge_2\wedge ge_3\wedge ge_4\}$ is a basis for $\Lambda^{4\prime}$, which lead to well-defined actions of G, $L_3: L_3\times\Lambda^3\to\Lambda_3$ and $L_4: L_4\times\Lambda^4\to\Lambda^4$. This can be summarized as follows.

Proposition: The diagram shown below commutes

$$\begin{array}{ccccccccc}
\Lambda^0 & \xrightarrow{d} & \Lambda^1 & \xrightarrow{d} & \Lambda^2 & \xrightarrow{d} & \Lambda^3 & \xrightarrow{d} & \Lambda^4 \\
\downarrow{\scriptstyle L_0} & & \downarrow{\scriptstyle L_1} & & \downarrow{\scriptstyle L_2} & & \downarrow{\scriptstyle L_3} & & \downarrow{\scriptstyle L_4} \\
\Lambda^{0\prime} & \xrightarrow{d'} & \Lambda^{1\prime} & \xrightarrow{d'} & \Lambda^{2\prime} & \xrightarrow{d'} & \Lambda^{3\prime} & \xrightarrow{d'} & \Lambda^{4\prime}
\end{array}$$

Proof: Again, the proof has three parts. First, observe that d and d' (the exterior derivative in the new coordinates), as well as L_k ($k=0,1,2,3,4$) are well defined on each chart of X. Second, note that d, d', and L_k ($k=0,1,2,3,4$) are well defined globally, using the transition functions to smoothly piece together the operators on overlapping coordinate charts. Third, the verification the diagram commutes is straightforward, because of the two preceding steps.

Since X has a well-defined inner product $\langle a,b\rangle$ is well defined, where either $a\in\Lambda^K$, $b\in\Lambda^K$, or $a\in\Lambda^K$, $*b\in\Lambda^{4-K}$. Both the real and imaginary parts of all these inner products remain invariant under the action of G. The Hamiltonian and Lagrangian functions result from forming linear combinations of these inner products.[2,4,5]

Example: In rectangular coordinates, the inner products invariant under the action of G are

(i) $\langle (g_e+ig_m), *[(s_m+is_e)\,dx\wedge dy\wedge dz\wedge icdt]\rangle = (g_e+ig_m)(s_m+is_e)$

(ii) $\langle c(A_{ex}+iA_{mx})\,dx + c(A_{ey}+iA_{my})\,dy + c(A_{ez}+iA_{mz})\,dz + (\varphi_e+i\varphi_m)\wedge icdt,$
$*[(J_{mx}+iJ_{ex})\,dy\wedge dz\wedge icdt + (J_{my}+iJ_{ey})\,dz\wedge dx\wedge icdt + (J_{mz}+iJ_{ez})\,dx\wedge dy\wedge icdt + c(\rho_e+i\rho_m)\,dx\wedge dy\wedge dz]\rangle$
$= c(A_{ex}+iA_{mx})(J_{mx}+iJ_{ex}) + c(A_{ey}+iA_{my})(J_{my}+iJ_{ey}) + c(A_{ez}+iA_{mz})(J_{mz}+iJ_{ez}) - (\varphi_e+i\varphi_m)c(\rho_e+i\rho_m),$

(iii) $\langle c(D_x+iB_x)\,dy\wedge dz + c(D_y+iB_y)\,dz\wedge dx + c(D_z+iB_z)\,dx\wedge dy + (E_x+iH_x)\,dx\wedge icdt + (E_y+iH_y)\,dy\wedge icdt + (E_z+iH_z)\,dz\wedge icdt,$
$*[c(D_x+iB_x)\,dy\wedge dz + (D_y+iB_y)\,dz\wedge dx + c(D_z+iB_z)\,dx\wedge dy + (E_x+iH_x)\,dx\wedge icdt + (E_y+iH_y)\,dy\wedge icdt + (E_z+iH_z)\,dz\wedge icdt]\rangle$
$= 2\{c(D_x+iB_x)(E_x+iH_x) + c(D_y+iB_y)(E_y+iH_y) + c(D_z+iB_z)(E_z+iH_z)\},$

(iv) $\|(g_e+ig_m)\|^2 = (g_e+ig_m)(g_e+ig_m),$

(v) $\| c(A_{ex}+A_{mx})\,dx + c(A_{ey}+iA_{my})\,dy + c(A_{ez}+iA_{mz})\,dz + (\varphi_e+i\varphi_m)icdt\|^2$
$= c^2[(A_{ex}+iA_{mx})^2 + (A_{ey}+iA_{my})^2 + (A_{ez}+iA_{mz})^2] + (\varphi_e+i\varphi_m)^2,$

(vi) $\| c(D_x+iB_x)\,dy\wedge dz + c(D_y+iB_y)\,dz\wedge dx + c(D_z+iB_z)\,dx\wedge dy + (E_x+iH_x)\,dx\wedge icdt + (E_y+iH_y)\,dy\wedge icdt + (E_z+iH_z)\,dz\wedge icdt\|^2$
$= c^2[(D_x+iB_x)^2 + (D_y+iB_y)^2 + (D_z+iB_z)^2] + (E_x+iH_x)^2 + (E_y+iH_y)^2 + (E_z+iH_z)^2,$

(vii) $\| (J_{mx}+iJ_{ex})\,dy\wedge dz\wedge icdt + (J_{my}+iJ_{ey})\,dz\wedge dx\wedge icdt + (J_{mz}+iJ_{ez})\,dx\wedge dy\wedge icdt + c(\rho_e+i\rho_m)\,dx\wedge dy\wedge dz\|^2$
$= (J_{mx}+iJ_{ex})^2 + (J_{my}+iJ_{ey})^2 + (J_{mz}+iJ_{ez})^2 + c^2(\rho_e+i\rho_m)^2,$

(viii) $\| (s_m+is_e)\,dx\wedge dy\wedge dz\wedge icdt\|^2 = (s_m+is_e)^2.$

Remark: (i), (iv), and (viii) are often overlooked invariants, (cf. Refs. 5, 6).

V. VARIATIONAL PRINCIPLES

For the sake of completeness, as well as to have an alternate interesting way in which to view the mathematical structure in electrodynamics, a Lagrangian and Hamiltonian formulation will now be discussed. The results presented here are more complete than any other of which the author is aware,[5,6] and illustrate a new relationship between dynamics based on exterior calculus and dynamics based on a calculus of variations approach. Since many excellent treatments[8,9] can be found in the literature on Lagrangian and Hamiltonian dynamics, but few good examples can be found on how to apply this knowledge, the general discussion is cursory, while the example is dwelt on at length.

The *Lagrangian function* L is defined as

$$L:\ \Lambda(T^*(X))\times\Lambda(T^*(X))\to R,$$
$$L=\tfrac{1}{2}RE(+\langle u_0, *\circ u_4\rangle + \langle u_1, *\circ u_3\rangle - \langle u_2, *\circ u_2\rangle - \langle u_3, *\circ u_1\rangle$$

TABLE I

Generalized coordinate	Components of associated generalized momentum			
	x	y	z	ict
g_e	J_{mx}	J_{my}	J_{mz}	$-ic\rho_m$
ig_m	iJ_{ex}	iJ_{ey}	iJ_{ez}	$-c\rho_e$

$$+\langle u_4, *\circ u_0\rangle),$$

where $RE(a+ib)=a,\ a,b\in \mathrm{R}$.

In order to specify L on a chart of a manifold, one must give the local coordinates of the chart, all elements in $\Lambda(T^*(X))$ and all partial derivatives of elements in $\Lambda(T^*(X))$ with respect to local coordinates. The elements in $\Lambda(T^*(X))$ are called *generalized coordinates* while the partial derivatives are called *generalized velocities*.

The *generalized momenta* are defined as the partials of L with respect to the generalized velocities. The *Hamiltonian function* H is derived from L by computing the inner product of the generalized velocities with their respective generalized momenta and then subtracting the Lagrangian L; this transformation is called a *Legendre transformation*. The Hamiltonian function H is specified on a chart of a manifold by specifying coordinates on the chart, the generalized coordinates and the generalized momenta.

Solutions to Maxwell's equations are extremals to the *action integral*

$$\int_1^2 L\, de_1\wedge de_2\wedge de_3\wedge de_4,$$

where the integral is evaluated along a space–time trajectory beginning at point 1 and ending at point 2, and $de_1\wedge de_2\wedge de_3\wedge de_4$ is a unit basis vector for Λ^4. For a more complete and precise discussion of how to evaluate this integral, the reader is referred to the bibliography (Spivak,[10] Loomis-Sternberg,[11] Warner[7]).

Given a Lagrangian function, a well-defined recipe due to Euler and Lagrange exists for finding the associated equations of motions whose solutions are extremals to the action integral. Given a Hamiltonian function, a well-defined formula due to Hamilton exists for finding the associated equations of motion. Since both these approaches are independent of the constitutive relations, but depend only on the underlying differentiable manifold and its associated vector fields, the resulting equations of motion are said, picturesquely, to be totally geometric or topological in nature, independent of matter or vacuum.

TABLE III

Generalized coordinate	Components of associated generalized momentum			
	x	y	z	ict
E_x	0	$-cA_{ez}$	cA_{ey}	0
E_y	cA_{ez}	0	$-cA_{ex}$	0
E_z	$-cA_{ey}$	cA_{ex}	0	0
icB_x	$-i\varphi_m$	0	0	cA_{ex}
icB_y	0	$-i\varphi_m$	0	cA_{ey}
icB_z	0	0	$-i\varphi_m$	cA_{ez}
iH_x	0	$-icA_{mz}$	icA_{my}	0
iH_y	icA_{mz}	0	$-icA_{mx}$	0
iH_z	$-icA_{my}$	icA_{mx}	0	0
cD_x	$-\varphi_e$	0	0	icA_{mx}
cD_y	0	$-\varphi_e$	0	icA_{my}
cD_z	0	0	$-\varphi_e$	icA_{mz}

Example: The Lagrangian function L is

$$L=-(cD_x\cdot E_x+cD_y\cdot E_y+cD_z\cdot E_z)+(cB_x\cdot H_x+cB_y\cdot H_y+cB_z\cdot H_z)-(g_m s_e-g_e s_m)-(cA_{mx}J_{ex}+cA_{my}J_{ey}+cA_{mz}J_{ez}+\varphi_e\cdot c\rho_e)+(cA_{ex}J_{mx}+cA_{ey}J_{my}+cA_{ez}J_{mz}+\varphi_m\cdot c\rho_m).$$

(A) The generalized coordinates are g_e and ig_m. The generalized velocities are all partials of g_e and ig_m with respect to x, y, z and ict. The x component of the generalized momentum associated with g_e is

$$x\text{ component}=\frac{\partial L}{\partial(\partial g_e/\partial x)}=\frac{\partial L}{\partial cA_{ex}}\cdot\frac{\partial cA_{ex}}{\partial(\partial g_e/\partial x)}=J_{mx}.$$

Note that to compute the generalized momentum it is necessary not only to compute $(\partial L/\partial cA_{ex})$ but also to know from Maxwell's equations that $\partial cA_{ex}/\partial(\partial g_e/\partial x)=+1$. In like manner it is straightforward to find all the generalized momenta, and the results are summarized in the Table I.

The Hamiltonian function H is

$$H_A=-(cB_x\cdot H_x+cB_y\cdot H_y+cB_z\cdot H_z)+(cD_x\cdot E_x+cD_y\cdot E_y+cD_z\cdot E_z)-(g_e s_m-g_m s_e)$$

and is independent of the generalized momentum. The Euler–Lagrange equations of motion are

$$g_e\colon\ s_m-\frac{\partial}{\partial x}(J_{mx})-\frac{\partial}{\partial y}(J_{my})-\frac{\partial}{\partial z}(J_{mz})-\frac{\partial}{\partial ict}-ic\rho_m=0,$$

TABLE II

Generalized coordinate	Components of associated generalized momenta			
	x	y	z	ict
cA_{ex}	0	E_z	$-E_y$	icB_x
cA_{ey}	$-E_z$	0	E_x	icB_y
cA_{ez}	E_y	$-E_x$	0	icB_z
$i\varphi_m$	$-icB_x$	$-icB_y$	$-icB_z$	0
icA_{mx}	0	iH_z	$-iH_y$	cD_x
icA_{my}	$-iH_z$	0	iH_x	cD_y
icA_{mz}	iH_y	$-iH_x$	0	cD_z
φ_e	$-cD_x$	$-cD_y$	$-cD_z$	0

TABLE IV

Generalized coordinate	components of associated generalized momentum			
	x	y	z	ict
J_{mx}	g_e	0	0	0
J_{my}	0	g_e	0	0
J_{mz}	0	0	g_e	0
$ic\rho_m$	0	0	0	$-g_e$
iJ_{ex}	ig_m	0	0	0
iJ_{ey}	0	ig_m	0	0
iJ_{ez}	0	0	ig_m	0
$c\rho_e$	0	0	0	$-ig_m$

TABLE V

Generalized Coordinate	Components of Associated Generalized momentum			
	x	y	z	ict
s_m	0	0	0	0
is_e	0	0	0	0

$$ig_m\colon\ is_e-\frac{\partial}{\partial x}(iJ_{ex})-\frac{\partial}{\partial y}(iJ_{ey})-\frac{\partial}{\partial z}(iJ_{ez})-\frac{\partial}{\partial ict}-c\rho_e=0$$

The Hamiltonian equations of motion are identical:

$$\frac{\partial H}{\partial g_e}=-s_m=-\left(\frac{\partial}{\partial x}J_{mx}+\frac{\partial}{\partial y}J_{my}+\frac{\partial}{\partial z}J_{mz}+\frac{\partial}{\partial ict}-ic\rho_m\right),$$

$$\frac{\partial H}{\partial ig_m}=-is_e=-\left(\frac{\partial}{\partial x}iJ_{ex}+\frac{\partial}{\partial y}iJ_{ey}+\frac{\partial}{\partial z}iJ_{ez}+\frac{\partial}{\partial ict}-c\rho_e\right).$$

Since the Hamiltonian is independent of the generalized momentum, the dual equations involved derivatives of H with respect to momenta are all zero. Note these equations are identical to those in Sec. 2, Eq. (4).

(B) The generalized coordinates and generalized momenta are tabulated in Table II.

The Hamiltonian function H is

$$\begin{aligned}H_B=&-(cD_x\cdot E_x+cD_y\cdot E_y+cD_z\cdot E_z)+(H_x\cdot cB_x+H_y\cdot cB_y\\&+H_z\cdot cB_z)-(g_es_m-g_ms_e)-(cA_{ex}\cdot J_{mx}+cA_{ey}\cdot J_{my}\\&+cA_{ez}\cdot J_{mz}+\varphi_m\cdot c\rho_m)+(cA_{mx}J_{ex}+cA_{my}J_{ey}\\&+cA_{mz}J_{ez}+\varphi_e\cdot c\rho_e).\end{aligned}$$

The Euler—Lagrange and Hamiltonian equations of motion are found in Sec. 2, Eq. (3).

(C) The generalized coordinates and generalized momenta are tabulated in Table III.

The Hamiltonian function H is

$$\begin{aligned}H_C=&-(cB_x\cdot H_x+cB_y\cdot H_y+cB_z\cdot H_z)+(cD_x\cdot E_x+cD_y\cdot E_y\\&+cD_z\cdot E_z)-(g_es_m-g_ms_e).\end{aligned}$$

The Euler—Lagrange and Hamiltonian equations of motion are found in Sec. 2, Eq. (2).

(D) The generalized coordinates and generalized momenta are tabulated in Table IV.

The Hamiltonian function H is

$$\begin{aligned}H_D=&(cD_x\cdot E_x+cD_y\cdot E_y+cD_z\cdot E_z)-(cB_x\cdot H_x+cB_y\cdot H_y\\&+cB_z\cdot H_z)+(cA_{mx}J_{ex}+cA_{my}J_{ey}+cA_{mz}J_{ez}+\varphi_e\cdot c\rho_e)\\&-(cA_{ex}J_{mx}+cA_{ey}J_{my}+cA_{ez}J_{mz}+\varphi_m\cdot c\rho_m).\end{aligned}$$

The Euler—Lagrange and Hamiltonian equations of motion are found in Sec. 2, Eq. (1).

(E) The generalized coordinates and generalized momenta are summarized in Table V.

The Hamiltonian function H is

$$\begin{aligned}H_E=&-L\\=&-(cB_x\cdot H_x+cB_y\cdot H_y+cB_z\cdot H_z)+(cD_x\cdot E_x+cD_y\cdot E_y\\&+cD_z\cdot E_z)-(g_es_m-g_ms_e)-(cA_{ex}J_{mx}+cA_{ey}J_{my}\\&+cA_{ez}J_{mz}+\varphi_mc\rho_m)+(cA_{mx}J_{ex}+cA_{my}J_{ey}+cA_{mz}J_{ez}\\&+\varphi_ec\rho_e).\end{aligned}$$

The Euler—Lagrange and Hamiltonian equations of motion are

$$g_e=0,\quad ig_m=0.$$

ACKNOWLEDGMENTS

The author appreciates several conversations with Professor R.W. Brockett of Harvard University for stimulating this work. Professor R. Hermann of Rutgers University and Dr. R. Brammer of NASA Goddard Spaceflight Center offered several concrete suggestions for improving the exposition of the ideas presented here, as did Dr. R. Ludwig and Dr. I. Sandberg of Bell Laboratories.

APPENDIX: MATHEMATICAL PRELIMINARIES

This section is largely tutorial, sketching some of the fundamental concepts of multilinear algebra, differentiable manifolds, and exterior calculus. Suitable references can be found in the bibliography (Nelson,[12] Warner,[7] Spivak,[10] Loomis-Sternberg[11]).

A. Multilinear algebra

Let R denote the real numbers, and let V and W be finite-dimensional real linear vector spaces. V^* denotes the dual space of V, consisting of all real valued functions of V. The *direct product* of V with W is denoted $V\times W$ and consists of all linear combinations of pairs (v,w), with $v\in V$ and $w\in W$. The k-fold direct product of V with itself, denoted V^k, the l-fold direct product of V^* with itself, denoted V^{*l}, and the mixed direct product of $V^k\times V^{*l}$ are defined in an identical fashion.

Let $S(V^k)$ be the subspace of V^k generated by the set of all elements of the form

$$\left.\begin{aligned}&(v_1,\ldots,v_i'+v_i'',\ldots,v_k)-(v_1,\ldots,v_i',\ldots,v_k)-(v_1,\ldots,v_i'',\ldots,v_k)\\&(v_1,\ldots,av_i,\ldots,v_k)-a(v_1,\ldots,v_i,\ldots,v_k)\end{aligned}\right\}\begin{aligned}&v_i,v_i',v_i''\in V,\quad i=1\cdots k,\\&a\in\mathrm{R}.\end{aligned}$$

The quotient space $T^k=V^k/S(V^k)$, the set $\{x\colon (y-x)\in S(V^k)$, for all $y\in V^k\}$, is called the set of kth *order contravariant tensors*. In a similar manner, the quotient space $T^{*m}=V^{*m}/S(V^{*m})$ can be defined, and is

called the set of *m*th *order covariant tensors*. Finally, the quotient space of mixed tensors, $T^{k,*m} = V^{k,*m}/S(V^{k,*m})$ can be defined in an analogous fashion. The terms contravariant and covariant will in general be dropped, being clear from context (cf. Spivak,[13] pp. 4—8 to 4—12). The direct sum $T(V)$, denoted by $\oplus$,

$$T(V) = T^{0,*0} \oplus T^{1,*0} \oplus T^{0,*1} \oplus \cdots = \bigoplus_{k=0}^{\infty} \bigoplus_{m=0}^{\infty} T^{k,*m}$$

where $T^{0,*0} \approx \mathrm{R}$, is called the *tensor algebra* of V. Consider an element in $T^k(V)$, denoted A; A is called *alternating* or *skew-symmetric* if

$$A(v_1, \ldots, v_i, \ldots, v_j, \ldots, v_k) = -A(v_1, \ldots, v_j, \ldots, v_i, \ldots, v_k) \quad \forall v_1, \ldots, v_j \in V.$$

The set of all alternating *k*th order tensors is denoted $\Lambda^k(V)$, and is clearly a subspace of $T^k(V)$. If $L: V \to W$ is a linear transformation, then $L^*: T^k(W) \to T^k(V)$ is defined by $(L^* \circ T^k)(v_1, \ldots, v_k) = T^k[L(v_1), \ldots, L(v_k)]$. In particular, if $L: V \to W$, then

$$L^*(u \wedge v) = (L^*(u)) \wedge (L^*(v));$$

$\Lambda(V) = \Lambda^0(V) \oplus \cdots \oplus \Lambda^n(V)$, n equals dimension of V, is the *contravariant exterior algebra* of V, while $\Lambda(V^*) = \Lambda^0(V^*) \oplus \cdots \oplus \Lambda^n(V^*)$ is the *covariant exterior algebra* of V [which is defined in a manner entirely analogous to $\Lambda(V)$]. This work will concentrate entirely on exterior algebra. Multiplication in the exterior algebra $\Lambda(V)$ is denoted by "$\wedge$" the *exterior* or *wedge product*, a natural generalization of the three-dimensional cross product operation on two vectors. The exterior algebra is a graded algebra: if $u \in \Lambda^k(V)$, $v \in \Lambda^l(V)$, then $u \wedge v \in \Lambda^{k+l}(V)$. The exterior product obeys the following properties:

$$\left.\begin{aligned} u \wedge v &= (-1)^{kl} v \wedge u, \\ (u_1 + u_2) \wedge v &= (u_1 \wedge v) + (u_2 \wedge v), \\ u \wedge (v_1 + v_2) &= (u \wedge v_1) + (u \wedge v_2), \\ u \wedge (v \wedge w) &= (u \wedge v) \wedge w, \\ (au) \wedge v &= u \wedge (av) = a[u \wedge v] \end{aligned}\right\} \begin{aligned} &u, u_1, u_2 \in \Lambda^k(V), \\ &v, v_1, v_2 \in \Lambda^l(V), \\ &w \in \Lambda^m(V), \\ &a \in \mathrm{R}. \end{aligned}$$

If $\{e_1, \ldots, e_n\}$ is a basis for V, then $\{e_{i_1} \wedge \cdots \wedge e_{i_l} \mid i_1, \ldots, i_l = 1, \ldots, n\}$ is a basis for $\Lambda^l(V)$. In particular, note that $e_{i_1} \wedge \cdots \wedge e_{i_n}$ $(i_1, \ldots, i_n = 1, \ldots, n)$ is a basis vector for $\Lambda^n(V)$. Since $\Lambda^n(V)$ is one-dimensional, the sign on this basis vector can be either positive or negative, corresponding to a choice in *orientation* (cf. "right-handed" and "left-handed" coordinates in R^3).

Let $\langle\ ,\ \rangle: V \times V \to \mathrm{R}$ be the standard sum-of-squares *inner product* on V, positive definite and symmetric in its arguments. Choose an orthonormal basis for V, $\{e_1, \ldots, e_n\}$. Let $a \in \Lambda^k(V)$, $b \in \Lambda^l(V)$,

$$a = \sum_{i_1 < \cdots < i_k} a(i_1, \ldots, i_k) e_{i_1} \wedge \cdots \wedge e_{i_k}, \quad a(i_1, \ldots, i_k) \in \mathrm{R},$$

$$b = \sum_{i_1 < \cdots < i_l} b(i_1, \ldots, i_l) e_{i_1} \wedge \cdots \wedge e_{i_l}, \quad b(i_1, \ldots, i_l) \in \mathrm{R},$$

the inner product of a and b, denoted $\langle a, b \rangle$ is defined by

$$\langle a, b \rangle = \begin{cases} \sum_{i_1 < \cdots < i_k} a(i_1, \ldots, i_k) b(i_1, \ldots, i_k) \langle e_{i_1}, e_{i_1} \rangle \cdots \langle e_{i_k}, e_{i_k} \rangle, & k = l, \\ 0, & k \neq l. \end{cases}$$

The *Hodge star operator*, denoted $*, *: \Lambda^k(V) \to \Lambda^{n-k}(V)$, is well defined by the requirement that for any orthonormal basis $e_1, \ldots, e_n$ of V,

$$*: (e_1 \wedge \cdots \wedge e_k) = \pm (e_{k+1} \wedge \cdots \wedge e_n),$$

where the plus sign is chosen if $+e_1 \wedge \cdots \wedge e_k \wedge e_{k+1} \wedge \cdots \wedge e_n$ is in the basis for $\Lambda^n(V)$, and the minus sign is otherwise chosen.

The requirement on the inner product and Hodge star operator that the basis be orthonormal can be relaxed, and the interested reader is referred to the bibliography (Warner,[7] Flanders,[1] Loomis-Sternberg,[11] Spivak[10]).

Example (R^3): Choose a rectangular set of orthonormal basis vectors $\{u_x, u_y, u_z\}$:

$\Lambda(\mathrm{R}^3)$	Basis
$\Lambda^0(\mathrm{R}^3)$	1
$\Lambda^1(\mathrm{R}^3)$	u_x, u_y, u_z
$\Lambda^2(\mathrm{R}^3)$	$u_y \wedge u_z, u_z \wedge u_x, u_x \wedge u_y$
$\Lambda^3(\mathrm{R}^3)$	$u_x \wedge u_y \wedge u_z$

Dual Forms

$*1 = u_x \wedge u_y \wedge u_z$,	$*u_y \wedge u_z = u_x$,
$*u_x = u_y \wedge u_z$,	$*u_z \wedge u_x = u_y$,
$*u_y = u_z \wedge u_x$,	$*u_x \wedge u_y = u_z$,
$*u_z = u_x \wedge u_y$,	$*u_x \wedge u_y \wedge u_z = -1$.

A zero form may be interpreted physically as a scalar, while a 1-form may be interpreted as a directed line segment, a 2-form as a directed area, and a 3-form as a directed volume.

Example (Space—Time): (For an extensive discussion of the mathematics underlying space—time, the reader is referred to Penrose.[3]) Choose a rectangular set of orthonormal basis vectors $\{dx, dy, dz, icdt\}$ where $i = \sqrt{-1}$ and $c =$ speed of light, with orientation $+dx \wedge dy \wedge dz \wedge icdt$:

Λ (Space–Time)	Basis
Λ^0	1
Λ^1	$dx,\ dy,\ dz,\ icdt$
Λ^2	$dy\wedge dz,\ dz\wedge dx,\ dx\wedge dy,\ dx\wedge icdt,\ dy\wedge icdt,\ dz\wedge icdt$
Λ^3	$dy\wedge dz\wedge icdt,\ dz\wedge dx\wedge icdt,\ dx\wedge dy\wedge icdt,\ dx\wedge dy\wedge dz$
Λ^4	$dx\wedge dy\wedge dz\wedge icdt$

Dual Forms

$$*1 = dx\wedge dy\wedge dz\wedge icdt, \qquad *dx\wedge icdt = dy\wedge dz,$$
$$*dx = dy\wedge dz\wedge icdt, \qquad *dy\wedge icdt = dz\wedge dx,$$
$$*dy = dz\wedge dx\wedge icdt, \qquad *dz\wedge icdt = dx\wedge dy,$$
$$*dz = dx\wedge dy\wedge icdt, \qquad *dy\wedge dz\wedge icdt = -dx,$$
$$*icdt = -dx\wedge dy\wedge dz, \qquad *dz\wedge dx\wedge icdt = -dy,$$
$$*dy\wedge dz = dx\wedge icdt, \qquad *dx\wedge dy\wedge icdt = -dz$$
$$*dz\wedge dx = dy\wedge icdt, \qquad *dx\wedge dy\wedge dz = icdt,$$
$$*dx\wedge dy = dz\wedge icdt, \qquad *dx\wedge dy\wedge dz\wedge icdt = 1.$$

Note: $*{}^{\circ}*{}^{\circ}u_k = (-1)^k u_k,\ u_k \in \Lambda^k,\ k = 0,1,2,3,4.$

B. Differentiable manifolds

Let X be a set, U an open subset of X, and m a map, $m: U \to V \subset \mathrm{R}^n$ where m is bijective (one-one and onto). The pair (m, U) defines a *chart* on X; m specifies *local coordinates* on a subset of X. Consider two charts on X, (m_1, U_1) and (m_2, U_2); suppose $m_1 m_2^{-1}, m_2 m_1^{-1}: \mathrm{R}^n \to \mathrm{R}^n$ are C^k functions, i.e., differentiable k times but not $(k+1)$. $m_1 m_2^{-1}$ and $m_2 m_1^{-1}$ are called *transition functions*. A collection of charts on a set X is denoted A; A is called an *atlas* for X if the chart domains cover X, and the associated transition functions have open domains and are C^∞. A *complete atlas* is the union of all possible atlases for a set X. A *differentiable manifold* is a set X together with a complete atlas. Intuitively, a differentiable manifold is a union of nondisjoint sets, each of which is locally diffeomorphic to R^n, which is pieced together by the transition functions.

Let X and Y be differentiable manifolds. Choose any chart on X and Y with coordinate maps m_x and m_y, respectively. Then $f: X \to Y$ is defined by the composite map $m_y^{-1} \circ f \circ m_x$. Let p be a point in R^n, v a vector in R^n. To every function f defined in the neighborhood of p, associate a number called the *directional derivative* of f in the direction v at p, denoted $D_v f(p)$ and defined by

$$D_v f(p) = \frac{d}{dt} f(p+tv)\bigg|_{t=0}$$

Consider now the manifold X; the *tangent vector* to X at p in the direction v is a map which associates with every C^∞ function f, defined on a neighborhood of p, a real number $D_v f(p)$ such that

(i) $f_1 = f_2$ implies $D_v f_1(p) = D_v f_2(p)$,

(ii) $D_v(f+g)(p) = D_v f(p) + D_v g(p)$,

(iii) $D_v(f\cdot g)(p) = [D_v f(p)]g(p) + f(p)[D_v g(p)]$.

The *tangent space* of X at p is the set of all tangent vectors, for all $v \in \mathrm{R}^N$. The tangent space of X at p can be shown to be a vector space, and thus has an associated dual vector space, called the *cotangent space* of X at p, the set of all linear functionals on the tangent space. The *tangent bundle* of a manifold X is the direct product of the set of all tangent spaces at all points $p \in X$; the *cotangent bundle* is the direct product of the set of all cotangent spaces at all points $p \in X$ with X. A *Riemannian differentiable manifold* is a differentiable manifold with a prescribed norm on the tangent bundle.

Example: Let X be a finite-dimensional vector space. Choose a basis for X, $\{e_1, \ldots, e_n\}$, so $x \in X$ can be expressed as $x = x_1 e_1 + \cdots + x_n e_n$. Define the coordinate map $m(x_1 e_1 + \cdots + x_n e_n) = (x_1, \ldots, x_n)$. An atlas for X is the set of coefficients, with respect to the basis $\{e_1 \cdots e_n\}$, of all points $x \in X$. A second atlas for X is the set of coefficients, with respect to a different basis $\{e_1', \ldots, e_n'\}$, of all points $x \in X$. The transition functions are given by a C^∞ linear transformation describing the change of basis. A complete atlas can be generated by considering all possible sets of basis vectors for X; thus, X is a differentiable manifold. The tangent space and cotangent space of X at a point p are clearly n-dimensional, so the tangent bundle and cotangent bundles are locally diffeomorphic to R^{2n}. Together with the standard Euclidean norm on the tangent bundle, X is a Riemannian differentiable manifold.

C. Exterior calculus

If $f: \mathrm{R}^n \to \mathrm{R}$ is a scalar differentiable function of n variables, then f is a *zero differential form*, or 0-form. The total differential of f, $df(x_1, \ldots, x_n) = (\partial f/\partial x_1)dx_1 + \cdots + (\partial f/\partial x_n)dx_n$ is called a *one differential form* or 1-*form* if each component $\partial f/\partial x_k$, $k = 1, \ldots, n$ is differentiable. Note that f may be considered in Λ^0, while df is an element of Λ^1. The *exterior derivative* generalizes the concept of a total differential using exterior algebra:

Theorem[10]: Let $u \in \Lambda^k$. Then the exterior derivative of u is $du \in \Lambda^{k+1}$, and is defined by

$$du = \sum_{i_1 < \cdots < i_k} du_{i_1,\ldots,i_k}\ dx_{i_1}\wedge\cdots\wedge dx_{i_k},$$

where $du_{i_1,\ldots,i_k}$ is the total differential of the i_j component of u, and the exterior derivative d obeys the following properties:

$$\left.\begin{array}{ll} \text{(i)} & d(u+v) = du + dv \\ \text{(ii)} & d(u\wedge v) = du\wedge v + (-1)^k u\wedge dv \\ \text{(iii)} & d(du) = 0 \Rightarrow d\circ d \cong 0 \end{array}\right\} \quad u \in \Lambda^k,\ \ v \in \Lambda^l.$$

A differential form u is called *closed* if $du=0$, *exact* if $dv=u$; it can be shown every exact form is closed, but the converse is not true. The *adjoint* δ of the exterior derivative is defined such that

$$\langle du, v\rangle = \langle u, \delta v\rangle, \quad u\in\Lambda^k, \quad v\in\Lambda^{k+1}.$$

It can be shown $\delta = *\circ d\circ *$. A differential form is called *coclosed* if $\delta u=0$, *coexact* if $\delta v=u$. The *Laplace–Beltrami operator* is defined as $\Delta = d\delta+\delta d$, and is linear, $\Delta:\Lambda^k\to\Lambda^k$ $(k=0,\ldots,n)$. Elements in the kernel of Δ are called *harmonic*, and the set of all such k-forms is denoted $H^k=\{u:\Delta u=0,\ u\in\Lambda^k\}$. It can be shown the Laplace–Beltrami operator is *elliptic* (Warner[7], pp. 250–251).

A question of great practical interest is solving $\Delta u = v$, given v subject to suitable boundary conditions. For the special case where the underlying manifold X is compact, this question has been answered by

Theorem[7] (Hodge–DeRham–Kodaira): $\Delta u=v$ has a unique solution $u\in\Lambda^k$ iff $v\in\Lambda^k$ is orthogonal to H^k. Furthermore, Λ^k can be decomposed into a direct sum of three mutually orthogonal vector spaces,

$$\begin{aligned}\Lambda^k &= H^k\oplus\Delta(\Lambda^k)\\ &= H^k\oplus(d\delta+\delta d)(\Lambda^k)\\ &= H^k\oplus d(\Lambda^{k-1})\oplus\delta(\Lambda^{k+1})\end{aligned}$$

and H^k is finite-dimensional.

Example ($\mathbf{R}^3$): $\Lambda^0\xrightarrow{d}\Lambda^1\xrightarrow{d}\Lambda^2\xrightarrow{d}\Lambda^3$.

For simplicity choose a rectangular orthonormal basis $\{dx,dy,dz\}$ with orientation $+dx\wedge dy\wedge dz$. Then

$$f_0\in\Lambda^0,\quad df_0=\frac{\partial f_0}{\partial x}dx+\frac{\partial f_0}{\partial y}dy+\frac{\partial f_0}{\partial z}dz\in\Lambda^1;$$

$$f_1=f_{1x}dx+f_{1y}dy+f_{1z}dz\in\Lambda^1,$$

$$\begin{aligned}df_1 &= \left(\frac{\partial f_{1x}}{\partial x}dx+\frac{\partial f_{1x}}{\partial y}dy+\frac{\partial f_{1x}}{\partial z}dz\right)\wedge dx\\ &\quad+\left(\frac{\partial f_{1y}}{\partial x}dx+\frac{\partial f_{1y}}{\partial y}dy+\frac{\partial f_{1y}}{\partial z}dz\right)\wedge dy\\ &\quad+\left(\frac{\partial f_{1z}}{\partial x}dx+\frac{\partial f_{1z}}{\partial y}dy+\frac{\partial f_{1z}}{\partial z}dz\right)\wedge dz\\ &= \left(\frac{\partial f_{1z}}{\partial y}-\frac{\partial f_{1y}}{\partial z}\right)dy\wedge dz+\left(\frac{\partial f_{1x}}{\partial z}-\frac{\partial f_{1z}}{\partial x}\right)dz\wedge dx\\ &\quad+\left(\frac{\partial f_{1y}}{\partial x}-\frac{\partial f_{1x}}{\partial y}\right)dx\wedge dy\in\Lambda^2;\end{aligned}$$

$$f_2=f_{2x}dy\wedge dz+f_{2y}dz\wedge dx+f_{2z}dx\wedge dy\in\Lambda^2,$$

$$\begin{aligned}df_2 &= \left(\frac{\partial f_{2x}}{\partial x}dx+\frac{\partial f_{2x}}{\partial y}dy+\frac{\partial f_{2x}}{\partial z}dz\right)\wedge dy\wedge dz\\ &\quad+\left(\frac{\partial f_{2y}}{\partial x}dx+\frac{\partial f_{2y}}{\partial y}dy+\frac{\partial f_{2y}}{\partial z}dz\right)\quad dz\quad dx\\ &\quad+\left(\frac{\partial f_{2z}}{\partial x}dx+\frac{\partial f_{2z}}{\partial y}dy+\frac{\partial f_{2z}}{\partial z}dz\right)\wedge dx\wedge dy\\ &= \left(\frac{\partial f_{2x}}{\partial x}+\frac{\partial f_{2y}}{\partial y}+\frac{\partial f_{2z}}{\partial z}\right)\wedge dx\wedge dy\wedge dz\in\Lambda^3;\end{aligned}$$

$$f_3\,dx\wedge dy\wedge dz\in\Lambda^3,\quad df_3=0.$$

Similarly, it can be shown that

$$\delta f_3=\frac{\partial f_3}{\partial x}dy\wedge dz+\frac{\partial f_3}{\partial y}dz\wedge dx+\frac{\partial f_3}{\partial z}dx\wedge dy\in\Lambda^2,$$

$$\delta f_2=\left(\frac{\partial f_{2z}}{\partial y}-\frac{\partial f_{2y}}{\partial z}\right)dx+\left(\frac{\partial f_{2x}}{\partial z}-\frac{\partial f_{2z}}{\partial x}\right)dy+\left(\frac{\partial f_{2y}}{\partial x}-\frac{\partial f_{2x}}{\partial y}\right)\in\Lambda^1,$$

$$\delta f_1=\frac{\partial f_{1x}}{\partial x}+\frac{\partial f_{1y}}{\partial y}+\frac{\partial f_{1z}}{\partial z}\in\Lambda^0,$$

$$\delta f_0=0.$$

Thus, the exterior derivative subsumes the operations of gradient, curl, and divergence. The Laplace–Beltrami operator $\Delta=d\delta+\delta d$ simplifies for this choice of basis. Define $D=-(\partial^2/\partial x^2+\partial^2/\partial y^2+\partial^2/\partial z^2)$, so

$$\Delta f_0=-\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)f_0=Df_0,$$

$$\Delta f_1=(Df_{1x})\,dx+(Df_{1y})\,dy+(Df_{1z})\,dz,$$

$$\Delta f_2=(Df_{2x})\,dy\wedge dy+(Df_{2y})\,dz\wedge dx+(Df_{2z})\,dx\wedge dy,$$

$$\Delta f_3=-\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)f_3\,dx\wedge dy\wedge dz=(Df_3)\,dx\wedge dy\wedge dz.$$

If the manifold is restricted to a compact subset of $\mathbf{R}^3$, then the Hodge–DeRham–Kodaira decomposition theorem shows that

$$\begin{aligned}f_0&=f_0^H\oplus\delta f_1^C,\\ f_1&=f_1^H\oplus df_0^E\oplus\delta f_2^C,\\ f_2&=f_2^H\oplus df_1^E\oplus\delta f_3^C,\\ f_3&=f_3^H\oplus df_2^E,\end{aligned}$$

where the superscripts H, E, and C denote harmonic, exact, and coexact, respectively. To be more explicit,

$$\begin{aligned}f_0&=f_0^H\oplus\delta(f_{1x}^C dx+f_{1y}^C dy+f_{1z}^C dz)\\ &=f_0^H\oplus\left[\frac{\partial}{\partial x}f_{1x}^C+\frac{\partial}{\partial y}f_{1y}^C+\frac{\partial}{\partial z}f_{1z}^C\right],\\ f_1&=f_1^H\oplus df_0^E\oplus\delta[f_{2x}^C dy\wedge dz+f_{2y}^C dz\wedge dx+f_{2z}^C dx\wedge dy]\\ &=f_1^H\oplus\left[\frac{\partial}{\partial x}f_0^E dx+\frac{\partial}{\partial y}f_0^E dy+\frac{\partial}{\partial z}f_0^E dz\right]\\ &\quad\oplus\left[\left(\frac{\partial}{\partial y}f_{2z}^C-\frac{\partial}{\partial z}f_{2y}^C\right)dx+\left(\frac{\partial}{\partial z}f_{2x}^C-\frac{\partial}{\partial x}f_{2z}^C\right)dy\right.\\ &\quad\left.+\left(\frac{\partial}{\partial x}f_{2y}^C-\frac{\partial}{\partial y}f_{2x}^C\right)dz\right],\\ f_2&=f_2^H\oplus d[f_{1x}^E dx+f_{1y}^E dy+f_{1z}^E dz]\oplus\delta[f_3^C dx\wedge dy\wedge dz]\\ &=f_2^H\oplus\left[\left(\frac{\partial}{\partial y}f_{1z}^E-\frac{\partial}{\partial z}f_{1y}^E\right)dy\wedge dz\right.\\ &\quad+\left(\frac{\partial}{\partial z}f_{1x}^E-\frac{\partial}{\partial x}f_{1z}^E\right)dz\wedge dx+\left(\frac{\partial}{\partial x}f_{1y}^E-\frac{\partial}{\partial y}f_{1x}^E\right)dx\wedge dy\\ &\quad\oplus\left[\frac{\partial}{\partial x}f_3^C dy\wedge dz+\frac{\partial}{\partial y}f_3^C dz\wedge dx+\frac{\partial}{\partial z}f_3^C dx\wedge dy\right],\\ f_3&=f_3^H\oplus d[f_{2x}^E dy\wedge dz+f_{2y}^E dz\wedge dx+f_{2z}^E dx\wedge dy]\\ &=f_3^H\oplus\left[\left(\frac{\partial}{\partial x}f_{2x}^E+\frac{\partial}{\partial y}f_{2y}^E+\frac{\partial}{\partial z}f_{2z}^E\right)dx\wedge dy\wedge dz\right].\end{aligned}$$

For the special case where the manifold is simply connected,

$$f_0^H=\text{const},\quad f_1^H=0,$$

$$f_2^H = 0, \quad f_3^H = (\text{const})\, dx \wedge dy \wedge dz.$$

In other words, any scalar 0-form can be written as the sum of a constant function plus the divergence of a function, any 1-form can be expressed as the curl of a vector valued function plus the gradient of a scalar function, any 2-form can be written as the curl of a vector valued function plus the gradient of a scalar function, and any 3-form can be written as the sum of a constant function plus the divergence of a vector valued function.

A second approach to this decomposition is to expand each f_k $(k=0,1,2,3)$ in eigenfunctions of the Laplace–Beltrami operator:

$$f_0 = f_0^H + \sum_{i=1}^{\infty} \langle f_0, \delta u_i^C \rangle \delta u_i^C,$$

$$f_1 = \sum_{i=1}^{\infty} \langle f_1, du_i \rangle du_i^{0E} + \sum_{j=1}^{\infty} \langle f_1, \delta u_j^{2C} \rangle \delta u_j^{2C},$$

$$f_2 = \sum_{i=1}^{\infty} \langle f_2, du_i^{1E} \rangle du_i^{1E} + \sum_{j=1}^{\infty} \langle f_2, \delta u_j^{3C} \rangle \delta u_j^{3C},$$

$$f_3 = f_3^H + \sum_{i=1}^{\infty} \langle f_3, du_i^{2E} \rangle du_i^{2E}.$$

The $\{u_j^i\}$, $j=1,2,\ldots$, $i=0,1,2,3$ are eigenfunctions of the Laplace–Beltrami operator

$$\Delta : \Lambda^i \to \Lambda^i, \quad i=0,1,2,3\ ,$$

$$\Delta u_j^i = \lambda_j u_j^i.$$

Various properties of these eigenfunctions are discussed in the text.

[1] H. Flanders, *Differential Forms* (Academic, New York, 1963), pp. 44–48.
[2] F.J. Dyson, Bull. Am. Math. Soc. 78, 635 (1972).
[3] R. Penrose, "Structure of Space–Time," in *Battelle Rencontres* (Benjamin, New York, 1968), pp. 121–235.
[4] A. Sommerfeld, *Electrodynamics* (Academic, New York, 1964), pp. 212–241.
[5] A.O. Barut, *Electrodynamics and Classical Theory of Fields and Particles* (MacMillan, New York, 1964), p. 121.
[6] C.W. Misner, K.S. Thorne, and J.A. Wheeler, *Gravitation* (Freeman, San Francisco, 1970, 1973), pp. 79–129.
[7] F.W. Warner, *Foundations of Differentiable Manifolds and Lie Groups* (Scott Foresman, Glenview, Ill., 1971).
[8] R. Abraham, *Foundations of Mechanics* (Benjamin, New York, 1967).
[9] H. Goldstein, *Classical Mechanics* (Addison-Wesley, Reading, Mass., 1950), Chap. 11.
[10] M. Spivak, *Calculus on Manifolds* (Benjamin, New York, 1965).
[11] L. Loomis and S. Sternberg, *Advanced Calculus* (Addison-Wesley, Reading, Mass., 1968).
[12] E. Nelson, *Tensor Analysis* (Princeton, U.P., Princeton, N.J., 1967).
[13] M. Spivak, *Differential Geometry* (Publish or Perish, Boston, 1970), Vol. I.