

Eigenvalue problem for an infinite tridiagonal matrix

C. Wongtawatnugool and S. Y. Wu

Department of Physics, University of Louisville, Louisville, Kentucky 40292

C. C. Shih

Department of Physics and Astronomy, The University of Tennessee, Knoxville, Tennessee 37916

(Received 22 July 1980; accepted for publication 20 November 1980)

A method is developed for the calculation of the eigenvectors of an infinite tridiagonal matrix. Possible application of this method to study the problem of localization in a disordered linear chain is also discussed.

PACS numbers: 02.10.Sp, 63.50.+x

I. INTRODUCTION

Many interesting physical systems may be described by a matrix of tridiagonal form. For such systems, the solution of the eigenvalue problem usually starts with the truncation of the matrix to finite but sufficiently large size. Then methods such as Dean's method of negative mode counting¹ can be used to determine the eigenvalues. Recently, Dy, Wu, and Wongtawatnugool² developed a scheme which leads directly to the determination of the eigenvector once the eigenvalue is determined. There are two key points to their approach. (1) The calculation of the eigenvector is carried out concurrently with the determination of the eigenvalue. Thus, there is no redundancy in computing separately the eigenvector after the eigenvalue is determined, resulting in considerable saving of computing time. (2) The boundary conditions are incorporated into the process of calculation of the eigenvalues and the eigenvectors. Therefore they will be satisfied to the same degree of accuracy as those in the calculation of the eigenvalues.

However, in many situations, it is crucial to understand the eigenvalue spectrum and the corresponding eigenstates of an infinite system. The conventional method for treating such situations is to allow N , the number of particles in the corresponding finite system, to increase and then to examine how the properties of the system change as $N \rightarrow \infty$. The computational problem encountered in a process such as this can be formidable. Hence, the problem of the approach to infinity is always difficult to handle.

In general, the eigenvalue spectrum of an infinite system consists of two parts: a series of discrete eigenvalues and a series of continuous bands. The most convenient way of studying the eigenvalue spectrum of an infinite system is to calculate the diagonal elements of the resolvent operators (the Green's function). The poles of the diagonal elements of the Green's function yield the discrete eigenvalues, and the imaginary part of the trace of the Green's function determines the continuous spectrum. But how do we calculate the eigenstates for an infinite system? Can we still use the method of Dy, Wu, and Wongtawatnugool (DWW)?

It turns out that the eigenstates corresponding to the discrete spectrum can indeed be calculated using the DWW method. However, the method needs to be modified for the eigenstates corresponding to the continuous spectrum. In this work we shall develop a scheme to calculate the eigen-

states for the continuous spectrum and discuss its relationship with that for the corresponding finite system. Possible application of the method to study the problem of localization in a disordered system is also discussed.

II. REVIEW OF THE DWW METHOD FOR A FINITE SYSTEM

For a system described by a tridiagonal matrix of the form

$$H_{i'i} = E_i \delta_{i,i'} - V_{i,i+1} \delta_{i+1,i'} - V_{i,i-1} \delta_{i-1,i'}, \quad (1)$$

the matrix elements of the resolvent operator

$R = (Z - H)^{-1}$, i.e., the Green's function, can be written as³

$$R_{ii} = (A_i - V_{i,i+1} \Delta_{i+1}^+ V_{i+1,i} - V_{i,i-1} \Delta_{i-1}^- V_{i-1,i})^{-1}, \quad (2)$$

$$R_{i'i} = R_{ii} V_{i,i\pm 1} \Delta_{i\pm 1}^\pm \dots V_{i'\mp 1,i'} \Delta_{i'}^\pm$$

(- for $i > i'$; + for $i < i'$),

or

$$= \Delta_{i'}^\pm V_{i,i\mp 1} \Delta_{i\mp 1}^\pm \dots V_{i'\pm 1,i'} R_{i'i'} \quad (- \text{ for } i < i'; + \text{ for } i > i'), \quad (3)$$

where $A_i = (Z - E_i)$ and

$$\Delta_{i\pm 1}^\pm = (A_i - V_{i,i\pm 1} \Delta_{i\pm 1}^\pm V_{i\pm 1,i})^{-1}. \quad (4)$$

If the eigenvector v is expanded in the representation u_i defining the matrix H_{ij} , we have

$$v = \sum_i c_i u_i, \quad (5)$$

The eigenvalue equation $(\lambda - H)v = R^{-1}v = 0$ can then be reduced to

$$\sum_i R_{ji}^{-1} c_i = 0, \quad (6)$$

where

$$R_{ji}^{-1} = (Z - E_j) \delta_{ji} - V_{j,j+1} \delta_{j+1,i} - V_{j,j-1} \delta_{j-1,i}.$$

For a $c_l \neq 0$, Eq. (6) may be rewritten as

$$\sum_i R_{ji}^{-1} b_i = 0, \quad (7)$$

with $b_i = c_i/c_l$ and in particular $b_l = 1$. The eigenvector v will be determined (except for the normalization constant) if all the b_i 's are determined.

In DWW's approach² Eq. (7) is split into two parts corresponding to $j > l + 1$ or $j < l - 1$. In this situation we obtain

$$\tilde{R}_{\pm}^{-1} b^{\pm} = x^{\pm}, \quad (8)$$

where

$$\tilde{R}_{\pm}^{-1} = \begin{pmatrix} R_{l_{\pm 1}, l_{\pm 1}}^{-1} & R_{l_{\pm 1}, l_{\pm 2}}^{-1} & 0 & \dots & \dots \\ R_{l_{\pm 2}, l_{\pm 1}}^{-1} & R_{l_{\pm 2}, l_{\pm 2}}^{-1} & R_{l_{\pm 2}, l_{\pm 3}}^{-1} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix},$$

$$b^{\pm} = \begin{pmatrix} b_{l_{\pm 1}} \\ b_{l_{\pm 2}} \\ \dots \\ \dots \end{pmatrix}, \quad x^{\pm} = \begin{pmatrix} -R_{l_{\pm 1}, l}^{-1} \\ 0 \\ \dots \\ \dots \end{pmatrix} = \begin{pmatrix} V_{l_{\pm 1}, l} \\ 0 \\ \dots \\ \dots \end{pmatrix}. \quad (9)$$

The column vectors b^{\pm} can then be expressed as

$$b^{\pm} = \tilde{R}^{\pm} x^{\pm} \quad (10)$$

or

$$b_{l_{\pm 1}} = \tilde{R}_{l_{\pm 1}, l_{\pm 1}}^{\pm} V_{l_{\pm 1}, l} = \Delta_{l_{\pm 1}}^{\pm} V_{l_{\pm 1}, l},$$

$$\dots$$

$$\dots$$

$$b_{l_{\pm r}} = \tilde{R}_{l_{\pm r}, l_{\pm 1}}^{\pm} V_{l_{\pm 1}, l}$$

$$= \Delta_{l_{\pm r}}^{\pm} V_{l_{\pm r}, l_{\pm r+1}} \dots \Delta_{l_{\pm 1}}^{\pm} V_{l_{\pm 1}, l},$$

$$\dots \quad (11)$$

Using the identity given in Eq. (3), we may also write

$$b_{l_{\pm r}} = R_{l_{\pm r}, l} / R_{ll}. \quad (12)$$

III. THE REQUIREMENT OF THE BOUNDARY CONDITION

The boundary condition of the eigenvalue problem discussed in the previous section is given at $j = l$. From Eq. (7) we have at $j = l$,

$$R_{l, l+1}^{-1} b_{l+1} + R_{ll}^{-1} + R_{l, l-1}^{-1} b_{l-1} = 0. \quad (13)$$

Substituting Eq. (12) into Eq. (13) we obtain

$$R_{l, l+1}^{-1} \frac{R_{l+1, l}}{R_{ll}} + R_{ll}^{-1} + R_{l, l-1}^{-1} \frac{R_{l-1, l}}{R_{ll}} = 0. \quad (14)$$

On the other hand, since R is the inverse of R^{-1} we should have

$$R_{l, l+1}^{-1} R_{l+1, l} + R_{ll}^{-1} R_{ll} + R_{l, l-1}^{-1} R_{l-1, l} = 1. \quad (15)$$

Dividing Eq. (15) by R_{ll} leads to

$$R_{l, l+1}^{-1} \frac{R_{l+1, l}}{R_{ll}} + R_{ll}^{-1} + R_{l, l-1}^{-1} \frac{R_{l-1, l}}{R_{ll}} = \frac{1}{R_{ll}}. \quad (16)$$

At first glance, Eqs. (14) and (16) seem to be in contradiction. However, for a finite system the eigenvalues are actually the poles of R_{ll} . Hence, Eq. (16) is identical to Eq. (14) when $Z = \lambda$, where λ is the eigenvalue of the system, because $1/R_{ll} \rightarrow 0$ at $Z = \lambda$. We should also note that at $Z = \lambda$, even though $R_{ll} \rightarrow \infty$, the ratio $R_{l_{\pm r}, l} / R_{ll}$ is in fact a finite quantity given by Eq. (11).

The situation is different, however, for an infinite system. Indeed, for the discrete spectrum, the eigenvalues still correspond to the pole of R_{ll} so that Eqs. (14) and (16) are consistent with each other. But for the continuous spectrum

R_{ll} does not go to infinity in the region defined by the continuous spectrum. This then indicates that in this situation Eq. (12) must be modified so that the boundary condition will be consistent with Eq. (16).

IV. THE EIGENSTATES FOR THE CONTINUOUS SPECTRUM

As pointed out earlier, the density of states for a continuous spectrum can be calculated in terms of the imaginary part of the trace of the Green's function. Specifically,⁴

$$\rho(E) = -\frac{1}{\pi} \lim_{\epsilon \rightarrow 0} \frac{1}{N} \text{Tr} R(E + i\epsilon). \quad (17)$$

In this sense, the existence of the continuous spectrum can be viewed as follows. For a finite system, when the degree of freedom N is allowed to increase the density of distribution of the discrete eigenvalues will increase. In certain regions of the spectrum it may happen that as $N \rightarrow \infty$, series of discrete poles will collapse into a series of branch cuts, resulting in a series of continuous bands.

To determine the eigenstates corresponding to the continuous spectrum we shall follow essentially the same reasoning. The equation defining the Green's function may be written as

$$\sum_r R_{ir}^{-1}(E + i\epsilon) R_{r'l} = \delta_{il}. \quad (18)$$

Consider the case when $i = i' = l$. Equation (18) will reduce to

$$R_{l, l+1}^{-1} R_{l+1, l} + R_{ll}^{-1}(E + i\epsilon) R_{ll} + R_{l, l-1}^{-1} R_{l-1, l} = 1. \quad (19)$$

The imaginary part of Eq. (19) is then

$$R_{l, l+1}^{-1} \text{Im} R_{l+1, l} + \text{Im}(R_{ll}^{-1} R_{ll}) + R_{l, l-1}^{-1} \text{Im} R_{l-1, l} = 0.$$

As $\epsilon \rightarrow 0$, this equation becomes

$$R_{l, l+1}^{-1} \text{Im} R_{l+1, l} + R_{ll}^{-1}(E) \text{Im} R_{ll} + R_{l, l-1}^{-1} \text{Im} R_{l-1, l} = 0$$

or

$$R_{l, l+1}^{-1} \frac{\text{Im} R_{l+1, l}}{\text{Im} R_{ll}} + R_{ll}^{-1} + R_{l, l-1}^{-1} \frac{\text{Im} R_{l-1, l}}{\text{Im} R_{ll}} = 0. \quad (20)$$

If one compares Eq. (20) with Eq. (14), one is tempted to propose that

$$b_{l_{\pm 1}} = \lim_{\epsilon \rightarrow 0} \frac{\text{Im} R_{l_{\pm 1}, l}}{\text{Im} R_{ll}}.$$

To see whether this is the case we shall split Eq. (18) into two parts: $i' = l, i > l$ and $i' = l, i < l$. If again, only the imaginary part of the equation is considered, we obtain as $\epsilon \rightarrow 0$,

$$R_{l_{\pm 1}, l}^{-1} \text{Im} R_{ll} + R_{l_{\pm 1}, l_{\pm 1}}^{-1} \text{Im} R_{l_{\pm 1}, l}$$

$$+ R_{l_{\pm 1}, l_{\pm 2}}^{-1} \text{Im} R_{l_{\pm 2}, l} = 0,$$

$$R_{l_{\pm 2}, l_{\pm 1}}^{-1} \text{Im} R_{l_{\pm 1}, l} + R_{l_{\pm 2}, l_{\pm 2}}^{-1} \text{Im} R_{l_{\pm 2}, l}$$

$$+ R_{l_{\pm 2}, l_{\pm 3}}^{-1} \text{Im} R_{l_{\pm 3}, l} = 0,$$

$$\dots \quad (21)$$

When Eq. (21) is divided by $\text{Im} R_{ll}$, it becomes

$$R_{l_{\pm 1}, l_{\pm 1}}^{-1} \frac{\text{Im} R_{l_{\pm 1}, l}}{\text{Im} R_{ll}} + R_{l_{\pm 1}, l_{\pm 2}}^{-1} \frac{\text{Im} R_{l_{\pm 2}, l}}{\text{Im} R_{ll}} = -R_{l_{\pm 1}, l}^{-1},$$

$$R_{l \pm 2, l \pm 1}^{-1} \frac{\text{Im}R_{l \pm 1, l}}{\text{Im}R_{ll}} + R_{l \pm 2, l \pm 2}^{-1} \times \frac{\text{Im}R_{l \pm 2, l}}{\text{Im}R_{ll}} + R_{l \pm 2, l \pm 3}^{-1} \frac{\text{Im}R_{l \pm 3, l}}{\text{Im}R_{ll}} = 0, \quad (22)$$

Comparison of Eq. (22) with Eq. (8) indicates that $b_{l \pm r}$ and $\text{Im}R_{l \pm r, l}/\text{Im}R_{ll}$, in fact, satisfy the same set of equations. Hence they must be proportional to each other. Since it had been shown by Dy, Wu, and Wongtawatnugool that $b_{l \pm r} = R_{l \pm r, l}/R_{ll}$ for a finite system, one may expect that for an infinite system the constant of proportionality is 1, or $b_{l \pm r} = \text{Im}R_{l \pm r, l}/\text{Im}R_{ll}$.

To check on this point consider the quantity $\text{Im}R_{l+1, l}/\text{Im}R_{ll}$. In the region defined by the continuous spectrum $\Delta_{l \pm 1}^{\pm}$ are complex quantities. If we denote $\Delta_{l+1}^{+} = a + ib$ and $\Delta_{l-1}^{-} = a' + ib'$, it can be shown that

$$\frac{\text{Im}R_{l+1, l}}{\text{Im}R_{ll}} = aV_{l+1, l} + V_{l+1, l}(A_l - V_{l, l+1}aV_{l+1, l} - V_{l, l-1}a'V_{l-1, l})[b/(V_{l, l+1}bV_{l+1, l} + V_{l, l-1}b'V_{l-1, l})].$$

For a finite system composed of N particles (regardless how large N is) $\Delta_{l \pm 1}^{\pm}$ must be real.⁵ Hence

$$\Delta_{l+1}^{+}(N) = a_N$$

and

$$\Delta_{l-1}^{-}(N) = a'_N.$$

The eigenvalues of the system are determined by the poles of $R_{ll}(N)$. Since $R_{ll}(N)$ is now given by

$$R_{ll}(N) = (A_l - V_{l, l+1}a_N V_{l+1, l} - V_{l, l-1}a'_N V_{l-1, l})^{-1},$$

hence, at any given eigenvalue of the system,

$$A_l - V_{l, l+1}a_N V_{l+1, l} - V_{l, l-1}a'_N V_{l-1, l} = 0.$$

This means that if we are examining the correspondence between the infinite system and the finite system we should have

$$\frac{\text{Im}R_{l+1, l}}{\text{Im}R_{ll}} \rightarrow a_N V_{l+1, l} = \Delta_{l+1}^{+}(N) V_{l+1, l} = \frac{R_{l+1, l}(N)}{R_{ll}(N)}. \quad (23)$$

Since $\text{Im}R_{l+1, l}/\text{Im}R_{ll}$ reduces to $R_{l+1, l}(N)/R_{ll}(N)$ for a finite system, we can then conclude that

$$b_{l+1} = \lim_{\epsilon \rightarrow 0} \frac{\text{Im}R_{l+1, l}}{\text{Im}R_{ll}}.$$

Similar argument can be used to show that in general

$$b_{l \pm r} = \lim_{\epsilon \rightarrow 0} \frac{\text{Im}R_{l \pm r, l}}{\text{Im}R_{ll}}. \quad (24)$$

V. AN ILLUSTRATIVE EXAMPLE: THE EIGENVECTORS OF A PERIODIC MONATOMIC CHAIN

To further establish the validity of Eq. (24) we use it to determine the eigenvectors of a periodic monatomic chain. The situation to be studied is the dynamics of a chain with

atomic mass m and nearest neighbor interaction characterized by the force constant γ . The analysis is also applicable for the electronic structure of a chain with an array of identical potentials.

Using the reduced unit of $\gamma/m = 1$, the equation describing the dynamics of the system can be written as

$$(2 - \omega^2)u_l = u_{l+1} + u_{l-1}. \quad (25)$$

For such a system, it can be seen that

$$\Delta_{l \pm 1}^{\pm} = \Delta_{l \pm 2}^{\pm} = \dots = \Delta,$$

where Δ satisfies the equation

$$\Delta = 1/(A - \Delta) \quad (26)$$

with

$$A = 2 - \omega^2.$$

Using Eqs. (2) and (3), we can write

$$R_{ll} = 1/(A - 2\Delta) = \Delta/(1 - \Delta^2)$$

and

$$R_{l \pm r, l} = \Delta^r R_{ll} = \Delta^{r+1}/(1 - \Delta^2). \quad (27)$$

Denoting

$$\Delta = ae^{i\alpha}$$

and

$$1 - \Delta^2 = be^{i\beta}$$

so that

$$\tan\beta = -a^2 \sin 2\alpha / (1 - a^2 \cos 2\alpha), \quad (28)$$

we obtain

$$b_{l \pm r} = a^r [\cos r\alpha + \sin r\alpha \cot(\alpha - \beta)]. \quad (29)$$

For the system under consideration, the region of continuous spectrum is determined by examining the frequency spectrum $\rho(\omega^2)$. It is easily seen that $\rho(\omega^2)$ does not vanish only in the interval $0 < \omega^2 < 4$. In this interval Δ is complex and is given by

$$\Delta = \frac{1}{2} \{ (2 - \omega^2) \pm i[\omega^2(4 - \omega^2)]^{1/2} \}.$$

The quantity a , the absolute value of Δ , is then 1. Using Eq. (28), it can be shown that

$$\beta = \alpha + \pi/2.$$

Substituting these results into Eq. (29), we obtain

$$b_{l \pm r} = \cos r\alpha, \quad (30)$$

where

$$\tan\alpha = [\omega^2(4 - \omega^2)]^{1/2} / (2 - \omega^2).$$

This is just the expected result.

VI. THE APPLICATION TO DISORDERED SYSTEMS

When studying the properties of disordered systems the most difficult problem is the determination of the eigenvalue spectrum and the eigenstates for an infinite system. Over the years various numerical schemes and analytic approximations were invented to handle the task.^{1,6,7} However, the calculation involved is usually quite substantial so that it always presents itself as a formidable problem. In particular, the determination of the eigenstates is most of the time difficult to deal with.

Because there really does not exist any means to calculate exactly the properties of an infinite disordered system, one of the most frequently used method is the numerical approach. This approach is to first solve the eigenvalue problem for a series of increasing but finite systems, and then to examine whether the calculated properties approach their respective limits. In fact, the numerical studies of the frequency spectra of disordered systems by Dean and his co-workers are typical examples of this kind of treatment.¹

One of the most challenging problems in disordered systems is the problem associated with the localization. Since the pioneering works of Mott and Twose,⁸ Borland,⁹ and Halperin,¹⁰ much effort has been devoted to the study of localization in one-dimensional disordered systems. But only recently has there been serious attempts to carry out direct eigenvector analysis of such problems.¹¹⁻¹⁵ The reason is that, in order to compute the eigenvectors of a long disordered chain, very precise determination of the eigenvalue is a necessity. Because of the intrinsic progression of errors in the computation, the numerical calculation of the eigenvectors becomes increasingly difficult for larger and larger systems. The method developed in Sec. IV, however, provides a scheme to avoid this major problem, i.e., the precise calculation of the eigenvalues.

From Eq.(24) it is seen that the eigenvector can be determined by

$$b_{l \pm r} = \lim_{\substack{N \rightarrow \infty \\ \epsilon \rightarrow 0}} \frac{\text{Im}R_{l \pm r, l}(N, E + i\epsilon)}{\text{Im}R_{ll}(N, E + i\epsilon)}, \quad (31)$$

where $R_{ll}(N)$ is the matrix element for the resolvent operator R for a system of N particles. For a given energy eigenvalue, as long as $\text{Im}R_{ll}(N, E + i\epsilon)$ does not vanish, Eq. (31) can be used to calculate the corresponding eigenvector.

The procedure of calculation of the eigenvector can now be set up as follows. For a system with N particles, a representative configuration is generated according to the law of distribution of disorder. Using Eqs. (2), (3), (4), and (31), the amplitudes of the eigenvector b_l can be calculated. There still remains the question of how to choose the parameter ϵ . The guideline can be constructed on the basis of the procedure used in the direct calculation of finite systems. Since the eigenvalues of a system depend on the degree of freedom, it is very difficult to follow the change of behavior of a particular eigenstate as $N \rightarrow \infty$. To circumvent this difficulty, for a system of N particles the properties of the eigenstates are averaged over all the eigenstates in the neighborhood of a chosen value of E . The interval ΔE is usually of the order $1/N$.¹⁴ In our treatment, the parameter ϵ can also be chosen as $1/N$. This is the equivalent to taking the average over the eigenvalue interval defined by $1/N$. The properties associated with the eigenstate can then be studied by examining how the amplitudes change as $N \rightarrow \infty$, $\epsilon \rightarrow 0$, but $N\epsilon = 1$. Work along this line is in progress.

Explicit eigenvector analysis of two- and three-dimensional systems may also be treated by this technique. This can be accomplished by first transforming a general matrix describing the two- or three-dimensional system into the tridiagonal form using the Lanczos recursion method¹⁶ pro-

posed by Haydock, Heine, and Kelly.¹⁷ The application of the recursion method to study the problem of localization has so far been concentrated on the general behavior of the localization,¹⁸⁻²¹ with practically no or little attention given to the nature of the localized states. The technique discussed here, however, may provide the scheme for a detailed amplitude analysis of the nature of the localized states.

VII. SUMMARY

In this work, a method is developed for the calculation of the eigenstate of an infinite tridiagonal matrix. Possible application of this method to study the problem of localization in a disordered linear chain is also discussed. There are two points which distinguish this method from the conventional numerical method. (1) There is no need to first numerically determine the eigenvalue (the eigenvalue spectrum is determined by the region where $\text{Im}R_{ll}$ does not vanish). Thus all the difficulties associated with the precise determination of the eigenvalue are circumvented. (2) There is no need to carry out the average over all the eigenstates in the eigenvalue interval defined by $1/N$, resulting in the saving of all the numerical calculation of those eigenstates.

ACKNOWLEDGMENTS

One of us (S. Y. Wu) would like to thank Professor H. S. Wu, Mr. Z. Y. Wong, and Mr. T. C. Mao of the Chinese University of Science and Technology for many stimulating discussions. He is also grateful to the Physics Department of the Chinese University of Science and Technology for the warm hospitality afforded to him during his visit to Hofei.

- ¹P. Dean, Rev. Mod. Phys. **44**, 127 (1972); see also, for example, L. Fox, *An Introduction to Numerical Linear Algebra* (Oxford U.P., New York, 1965).
- ²K. S. Dy, S. Y. Wu, and C. Wongtawatnugool, J. Phys. C **12**, L141 (1979).
- ³S. Y. Wu, C. C. Tung, and M. Schwartz, J. Math. Phys. **15**, 938 (1974).
- ⁴J. S. Langer, J. Math. Phys. **2**, 584 (1961).
- ⁵S. Y. Wu, Phys. Status Solidi (B) **74**, 349 (1976).
- ⁶R. J. Elliott, J. A. Krumhansl, and P. L. Leath, Rev. Mod. Phys. **46**, 465 (1974).
- ⁷S. Y. Wu, S. Bowen, and K. S. Dy, CRC Crit. Rev. Solid State Sci. (1980).
- ⁸N. F. Mott and W. D. Twose, Adv. Phys. **10**, 107 (1961).
- ⁹R. E. Borland, Proc. R. Soc. London, Ser. A **274**, 529 (1963).
- ¹⁰B. I. Halperin, Adv. Chem Phys. **13**, 127 (1967).
- ¹¹G. Theodorou and M. H. Cohen, Phys. Rev. B **13**, 4597 (1976).
- ¹²K. N. Economou and M. H. Cohen, Phys. Rev. B **4**, 396 (1971).
- ¹³J. C. Kimball, J. Phys. C **11**, 1367 (1978).
- ¹⁴C. C. Shih, to appear in J. Phys. C (1980).
- ¹⁵S. Y. Wu and Z. B. Zheng (to be published).
- ¹⁶See, for example, L. Fox, *An Introduction to Numerical Linear Algebra* (Oxford U.P., Oxford, 1965).
- ¹⁷R. Haydock, V. Heine, and M. J. Kelly, J. Phys. C **8**, 2591 (1975).
- ¹⁸S. Yoshimo and M. Okazaki, Solid State Commun. **20**, 81 (1976).
- ¹⁹J. Stein and U. Krey, Solid State Commun. **27**, 797 (1978).
- ²⁰R. Haydock, Philos. Mag. **B37**, 97 (1978).
- ²¹D. Mattis and R. Raghvan, Phys. Lett. A **75**, 313 (1980).

Young-tableau methods for Kronecker products of representations of the classical groups

Mark Fischler

Fermi National Accelerator Laboratory, Batavia, Illinois 60510

(Received 18 July 1980; accepted for publication 17 October 1980)

Diagrammatic methods for decomposing Kronecker products of arbitrary representations of any of the classical groups are presented. For convenience, efficient ways of computing the dimensions and quadratic Casimir's $C_2(R)$ are also given. These methods seem more useful for hand calculations than the method of Schur functions (or characteristic polynomials). An appendix presents the Kronecker products for any two representations of dimension ≤ 100 .

PACS numbers: 02.20.Qs

INTRODUCTION

The particle physicist looking at the theory of groups is generally interested in certain "practical" questions concerning the representations of the groups. Among these questions are

(a) What groups are available?

(b) What representations exist for a given group, and what is their nature?

(c) Branching rules: How representation R of group G breaks into representations S_i of subgroup H .

(d) Kronecker products: How $R_1 \otimes R_2$ breaks into irreducible representations $S_i \oplus S_2 \oplus \dots \oplus S_n$.

(e) "Clebsch-Gordan" coefficients for $R_1 \otimes R_2$: This, of course, needs the answer to (d) as a starting point.

There is a tendency to assume that mathematicians have addressed and solved these "practical" questions, yet it is not easy to find answers in the literature. The available groups (in the sense of having finite-dimensional representations) are well known: $SU(N)$, $SO(N)$, $Sp(N)$, the five exceptional groups, and products of these groups. Questions (b) and (c) are answered in table form in Patera and Sankoff¹; but these tables give no insight as to how the representations and branching rules are obtained. A partial table of Kronecker products exists,² but it suffers the same flaw, and also omits some important groups, for example, $SO(10)$, and lists no spinors at all. The problem of "Clebsch's" is a most difficult question in practice (although simple in theory once the Kronecker product is understood), and will not be addressed here.

Many physicists are familiar with Young-tableaux methods for finding the dimensions of representations and decomposing Kronecker products in $SU(N)$. This work generalizes these procedures to the groups $SO(2N+1)$, $Sp(2N)$, $SO(2N)$ and G_2 . The methods mathematicians describe use "characteristic functions"³ or Schur functions^{2,4} and are both nonintuitive and hard to learn to apply. The tableau method has the additional advantage that one can check whether $R_1 \otimes R_2$ contains a particular R_3 , without having to do the full product.

We have tried to make these rules as simple and "cook-booklike" as possible. Actually drawing out the diagrams is easier than working with lists of numbers, but these diagrams can't appear in the text, so they are represented by a

string of numbers in parentheses, with perhaps a symbol (\uparrow , \downarrow or $*$) in front, describing the number of boxes in each row. Representations can also be described by the Dynkin numbers, which we put in brackets []; this notation is standard and is how they appear in Ref. 1. The notation $(abc\dots n)$ matches that in Ref. 2 for nonspinors; we feel our notation for spinors is more convenient for a reason described below.

Our method of getting the dimension of an $SU(N)$ representation may differ from the "product of boxes over product of hooks" rule familiar to some physicists. It is, however, equally easy to apply, and falls into the same pattern as the other groups $SO(2N)$, $SO(2N+1)$, $Sp(2N)$, G_2 and F_4 . The six rules for Kronecker products may look imposing, but Rules 1-3 cover all but certain $SO(2N)$ cases, and in any event, these rules are easier to use than to concisely describe.

In the literature,¹ it is advised that the practical way of multiplying two representations is to multiply their dimensions, and look for a set of irreducible representations whose dimensions total that number, resolving ambiguities by using "Dynkin indices" (values of the quadratic Casimir operator). This method works for the few smallest representations, but for larger numbers it becomes fantastically cumbersome and ambiguous [e.g. in $SO(7)$ there is a 35-dimensional representation, a 27 representation, a 7 representation, and the trivial 1 representation, $35 = 27 + 7 + 1$. Their Dynkin indices are 20, 18, 2, and 0, respectively. Thus whenever a 35-dimensional representation appears in $R_1 \otimes R_2$, it could be replaced by $27 + 7 + 1$]. The rules set forth below are trivial to apply in these low-dimension cases, and are unambiguous in all cases. Dimensionally checking the result is, of course, still useful to prevent errors.

We append to this article a list of decompositions of all products where R_1 and R_2 are both ≤ 100 , and up to 210 for $SO(10)$, which is of special interest to grand unification theorists. We omit $SU(N)$, which is easy to decompose using rules 1 and 2. The $Sp(2N)$ products appear in Ref. 2 and are included here for completeness.

REPRESENTATIONS AND THEIR DIMENSIONS

A representation of a simple group of rank r can unambiguously be specified by a set of r integers corresponding to the r simple roots of the group. For example, in $SU(3)$, $[1,0]$ is the 3, $[0,1]$ the $\bar{3}$, and $[1,1]$ the 8; in $SO(10)$, $[10000]$ is the 10,

and $[00010]$, $[00001]$ are the $\overline{16}$ and 16 . This is how the representations are listed in Ref. 1, and we always will use square brackets and integers not separated by commas when referring to such a specification (we have had no occasion to look at any representation in which a number is more than 9 in this specification scheme).

It is well known to physicists, at least for $SU(N)$, that it is often more convenient to specify representations by "Young tableaux." In the case of $SU(N)$, the Young tableau corresponding to $[a_1 a_2 \dots a_{n-1}]$ consists of a_{n-1} columns of $n-1$ boxes, followed on the right by a_{n-2} columns of $n-2$ boxes...with lastly a_1 "columns" of one box each. Thus in $SU(6)$, for example, $[21031]$ is drawn as shown in Fig. 1. We will find it convenient to describe a tableau by listing in parentheses the number of boxes appearing in each row. $[21031]$ in $SU(6)$ is then written as (75441) . This notation should enable the reader to easily draw out any tableau in an example here.

The advantages of using such tableaux are threefold: In terms of the tableaux, one can compute the dimensionality of the representation, compute Kronecker products of two representations, and identify the symmetry properties of a representation (two boxes in a row mean two symmetric indices; boxes in a column imply antisymmetric indices). The justification for our particular way of defining tableaux for $SO(N)$ and $Sp(2N)$ is that we want to preserve the first two properties; the third can't be kept when spinor representations are involved.

For $Sp(2N)$ the tableau is the same as in the $SU(N)$ case. It will be seen, however, that where $[a, b, \dots, z]$ and $[z, \dots, b, a]$ are conjugate representations in $SU(N)$, they are not related in $Sp(2N)$.

$SO(2N+1)$ has the property of including spinors. The last number z in $[a, b, c, \dots, z]$ will determine if the representation is a spinor: if z is odd, it is a spinor. A pair of spinor indices can form vectorlike indices. Thus, if z is even, the tableau will contain $z/2$ columns of N boxes [as opposed to z such columns in, say, $Sp(2N)$]. If z is odd, the tableau will look the same: there are $(z-1)/2$ columns of N boxes, and to indicate a spinor is being described, an arrow is added to the notation. For example, as in Fig. 2, in $SO(7)$, $[123] = (\uparrow 431)$. $[002]$ would be (111) while $[003]$ is $(\uparrow 111)$ and $[001]$ is $(\uparrow 000)$.

$SO(2N)$ also has spinors; it has the added complexity of the last 2 roots referring to spinor indices. Let the representation be $[a, b, \dots, y, z]$, where $z \geq y$. Then if $y+z$ is odd, it is a spinor indicated by an upward pointing arrow. There are y columns of $N-1$ boxes, and $(z-y)/2$ [or $(z-y-1)/2$ in the case of a spinor] columns of N boxes. What is happening is that pairs of one of each type of spinor indices form vector indices of one kind, and then excess pairs of one type of spinor index form other vector indices. Thus, as in Fig. 3,

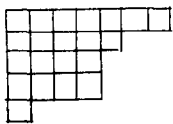


FIG. 1. $[21031]$ in $SU(6)$ can be written as (75441) .

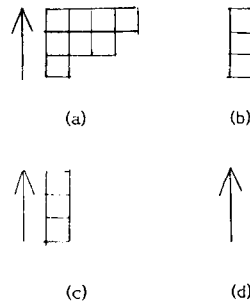


FIG. 2. (a) $[123] = (\uparrow 431)$ in $SO(7)$. (b) $[002] = (111)$. (c) $[003] = (\uparrow 111)$. (d) $[001] = (\uparrow 000)$.

$[00014]$ in $SO(10)$ becomes $(\uparrow 22221)$.

When $y > z$, the conjugate representation is formed: In $SO(10)$, 16 is $[00001]$ and $\overline{16}$ is $[00010]$. In this case, there are z columns of $N-1$ boxes, and $(y-z)/2$ (or $(y-z-1)/2$) columns of N boxes. When taking Kronecker products, it is important to know whether you are doing $R \times R$ or $R \times \overline{R}$, so we distinguish the $y > z$ representations by a down arrow if spinors $\{[00021] = (\downarrow 1110)\}$ or a star if the representation is a nonspinor $\{[10040] = (*3222)\}$.

To find the dimension of a representation $(a_1 a_2 \dots a_n)$ in a group one follows the following prescription: Add to the a_i [or twice a_i , if the group is $SO(N)$] some simple set of numbers (again dependent on the group), to get l_i . Form the product of some combination of the l_i , their differences $\Delta_{ij} = l_i - l_j$ ($i > j$) and their sum $\epsilon_{ij} = l_i + l_j$ ($i > j$), and divide by a specified denominator, which is the same as the numerator for all the $a_i = 0$. The specifics of this process are given in Table I. The process is illustrated for the representations $(\uparrow 1000)$ in $SO(8)$ and $SO(9)$ (Fig. 4); the numbers down the left side are common to any $SO(8)$ [$SO(9)$] representation, and the $+1$'s are because this representation is a spinor.

When the group in question is G_2 , two integers $[p, q]$ will label the representation, and the dimension can be computed from the Young diagram [which is $(p+q, q)$] by labelling the side with 1, 2 so that $l_i = 2 - i + a_i$, and forming the numerator $\prod l_i \prod \Delta_{ij} \prod \epsilon_{ij} \times (2l_1 + l_2)(2l_2 + l_1)$. The denominator is 120. Equivalently, the dimension is given by

$$(p+1)(q+1)(p+q+2)(p+2q+3)(p+3q+4) \times (2p+3q+5) \text{ divided by } 120.$$

For representations of F_4 , there are nonspinors $[a, b, 2c, d]$ and spinors $[a, b, 2c+1, d]$. The diagram for a nonspinor has the form $(d+a+2b+3c, a+b+c, b+c, c)$ and for a spinor $(\uparrow d+a+2b+3c+1, a+b+c, b+c, c)$. Equivalently, $(w, x, y, z) = [x-y, y-z, 2z+1]$ for spinor, $a-b-c-d(-1)$. Notice that the first row is always at least as long as the sum of the lengths of the other three rows. To find the dimension of a representation, write 11, 5, 3, 1

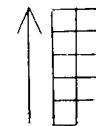


FIG. 3. $[00014] = (\uparrow 22221)$ in $SO(10)$.

TABLE I.

	Group	l_i	Numerator	Denominator
	SU(N)	$l_i = a_i + N - i$	$\prod_i \Pi \Delta_{ij}$	$1!2!\dots(N-1)!$
	Sp($2N$)	$l_i = a_i + N - i$	$\prod_i \Pi \Delta_{ij} \Pi \epsilon_{ij}$	$1!3!\dots(2N-1)!$
Nonspinor	SO($2N+1$)	$l_i = 2a_i + 2N + 1 - 2i$	$\prod_i \Pi \Delta_{ij} \Pi \epsilon_{ij}$	$2^{N(N-1)} 1!3!5!\dots(2N-1)!$
Spinor	SO($2N+1$)	$l_i = 2a_i + 2N + 2 - 2i$	same	same
Nonspinor	SO($2N$)	$l_i = 2a_i + 2N - 2i$	$\Pi \Delta_{ij} \Pi \epsilon_{ij}$	$2^{N(N-1)} (N-1)! 1!3!5!\dots(2N-3)!$
Spinor	SO($2N$)	$l_i = 2a_i + 2N + 1 - 2i$	same	same

down the left side and add two per box, plus 1 more if it a spinor. ($l_i = 2a_i + 11, 5, 3$, or 1 for $i = 1, 2, 3$ or 4, + 1 for a spinor or 0 for a nonspinor.) Then form the numerator:

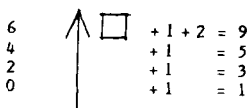
$$\prod_i \Pi \Delta_{ij} \Pi \epsilon_{ij} (l_1 + l_2 + l_3 + l_4)(l_1 + l_2 + l_3 - l_4) \times (l_1 + l_2 - l_3 + l_4) \times (l_1 + l_2 - l_3 - l_4)(l_1 - l_2 + l_3 + l_4) \times (l_1 - l_2 + l_3 - l_4) \times (l_1 - l_2 - l_3 + l_4)(l_1 - l_2 - l_3 - l_4).$$

Note that because $a_1 \geq a_2 + a_3 + a_4$, all of those are positive. The denominator is, as usual the numerator with $a_1 = a_2 = a_3 = a_4 = 0$, which works out to be $11!9!!25 \times 2^{25}$. Of course, $\dim [pqrs]$ can be written as a polynomial (of degree 24) in p, q, r and s , but this is not very illuminating or convenient.

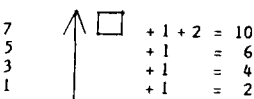
KRONECKER PRODUCTS

When taking the Kronecker product of two representations, arrange the less complicated representation on the right, and label it with an a in each box in the first row, b 's in the second row, etc. Then follow the rules set down below. Note that one can check the result by seeing whether the sum of the dimensions of the results is equal to the product of the dimensions of the representations being multiplied. Also, it is sometimes easier to use the following trick than to multiply out explicitly: Say you need $R_1 \otimes R_2$, and you know that $S_1 \otimes S_2 = R_2 + T_1 + T_2 + \dots$, where the S_i and T_i are all much simpler than R_2 . Then one may write $R_2 = S_1 \otimes S_2 - T_1 - T_2 \dots$ and do $(R_1 \otimes S_1) \otimes S_2$, subtracting the results of $R_1 \otimes T_1 + R_2 \otimes T_2 + \dots$. This trick will be illustrated below.

Two techniques were utilized in deriving these rules.



(a)



(b)

FIG. 4. (a) $(\uparrow 1000)$ in SO(8) has dimension $56 = \frac{14 \cdot 12 \cdot 10 \cdot 8 \cdot 6 \cdot 4 \cdot 4 \cdot 6 \cdot 8 \cdot 2 \cdot 4 \cdot 2}{2^{12} 3! 1! 3! 5!}$.

(b) $(\uparrow 11000)$ in SO(9) has dimension $128 = \frac{10 \cdot 6 \cdot 4 \cdot 2 \cdot 16 \cdot 14 \cdot 12 \cdot 10 \cdot 8 \cdot 6 \cdot 4 \cdot 6 \cdot 8 \cdot 2 \cdot 4 \cdot 2}{2^{12} 1! 3! 5! 7!}$.

Careful manipulations of tensors and group invariants can indicate the procedure when there are no spinor indices (or implied spinor indices). When spinors are present, it is possible to use the trick described above to determine what the product is; one can then carefully note for the general cases which representations will remain after subtracting $R_1 \otimes T_1 + R_2 \otimes T_2 + \dots$. This procedure could in principle have become prohibitively cumbersome, but any combination rules simple enough to be practical to apply are also relatively easy to derive in this way.

Rule 1: Adding one box: One tacks a single box onto the end of any row (including a row of 0 length) in all ways so as to leave a correct tableau for the particular group (no row longer than the one above, and the number of rows not exceeding the rank of the group). For example (see Fig. 5) in SU(4), $(110) \otimes (100)$ contains (210) and (111) . For notational convenience, we will write the operation of appending an "a" box in the n th position of the k th row as $\{a \rightarrow n, k\}$. Thus in this example, we have $\{a \rightarrow 2, 1\}$ and $\{a \rightarrow 1, 3\}$.

Rule 1a: In Su(N) you can add the box to the n th row (the rank is $n - 1$) and cancel that whole column. This corresponds to contracting n indices via an epsilon symbol. Thus in SU(3), $(1,1) \otimes (1,0)$ contains $(0,0)$ via $\{a \rightarrow 1, 3$ (elim. col. 1)}.

Rule 1b: In SO($2N$), SO($2N + 1$) or Sp($2N$), one may also use the added box to cancel a box in the existing tableau. For example, in SO(10), $(11000) \otimes (10000)$ contains, via $\{a \rightarrow 1, 2$ cancel}, (10000) . This corresponds to contraction with δ_{ab} , an invariant in SO(N), or with Ω_{ab} , the antisymmetric invariant in Sp($2N$).

Rule 1c: In SO($2N$) or SO($2N + 1$), if R_1 is a spinor and does not contain any N -box column, you may also use the added box to simply flip the direction of the spinor-indicating arrow. In SO($2N + 1$) this means simply "absorbing" the box in the spinor arrow. For example, in SO(8), $\uparrow(2000) \otimes \uparrow(1000)$ contains, via $\{a \rightarrow \uparrow\}$, $(\uparrow 2000)$. In SO(7), $(\uparrow 200) \otimes (100)$ contains $(\uparrow 200)$.

Rule 1d: Only in SO($2N + 1$), one may also "merge" the added box with the last box in the N th row. Thus in SO(7), $(222) \otimes (100)$ contains, via $\{a \rightarrow 2, 3$ merge}, (222) .

Rule 1e: In SO($2N$), when adding a box at the N th row in the 1st column $\{a \rightarrow 1, N\}$, both $(abc\dots 1)$ and $(*abc\dots 1)$ appear in the result. For example, in SO(10), (11110) is $[00011]$ and

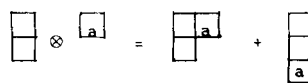


FIG. 5. In SU(4), $(110) \otimes (100) = (210) + (111)$.

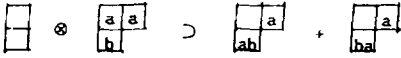


FIG. 6. In $Sp(6)$, $(110) \otimes (210)$ contains two (210) 's.

(10000) is [10000]. In their product, since they are both self-conjugate (under exchange of the last two Dynkin numbers) you would get both [00002] and [00020], that is, both (11111) and (*11111).

Rule 2: Adding more than one box. Label the boxes in the top row "a," the next row "b," etc. Add each box one by one, always in a one-box-permissible way, the top row first, then the 2nd row, etc., and such that reading from right to left and then up to down, the number of *a* boxes encountered is always \geq the number of *b*'s, \geq number of *c*'s, and so on. Two representations are distinct if the *a, b, c, ...* labelling differs. For instance, in $SU(4)$, $(210) \otimes (210)$ contains both a (321) from $\{a \rightarrow 3, 1; a \rightarrow 2, 2, b \rightarrow 1, 3\}$ and a (321) from $\{a \rightarrow 3, 1; a \rightarrow 1, 3; b \rightarrow 2, 2\}$. Also, no two *a*'s (or *b*'s or *c*'s...) may appear in different rows of the same column. Such a representation would be both symmetric and antisymmetric in those two indices. Rules 1, 1a and 2 fully cover the case of $SU(N)$.

Rule 2a: In $Sp(2N)$, $SO(2N + 1)$ or $SO(2N)$, you may use a box from R_2 to replace a box in R , that was previously cancelled. Thus $(200) \times (110)$ contains (200) via $\{a \rightarrow 2, 1$ cancels; $b \rightarrow 2, 1\}$. For the purposes of Rule 2, these would count as an "a" and a "b" simultaneously.

Rule 2b: Rarely, when applying Rule 2a using boxes of two different rows, it will be found that Rule 2 is satisfied (the right to left and up to down part) whether the labelling of the readded box is *ab* or *ba*. For example, in $Sp(6)$, $(110) \otimes (210)$ contains, via $\{a \rightarrow 2, 1; a \rightarrow 1, 2$ cancel; $b \rightarrow 1, 2\}$ the representation (210). As can be seen from Fig. 6, Rule 2b applies here. In this case, two (210)'s appear in the result.

Rule 2c: A box may never cancel a previously added box. This operation would correspond to taking a trace (or symplectic trace by contracting with Ω_{ab}) over two indices which both appear in R_2 , but R_2 is irreducible, so the operation gives zero.

Rule 2d: In $Sp(2N)$, two boxes from different rows may not cancel and readd a box in the *N*th row. Thus in $Sp(8)$, $(1111) \otimes (1100)$ does not contain (1111).

Rule 2e: Up to one "a," one "b" ... may be absorbed by a spinor line on $SO(2N)$ or $SO(2N + 1)$.

Rule 2f: When cancelling boxes, you may anticipate future cancellations. An example of this should explain: In $SO(9)$, $(2110) \otimes (1100)$ contains (2000) via $\{a \rightarrow 1, 2$ cancels; $b \rightarrow 1, 3$ cancels} (see Fig. 7). When a box is subsequently readded, if an *N*th row box was cancelled in $Sp(2N)$ Rule 2d applies and the tableau should not appear in the result. In Fig. 8, $(1111) \otimes (1110)$ is done in $Sp(8)$. Note that (1110) does not appear in the answer.

Rules 1–2f fully cover $Sp(2N)$. When spinors or repre-

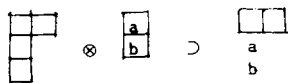


FIG. 7. $(2110) \otimes (1100)$ in $SO(9)$ contains (2000).

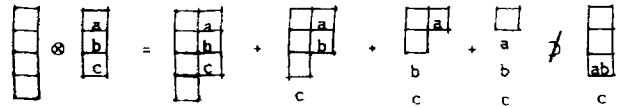


FIG. 8. In $Sp(8)$, $(1111) \times (1110)$ does not contain (1110).

sentations with $N - 1$ or more rows in $SO(2N)$ or $SO(2N + 1)$ appear, one must also apply the following rules:

Rule 3: [Rules 3–3d apply to $SO(2N + 1)$] No two boxes from the same row may merge together. For example, in $SO(7)$, $(100) \otimes (111)$ contains (111) via $\{a \rightarrow 2, 1; b \rightarrow 3, 1; c \rightarrow 3, 1$ merge}. But $(110) \otimes (200)$ does not contain (111) via $\{a \rightarrow 3, 1; a \rightarrow 3, 1$ merge}.

Rule 3a: "Merging" boxes, in actuality, is adding the boxes in the $N + 1$ st row, and using the $2N + 1$ -index epsilon symbol $SO(2N + 1)$ invariant to reduce the column of X boxes to $2N + 1 - X$ boxes. It is necessary to use this more cumbersome point of view when R_2 contains a column of N boxes. For example, in $SO(7)$, $(111) \otimes (111)$ contains (100) via $\{a \rightarrow 4, 1; b \rightarrow 5, 1; c \rightarrow 6, 1$ contract $\epsilon\}$, which would not be obtained by any combination of merging and cancelling.

Rule 3b: When R_1 is a spinor, boxes absorbed by the spinor line as per Rule 1c count as being put in the $N + 1$ st row for the purposes of the right to left and up to down part of Rule 2. This also applies to $SO(2N)$.

Rule 3c: When R_2 (but not R_1) is a spinor, the resulting representations are all spinors and are formed by multiplying as if R_2 was a nonspinor, adding the spinor line, and for each result, removing zero or one box per row. Figure 9 illustrates this for $SO(7)$ $(110) \times (\uparrow 100)$: The (210) in $(110) \times (100)$, for instance leads to (1210), ($\uparrow 110$), (1200) and ($\uparrow 100$). When applying this rule, sometimes one gets repeated diagrams. In this case, discard one of each group of identical diagrams. In the example shown, of the 3 ($\uparrow 100$)'s, only two are kept. This rule will also apply to $SO(2N)$ when R_1 has less than N rows.

Rule 3d: When both R_1 and R_2 are spinors, the answers will, of course, be nonspinors. Multiply the nonspinor parts,

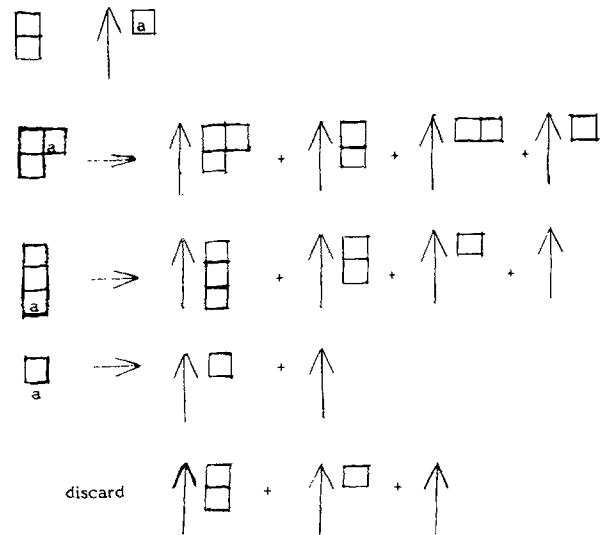


FIG. 9. $(110) \times (\uparrow 100)$ in $SO(7) = (\uparrow 210) + (\uparrow 110) + (\uparrow 200) + (\uparrow 100) + (\uparrow 111) + (\uparrow 1000) + (\uparrow 100)$.

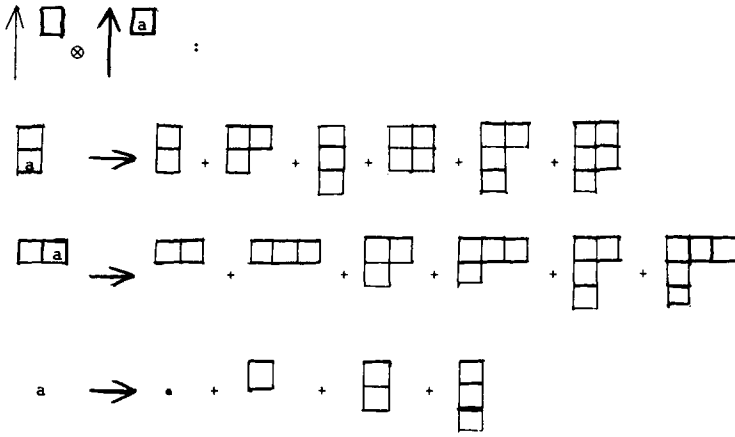


FIG. 10. $(\uparrow 100) \otimes (\uparrow 100)$ in $SO(7) = (110) + (210) + (111) + (220) + (211) + (221) + (200) + (300) + (210) + (310) + (211) + (311) + (000) + (100) + (110) + (111)$.

and then for each result, add zero or one box per row in all possible ways. This is illustrated in Fig. 10 which does $(\uparrow 100) \times \uparrow(100)$ in $SO(7)$. Note that the repeated diagrams are all counted. Applying Rule 3d, one can immediately see that $(\uparrow 0000\dots) \times (\uparrow 0000\dots) = (0000\dots) + (1000\dots) + (1100\dots) + \dots + (11\dots 11)$.

Rules 1–3d fully cover $SO(2N + 1)$. Rules 4 and 5 apply to $SO(2N)$.

Rule 4: When at least one of R_1 and R_2 is a nonspinor and contains no column of N rows, use Rules 4a–4f.

Rule 4a: When a box is being added to the N th row, 1st column it can stand for either $[\dots, 0, 2]$ or $[\dots, 2, 0]$. Thus the representation is counted twice, $(a, b, c\dots)$ and $(*a, b, c\dots)$. This is a generalization of Rule 1e.

Rule 4b: When R_1 is a spinor, or already has N -box columns, the doubling in Rule 4a does not apply. For example, in $SO(10)$,

$$(\uparrow 11110) \otimes (10000) = (\uparrow 11100) + (\uparrow 21110) + (11110) + (\uparrow 11111);$$

There is *not* also $(\downarrow 11111)$.

Rule 4c: When cancelling and readding in the $(N, 1)$ position, this doubling does *not* apply. Thus, in $SO(10)$, $(11111) \otimes (11000)$ contains (11111) but not $(*11111)$.

Rule 4d: The spinor line in R_1 changes direction once for each box “absorbed” in it.

Rule 4e: When R_1 and R_2 have long enough columns, one may also form columns of $M > N$ boxes and use the $2N - M$ index epsilon symbol to create $2N - M$ box columns. This result is not distinct from what would be gotten by cancellations only, without readds. Figure 11 shows how this works: In $SO(8)$, $(1110) \otimes (1110)$ contains (1100) via $\{a \rightarrow 2, 1 \text{ cancels}; b \rightarrow 2, 1; c \rightarrow 3, 1 \text{ cancels}\}$ and another (1100) via $\{a \rightarrow 4, 1; b \rightarrow 5, 1; c \rightarrow 6, 1 \text{ epsilon}\}$. Yet in $(1111) \otimes (1110)$, the (1000) from $\{a \rightarrow 2, 1 \text{ cancels}; b \rightarrow 3, 1 \text{ cancels}; c \rightarrow 4, 1 \text{ cancels}\}$ is the only one counted; $\{a \rightarrow 5, 1; b \rightarrow 6, 1; c \rightarrow 7, 1 \text{ epsilon}\}$ is the same (1000) .

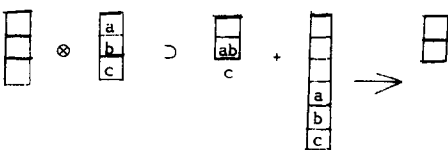


FIG. 11. In $SO(8)$ $(1110) \otimes (1110)$ contains two (1100) 's.

Rule 4f: When R_2 (but not R_1) is a spinor, and R_1 has no N -box columns, do the multiplication just as in $SO(2N + 1)$, using Rule 3c. When an odd number of boxes was eliminated, the resulting spinor arrow is in the opposite direction from that of R_2 if R_2 had no N -box columns.

This covers $SO(2N)$ except when a circumstance peculiar to $SO(2N)$ occurs: If both R_1 and R_2 are not self-conjugate (they are spinors or have N rows), then it makes a difference whether you multiply R_1 by R_2 or by its conjugate. For instance, in $SO(10)$, it is well known that $16 \otimes 16 = 1 + 45 + 210$, while $16 \otimes \overline{16} = 10 + 120 + 126$. This phenomenon, covered in Rule 5, is distinct from that of rule 4, although the underlying reason for both is the two distinct (yet isomorphic) types of spinors available.

Rule 5: $(\uparrow 000\dots 00) \otimes (\uparrow 000\dots 00) = (11\dots 111111) + (11\dots 111100) + (11\dots 110000) + \dots$, while $(\uparrow 0000\dots 00) \otimes (\downarrow 000\dots 00) = (11\dots 11110) + (11\dots 11000)$

$+ (11\dots 100000) + \dots$. This is illustrated for $SO(10)$: $16 \otimes 16 = 1 + 45 + 210$, while $16 \otimes \overline{16} = 10 + 120 + 126$. It is interesting to note that in $SO(4N)$, a spinor \otimes itself contains the 1 representation, and a spinor \otimes its conjugate contains the “vector” $4N$, while in $SO(4N + 2)$, spinor \otimes spinor contains 1, while spinor \otimes spinor contains the vector. This pattern is easy to verify for $SO(4)$ [isomorphic to $SU(2) \times SU(2)$: $(\uparrow 00) \rightarrow \{\frac{1}{2}, 0\}$ and $(\downarrow 00) \rightarrow \{0, \frac{1}{2}\}$ so $\uparrow \times \uparrow = \{1, 0\} + \{0, 0\} = (11) + (00)$ while $\uparrow \times \downarrow = (10)$] and for $SO(6)$ [isomorphic to $SU(4)$: $\uparrow \rightarrow 4, \downarrow \rightarrow \overline{4}$, so $\uparrow \times \uparrow$ contains a 6, the vector in $SO(6)$, while $\uparrow \times \downarrow$ has a 1]. It is also obvious in $SO(8)$, the first nontrivial $SO(2N)$, because of the symmetry of $SO(8)$ which says $[abcd]$ and $[cbad]$ (and

$[abdc] = [\overline{abcd}]$, of course) are isomorphic. Thus $[1000]$ looks like $[0001]$: the vector and spinor $8_v, 8_s$ are alike as 8_s and $\overline{8}_s$. So, when taking $8_s \times 8_s$, this can't contain 8_v , because if it did, the symmetry would tell you that $8_v \times 8_v$ contains $\overline{8}_s$, a contradiction since two nonspinors can't produce a spinor. This symmetry property is useful for

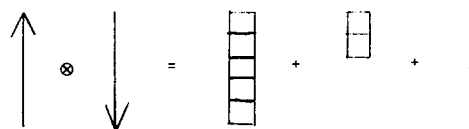


FIG. 12. $16 \times \overline{16}$ in $SO(8)$.

doing SO(8) Kronecker products. For instance, $(1111) \times (*1111)$ is $[0002] \times [0020]$ which is related to $[0002] \times [2000]$ or $(1111) \times (2000)$, an easier product to take. For this reason, SO(8) is also a good "laboratory" for seeing how complicated SO(2N) representations multiply; for the remainder of Rule 5, SO(6) and SO(4), where the results are easy to derive in a different way, can also be used to illustrate the various rules.

Rule 5a: When R_2 is an elementary spinor, $[000\dots 01] = (\uparrow 000\dots)$ or $[00\dots 10] = (\downarrow 00\dots)$ and R_1 is a non-spinor with N rows (if R_1 has fewer rows, see Rule 3c). For the sake of illustration, we will assume R_1 is of the type $[ab\dots yz]$, $z > y$. If $y > z$, then one can still use these rules by multiplying the conjugates of R_1 and R_2 , and conjugating the answer: e.g., $[0002] \otimes [0001] = [0003] + [0021] + [0101] + [1010] + [0001]$ in SO(8), so $[0020] \otimes [0010] = [0030] + [0012] + [0110] + [1001] + [0010]$. The rule is to eliminate zero or one box per row in R_1 in all possible ways such that an even (odd) number of boxes are eliminated if R_2 is $\uparrow(\downarrow)$. Assign a \uparrow spinor orientation to each of the results. (The z in R_1 , which must be at least $y + 2$ since R_1 is a non-spinor, dominates even if R_2 is \downarrow .)

Rule 5b: When R_1 is a spinor and R_2 is an elementary spinor, add zero or one box per row in R_1 in all possible ways such that the total number of boxes added is of the same parity as $N(N + 1)$ if the two spinor arrows are in the same (opposite) direction. The results are non-spinors, and if $R_1 = (\downarrow abc\dots)$ they all have $y \geq z$; if $R_1 = (\uparrow ab\dots)$ they all have $z \geq y$. This rule is shown in Fig. 13, wherein $560 \otimes 16$ is done for SO(10). The 560 is $[01001]$ or $(\uparrow 11000)$; the 16 is $(\downarrow 00000)$ or $[00010]$. The result is $(11000) + (11110) + (21100) + (22000)$

$+ (22200) + (22110) + (21111)$. 560×16 would contain $(*21111) = [10020]$ rather than (21111) .

Rules 5c-5g will cover the cases where R_2 is a nonelementary spinor and R_1 contains N rows. It will be more practical, however, to treat these cases by writing R_2 as $\uparrow \times S$ -various smaller spinors. For example, in SO(8), $(2111) \times (\uparrow 1100) = (2111) \times \{(\uparrow 0000) \times (1100) - (\downarrow 1000) - (\uparrow 0000)\} = (2111) \times \{(1100) \times (\uparrow 0000) - (1000) \times (\downarrow 0000) + (\uparrow 0000) - (\uparrow 0000)\}$. Each of the two resulting triple products is easy to do. Figure 14 shows $(1111) \otimes (\uparrow 1000)$ in SO(8) graphically, to illustrate how simple the tableau method makes things.

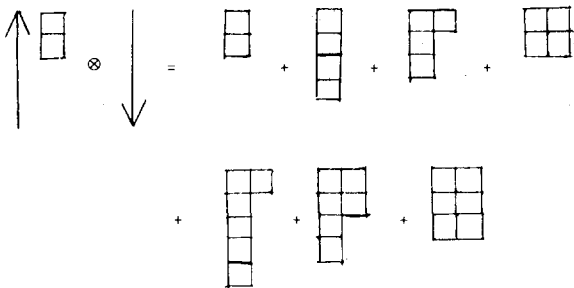


FIG. 13. $(\uparrow 11000) \otimes (\downarrow 00000)$ in SO(10) = $(11000) + (11110) + (21100) + (22000) + (22200) + (22110) + (21111)$.

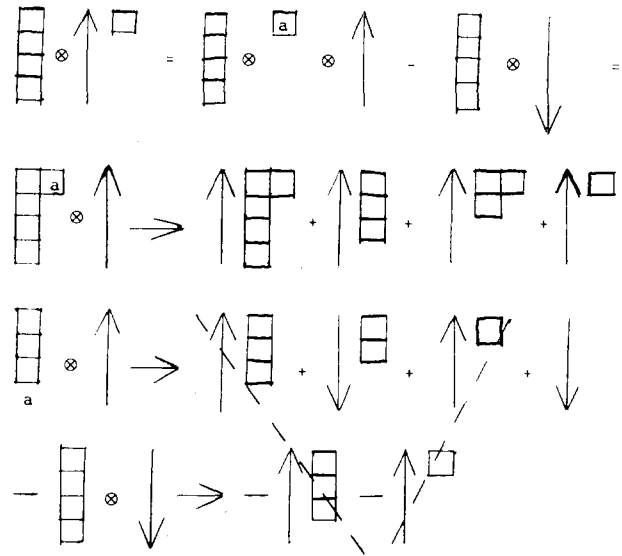


FIG. 14. In SO(8), $(1111) \otimes (\uparrow 1000) = (1111) \otimes \{(\uparrow 1000) \otimes (\uparrow 0000) - (\downarrow 0000)\} = (\uparrow 2111) + (\uparrow 1110) + (\uparrow 2100) + (\uparrow 1000) + (\uparrow 1100) + (\uparrow 0000)$.

Rule 5c: When R_1 is a non-spinor of N rows and R_2 is a spinor containing less than N rows: Combine the non-spinor parts normally, and eliminate one box each from an even number of rows (if R_2 is \uparrow ; an odd number if \downarrow). All the resulting representations are \uparrow , except if in combining the non-spinor part, you get a representation with less than N rows (the bottom boxes were all killed). In that case, for that representation, the results are all \downarrow , and the number of boxes eliminated is of the other parity (odd if R_2 is \uparrow). This process is illustrated in Fig. 15, in which in SO(8), $(1111) \times (\uparrow 1000)$ is done directly. $(1111) \times (\uparrow 1000) = (2111) + (1110)$, $(2111) \rightarrow (\uparrow 2111) + (\uparrow 2100) + (\uparrow 1110) + (\uparrow 1000)$, $(1110) \rightarrow (\downarrow 1100) + (\downarrow 0000)$. After doing this, certain representations may have to be discarded, as outlined in 5d and 5e.

Rule 5d: When a representation can be arrived at in two or more ways, by elimination of boxes in two or more different representations, discard one of those representations. For example, in SO(8) $(4321) \times (\uparrow 1000)$, $(5321) \rightarrow (\uparrow 4311) + \text{others}$, $(4421) \rightarrow (\uparrow 4311) + \text{others}$, but only one of these $(\uparrow 4311)$'s appears in the result.

Rule 5e: When all the eliminated boxes are the boxes coming from R_2 , eliminate one of that kind of representation, even if there appear no others. For example, in SO(8),

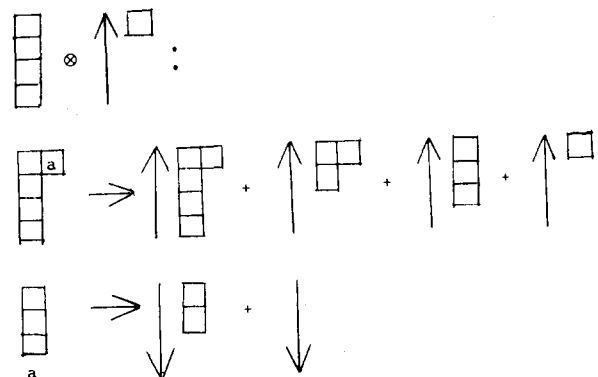


FIG. 15. Doing $(1111) \otimes (\uparrow 1000)$ using Rule 5c.

(1111) × (↑1000): (1111) × (1000) contains, via $\{a \rightarrow 1, 2\}$, (2111). (2111) → (↑1111) by eliminating the box at the end of row 1, which came from R_2 . Thus (1111) × (↑1000) does not contain this (↑1111). On the other hand (2111) × (↑1000) can form (↑2111) in two such ways: (3111) → (↑2111) and (2211) → (↑2111), so one of these is discarded, and one appears in the answer.

Rule 5f: When R_2 is a spinor with N rows and R_1 is a spinor with less than N rows: Multiply the nonspinor parts. To each result add an odd or even number of boxes in all possible ways, adding up to one per row. The number of boxes added is of the same parity as N if R_1 and R_2 are both ↑ (or both ↓) and the opposite parity otherwise. [Here is another case in which $SO(4N + 2)$ differs from $SO(4N)$.] Then eliminate one of each set of duplicated representations as in 5d, and one of each type of representation formed by adding, in the 2nd step, boxes that were eliminated in the 1st (as in Rule 5e). Also eliminate one of each type of representation wherein just one entire column was added in the two steps.

Rule 5g: Multiplying two N -row representing is best done via the procedure in Rule 5c. However, when both R_1 and R_2 are of the form $[00 \dots 02m]$ or $[00 \dots 2m]$, a simple pattern emerges: start with the representation formed by adding $[00 \dots 2m]$ (R_1) to $[00 \dots 2m]$ or $[00 \dots 2m, 0]$ (R_2), to get $[00 \dots 2m, 2m]$ or $[00 \dots 4m]$ and eliminate pairs of vertically touching boxes. This is illustrated in Fig. 16 for $[00002] \times [00002]$ in $SO(10)$:

$(11111) \times (11111) = (22222) + (22211) + (22200) + (21111) + (21100) + (20000)$. This process is applicable to any R_1 and R_2 of the form $[000 \dots K]$ or $[00 \dots K 0]$, with the 1st M columns ($M = |K/2(R_1) - K/2(R_2)|$) untouched by the elimination process: $[00004] \times [00020]$ in $SO(10)$ is shown in Fig. 17: $(22222) \times (*11111) = (33331) + (33221) + (22221) + (33111) + (22111) + (11111)$.

Rule 6: Products of representations of G_2 : the simplest diagrammatic means of multiplying two representations of G_2 relies on the fact that $SU(3)$ is a subgroup of G_2 . The procedure entails four steps: (1) Break R_1 and R_2 down into their $SU(3)$ content $(S_{11} + S_{12} + S_{13} \dots) \times (S_{21} + S_{22} + S_{23} \dots)$. (2) Multiply these $SU(3)$ representations to get $(S_{11} \times S_{21}) + (S_{11} \times S_{22}) + \dots = T_1 + T_2 + \dots$. (3) Choose a particular T_i to be part of the content of some G_L representations. (4) From the set $\{T_i\}$, eliminate the content of that representation. Repeat steps 3 and 4 until no representations are left over. Steps 1–4 are each explained in some detail in Rules 6a–6d.

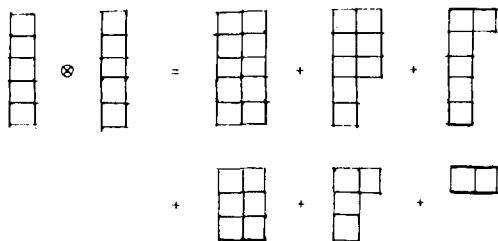


FIG. 16. $[00002] \otimes [00002]$ (126×126) in $SO(10) = (22222) + (22211) + (22200) + (21111) + (21100) + (20000)$: $126 \times 126 = 2772 + 6930 + 4125 + 1050 + 945 + 54$.

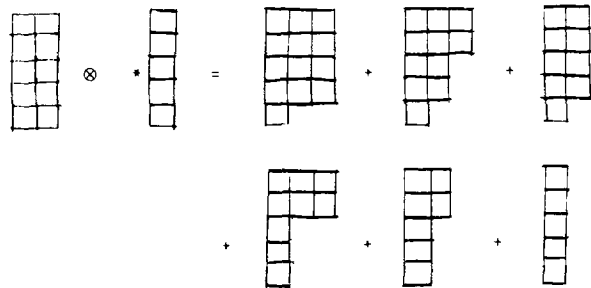


FIG. 17. $(22222) \times (*11111) = (33331) + (33221) + (22221) + (33111) + (22111) + (11111)$. Note that the first column remains untouched.

Rule 6a: To get the $SU(3)$ content of a G_2 representation: write the symbols p , q and o , one in each box in the Young diagram, in all ways such that within a row, the p 's precede the q 's which precede the o 's and within a column, p is above q , which is above o . Both boxes in one column may not contain the same symbol. Each of these arrangements becomes one $SU(3)$ representation, $[p, q]$ with $p =$ the number of p 's appearing in the labelled tableau, and $q =$ the number of q 's. This process is illustrated in Fig. 18 for $7 = [10] \rightarrow 3 + \bar{3} + 1$ and in Fig. 19 for $[11]$ (64 dimensions in G_2) $\rightarrow 15 + 6 + 15 + 8 + 8 + 3 + \bar{6} + \bar{3}$.

Rule 6b: Of course, $SU(3)$ Kronecker products are easy to do. But a further factor of two in time spent can be saved if one makes use of the fact that the G_2 representation always contains both R and \bar{R} of $SU(3)$.

Rule 6c: The G_2 representation to eliminate first is found by picking the remaining $SU(3)$ representation with the longest first row (when two have equal first rows, the longest second row). This is the T_i of step 3. Call this representation $(a + b, b)$ or $[a, b]$ of $SU(3)$ (b will always be greater than a); then the G_2 representation to eliminate this is (b, a) . The reason this representation is chosen is that the minimal G_2 representation that contains $(a + b, b)$ is (b, a) . (b, a) decomposes into $SU(3)$ representations $(p + q, q)$ with $p + q \leq a + b$ and if $p + q = a + b$, $q \leq b$. Thus by eliminating any of the other representations, one would never get a G_2 representation that includes T_i . So, eventually, one will have to use (b, a) in G_2 . This representation, however, will decompose into others of the remaining representations, so one should eliminate those first. For example, when doing 14×7 in G_2 , one has the $SU(3)$ representation (32) , 3 (21's), as well as others. If you tried to cover the (21)'s first by using 3 (11)'s of G_2 $\{(11)$ in G_2 contains (21) in $SU(3)\}$, then later, when taking care of the (32), you would find you still have to eliminate two more

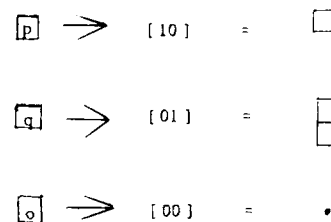


FIG. 18. $[10] = (10)$ in $G_2 = [10] = (10) + [01] = (11) + [00] = (00)$ in $SU(3)$.

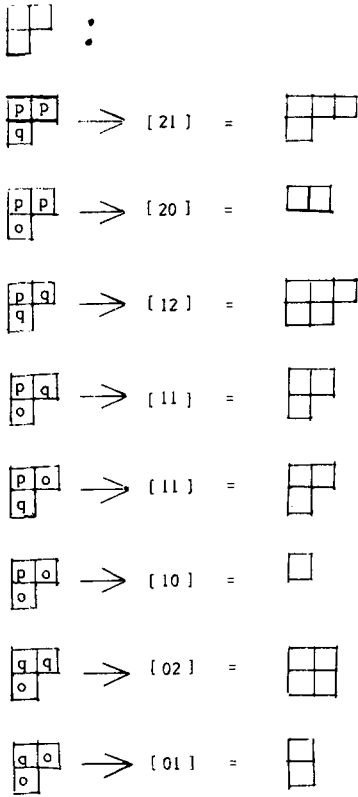


FIG. 19. $[11] = [21]$ in G_2 contains $(31) + (20) + (32) + (21) + (21) + (10) + (22) + (11)$ of $SU(3)$.

(21) 's { which are no longer available }. But if you start with the (32) , the dilemma can't occur.

To illustrate this procedure, $[01] \times [20]$ (or 14×27) is done below: Step 1: $G_2[01] \rightarrow [11] + [10] + [01]$, $G_2[20] \rightarrow [20] + [11] + [10] + [02] + [01] + [00]$. Step 2: the product of the 1st set with the 2nd set is $\{([31] + [12] + [20] + [01]) + ([30] + [21]) + ([21] + [10])\} + \{([22] + [30] + [03] + [11] + [00]) + ([21] + [02] + [10]) + ([12] + [20] + [01])\} + \{([21] + [02] + [10]) + ([01] + [20]) + ([11] + [00])\} + \{([13] + [21] + [02] + [10]) + ([12] + [01]) + ([03] + [12])\} + \{([13] + [20] + [01]) + ([10] + [02]) + ([11] + [00])\} + \{[11] + [10] + [01]\}$.

Step 3: The 1st representation to eliminate is $[13]$ because 4 is the biggest sum, and among the representations with a sum of 4, $[13]$ has the biggest 2nd number, i.e. $[13] = (43)$ while $[22] = (42)$. $[13] \rightarrow (31)G_2 = [21]G_2$.

Step 4: $[21]G_2 \rightarrow [31] + [30] + [22] + [21] + [21] + [20] + [13] + [12] + [12] + [11] + [11] + [10] + [03] + [02] + [01]$. Of course, the $[13]$ (and $[31]$) are what we picked $[21]G_2$ for. Cancelling these 15 representations, we return to step 3: the $[31]$, $[13]$ and, by accident, the only $[22]$ have all been eliminated; next is $[03] = (33)$, since one of them is left.

$[03] \rightarrow (30)G_2 = [30]G_2 \rightarrow [30] + [21] + [20] + [12] + [11] + [10] + [03] + [02] + [01] + [00]$. Eliminate these 10, and repeat step 3. Completing the process, we find that, in G_2 , $[20] \times [01] = [21] + [30] + [11] + [20] + [01] + [10]$, or $27 \times 14 = 189 + 77 + 64 + 27 + 14 + 7$ in terms of dimensions.

It is not as easy to formulate rules for the other exceptional groups. (F_4 representations may be decomposed into $SO(9)$ representations² and multiplied as in Rule 6, but the simple decomposition rule we have put forth for G_2 is absent in this case.)

AN APPLICATION IN GRAND UNIFIED THEORIES

It is currently popular to postulate that the grand unified group is some large group [usually $SU(N)$ or $SO(2N)$ with $N \geq 5$]. The symmetry is then broken in some series of steps down to $SU(3) \times SU(2) \times U(1)$. It has been proposed⁴ that instead of elementary Higgs, the scalars should be composites made of two fermions. Dimopoulos and Susskind⁵ give a "rule" for determining which fermions might condense out: Say the fermions are in representations R_1, R_2, \dots, R_n . The condensate will form in the "most attractive channel" (this assumes that one gluon exchange is the important process). To determine the relative attractiveness of channels

$R_i + R_j \rightarrow S$ (where $S \in (R_i \otimes R_j)$, one compares $C_2(S) - C_2(R_i) - C_2(R_j)$). Thus to apply the maximally attractive channel (MAC) prescription, one needs to know $R_i \otimes R_j$ and the quadratic Casimir's for the relevant representations. The rules for getting $R_i \otimes R_j$ are given above. Since these large groups don't appear in Ref. 1, it would be useful to find a way to compute $C_2(R)$. Fortunately, $C_2(R)$ is, for the classical groups, a polynomial of degree 2 in the indices $[a_n]$ which describe the representation. For $SU(N+1)$, with $R = [a_1 a_2 a_3 \dots a_N]$,

$$C_2(R) = \left(\sum_{k=1}^N k(N-k+1)(a_k^2 + (N+1)a_k) + 2 \sum_{i=2}^N \sum_{j<i} j(N-i+1)a_i a_j \right) 2(N+1).$$

When R is the defining $(N+1)$ -dimensional representation, $C_2(R)$ is normalized to $N(N+2)/2(N+1)$. Thus for $SU(3)$, $C_2(R) = \frac{4}{12}(a_1^2 + 3a_1 + a_2^2 + 3a_2 + a_1 a_2)$, contrary to the formula given in Ref. 1. For $SO(2N)$, with $R = [a_1 a_2 \dots a_{N-2} y z]$,

$$C_2(R) = \left(\sum_{k=1}^{N-2} \{ka_k^2 + [2kN - k(k+1)]a_k\} + \frac{1}{4}[Ny^2 + (2N-2)Ny + Nz^2 + (2N-2)Nz] + 2 \sum_{i=2}^{N-2} \sum_{j<i} ja_i a_j + \sum_{k=1}^{N-2} ka_k(y+z) \right) + (N/2 - 1)yz (2N^2 - N).$$

For $Sp(2N)$, with $R = [a_1 a_2 \dots a_N]$,

$$C_2(R) = \left(\sum_{k=1}^N [ka_k^2 + (2N+1-k)a_k] + \sum_{i=1}^{N-1} \sum_{j=i+1}^N 2ia_i a_j \right) 2N(2N+1).$$

And for $SO(2N+1)$, with $R = [a_1 a_2 \dots a_{N-1} z]$,

$$C_2(R) = \left(\sum_{k=1}^{N-1} [ka_k^2 + (2N-k)a_k] + (Nz^2 + 2N^2z)/4 + \sum_{k=1}^{N-1} ka_k z \sum_{i=1}^{N-2} \sum_{j=i+1}^{N-1} 2ia_i a_j \right) N(2N+1).$$

The Dynkin index (I in Ref. 1) is given by $I(R) = C_2(R) \times \dim(R)$. These values of $C_2(R)$ are normalized such that C_2 of an elementary spinor = C_2 of the defining ($2N$ -dimensional) representation of $Sp(2N) = 1$. For the purposes of MAC analysis, this normalization can be specified arbitrarily.

In principle, one would also like to have at hand rules for dimensions, Kronecker products, and $C_2(R)$ for the exceptional groups. However, the procedures would certainly contain enough different cases and exceptions that they would not shed much light on matters, and would be sufficiently cumbersome to preclude any thought of hand calculations. In these cases, the use of Schur functions is required.

APPENDIX: KRONECKER PRODUCTS

For compactness, the representations are referred to by their dimension, when no ambiguity results and the dimension is less than 1000 so that the representation appears in Ref. 1.

SO(7)

$$\begin{aligned}
 7 \times 7 &= 1 + 21 + 27, \\
 7 \times 8 &= 8 + 48, \\
 7 \times 21 &= 7 + 35 + 105, \\
 7 \times 27 &= 7 + 77 + 105, \\
 7 \times 35 &= 21 + 35 + 189, \\
 7 \times 48 &= 8 + 48 + 112 + 168_s, \quad 112 = [011], \quad 112' = [003], \\
 7 \times 77 &= 27 + 182 + 330, \quad 168_s = [201], \quad 168 = [020]; \\
 8 \times 8 &= 1 + 7 + 21 + 35, \\
 8 \times 21 &= 8 + 48 + 112, \\
 8 \times 27 &= 48 + 168_s, \\
 8 \times 35 &= 8 + 48 + 112 + 112', \\
 8 \times 48 &= 7 + 21 + 27 + 35 + 105 + 189, \\
 8 \times 77 &= 168_s + 448; \\
 21 \times 21 &= 1 + 21 + 27 + 35 + 168 + 189, \\
 21 \times 27 &= 21 + 27 + 189 + 330, \\
 21 \times 35 &= 7 + 21 + 35 + 105 + 189 + 378, \\
 21 \times 48 &= 8 + 48 + 48 + 112 + 112' + 168_s + 512, \\
 21 \times 77 &= 77 + 105 + 616 + 819; \\
 27 \times 27 &= 1 + 21 + 27 + 168 + 182 + 330, \\
 27 \times 35 &= 35 + 105 + 189 + 616, \\
 27 \times 48 &= 8 + 48 + 112 + 168_s + 448 + 512, \\
 27 \times 77 &= 7 + 77 + 105 + 378 + 693 + 819; \\
 35 \times 35 &= 1 + 7 + 21 + 35 + 105 + 168 + 189 + 294 \\
 &+ 378, \\
 35 \times 48 &= 8 + 48 + 48 + 112 + 112 + 112' + 168_s \\
 &+ 512 + 560, \\
 35 \times 77 &= 189 + 330 + 616 + [302] = 1560; \\
 48 \times 48 &= 1 + 7 + 21 + 21 + 27 + 35 + 35 + 77 + 105 \\
 &+ 105 + 168 + 189 + 189 + 330 + 378 + 616, \\
 48 \times 77 &= 48 + 168_s + 448 + 512 + [401] \\
 &= 1008 + [211] = 1512, \\
 77 \times 77 &= 1 + 21 + 27 + 168 + 330 + 714 + 825 \\
 &+ [410] = 1750 + [600] = 1911.
 \end{aligned}$$

SO(9)

$$\begin{aligned}
 9 \times 9 &= 1 + 36 + 44, \\
 9 \times 16 &= 16 + 128, \\
 9 \times 36 &= 9 + 84 + 231,
 \end{aligned}$$

$$\begin{aligned}
 9 \times 44 &= 9 + 156 + 231, \\
 9 \times 84 &= 36 + 126 + 594; \\
 16 \times 16 &= 1 + 9 + 36 + 84 + 126, \\
 16 \times 36 &= 16 + 128 + 432, \\
 16 \times 44 &= 128 + 576, \\
 16 \times 84 &= 16 + 128 + 432 + 768; \\
 36 \times 36 &= 1 + 44 + 126 + 495 + 594, \\
 36 \times 44 &= 36 + 44 + 594 + 910, \\
 36 \times 84 &= 9 + 84 + 126 + 231 + 924 + [0110] = 1650; \\
 44 \times 44 &= 1 + 36 + 44 + 450 + 495 + 910, \\
 44 \times 84 &= 84 + 231 + 924 + [2010] = 2457, \\
 84 \times 84 &= 1 + 44 + 84 + 126 + 495 + 594 + 924 \\
 &+ [0020] = 1980 + [0102] = 2772.
 \end{aligned}$$

SO(11)

$$\begin{aligned}
 11 \times 11 &= 1 + 55 + 65, \\
 11 \times 32 &= 32 + 320, \\
 11 \times 55 &= 11 + 165 + 429, \\
 11 \times 65 &= 11 + 275 + 429; \\
 32 \times 32 &= 1 + 11 + 55 + 165 + 330 + 462, \\
 32 \times 55 &= 32 + 320 + [01001] = 1408, \\
 32 \times 65 &= 320 + [20001] = 1760; \\
 55 \times 55 &= 1 + 55 + 65 + 330 + [02000] \\
 &= 1144 + [10100] = 1430, \\
 55 \times 65 &= 55 + 65 + 1430 + [21000] = 2025, \\
 65 \times 55 &= 1 + 55 + 65 + 935 + 1144 + 2025.
 \end{aligned}$$

SO(13)

$$\begin{aligned}
 13 \times 13 &= 1 + 78 + 90, \\
 13 \times 64 &= 64 + 768, \\
 13 \times 78 &= 13 + 286 + 715, \\
 (715 &= [110000]; 715' = [000100]) \\
 13 \times 90 &= 13 + 442 + 715; \\
 64 \times 64 &= 1 + 13 + 78 + 286 + 715 + [000010] = \\
 &1287 + [000002] = 1716, \\
 64 \times 78 &= 64 + 768 + [010001] = 4160, \\
 64 \times 90 &= 768 + [200001] = 4992; \\
 78 \times 78 &= 1 + 78 + 90 + 715' + [020000] \\
 &= 2275 + [101000] = 3925, \\
 78 \times 90 &= 78 + 90 + [210000] = 2927 + 3925, \\
 90 \times 90 &= 1 + 78 + 90 + 2275 + [400000] \\
 &= 2629 + 2927.
 \end{aligned}$$

SO(2N + 1) N > 6

$$(2N + 1) \otimes (2N + 1) = 1 \otimes N(2N + 1) \otimes (2N^2 + 3N).$$

SO(8)

$$\begin{aligned}
 \{ 8 \times 8 &= 1 + 28 + 35 \quad (R_1 = R_2), \\
 \{ 8 \times 8 &= 8 + 56 \quad (R_1 \neq R_2), \\
 8 \times 28 &= 8 + 56 + 160, \\
 \{ 8 \times 35 &= 8 + 112 + 160, \\
 \{ 8 \times 35 &= 56 + 224, \\
 \{ 8 \times 56 &= 28 + 35 + 35 + 350, \\
 \{ 8 \times 56 &= 8 + 56 + 160 + 224; \\
 28 \times 28 &= 1 + 28 + 35 + 35 + 35 + 300 + 350, \\
 28 \times 35 &= 28 + 35 + 350 + 560, \\
 28 \times 56 &= 8 + 56 + 56 + 160 + 224 + 224 + 840; \\
 \{ 35 \times 35 &= 1 + 28 + 35 + 294 + 300 + 567, \\
 \{ 35 \times 35 &= 35 + 350 + 840,
 \end{aligned}$$

$$\begin{aligned}
\{35 \times 56 &= 8 + 56 + 160 + 224 + 672 + 840, \\
\{35 \times 56 &= 56 + 160 + 224 + 224 + [2011] \\
&= [\text{BK}:1296] \text{ (or } [1021] \text{ or } [1012]), \\
56 \times 56 &= 1 + 28 + 28 + 35 + 35 + 35 + 300 + 350 + 350 + 567 + 567 + 840, \\
56 \times 56 &= 8 + 56 + 56 + 112 + 160 + 160 + 224 + 224 + 840 + 1296.
\end{aligned}$$

SO(10)

Representations appearing here are:

1 = [00000],	1050 = [10002],	5280 = [10003],
10 = [10000],	1200 = [00101],	5940 = [01011],
16 = [00001],	1386 = [21000],	6930 = [00013],
45 = [01000],	1440 = [00012],	7644 = [03000],
54 = [20000],	1728 = [10011],	7920 = [40001],
120 = [00200],	1782 = [50000],	8085 = [20011],
126 = [00002],	2640 = [30001],	8800 = [10101],
144 = [10001],	2772 = [00004],	8910 = [00022],
210 = [00011],	2970 = [01100],	10560 = [00111],
210' = [30000],	3696 = [01002],	11088 = [10012],
320 = [11000],	3696 _s = [11001],	12870 = [41000],
560 = [01001],	4125 = [00200],	4784 = [30100],
660 = [40000],	4290 = [60000],	15120 = [21001],
672 = [00003],	4312 = [20100],	16380 = [22000],
720 = [20001],	4410 = [12000],	17325 = [30002],
770 = [02000],	4608 = [31000],	21860 = [30011],
945 = [10100],	4950 = [20002],	

$$\begin{aligned}
10 \times 10 &= 1 + 45 + 54, \\
10 \times 16 &= \overline{16} + 144, \\
10 \times 45 &= 10 + 120 + 320, \\
10 \times 54 &= 10 + 210' + 320, \\
10 \times 120 &= 45 + 210 + 945, \\
10 \times 144 &= 16 + 144 + 560 + 720, \\
10 \times 210 &= 120 + 126 + 126 + 1728, \\
10 \times 210' &= 54 + 660 + 1386; \\
\{16 \times \overline{16} &= 1 + 45 + 210 \quad ([00001] \times [00010]), \\
\{16 \times 16 &= 10 + 120 + 126 \quad ([00001] \times [00001]), \\
16 \times 45 &= 16 + \overline{144} + 560, \\
16 \times 54 &= 144 + 720, \\
16 \times 120 &= \overline{16} + 144 + \overline{560} + 1200, \\
\{16 \times 126 &= 144 + 672 + 1200 \quad ([00001] \times [00002]), \\
\{16 \times \overline{126} &= 16 + 560 + 1440 \quad ([00001] \times [00020]), \\
\{16 \times \overline{144} &= 10 + 120 + 126 + 320 + 1728 \quad ([00001] \times [10010]), \\
\{16 \times 144 &= 45 + \overline{54} + 210 + \overline{945} + 1050 \quad ([00001] \times [10001]), \\
16 \times 210 &= 16 + 144 + 560 + \overline{1200} + 1440, \\
16 \times 210' &= 720 + 2640; \\
45 \times 45 &= 1 + 45 + 54 + 210 + 770 + 945, \\
45 \times 54 &= 45 + 54 + 945 + 1386, \\
45 \times 120 &= 10 + 120 + 126 + 320 + 1728 + 2970, \\
45 \times 126 &= \overline{120} + 126 + 1728 + 3696, \\
45 \times 144 &= \overline{16} + 144 + 144 + 560 + \overline{720} + 1200 + 3696_s, \\
45 \times 210 &= 45 + 210 + 210 + 945 + 1050 + 1050 + 5940, \\
45 \times 210' &= 210' + 320 + 4312 + 4608; \\
54 \times 54 &= 1 + 45 + 54 + 660 + 770 + 1386, \\
54 \times 120 &= 120 + 320 + 1728 + 4312, \\
54 \times 126 &= \overline{126} + 1728 + \overline{4950}, \\
54 \times 144 &= \overline{16} + 144 + \overline{560} + \overline{720} + 2640 + 3696, \\
54 \times 210 &= 210 + 945 + 1050 + 1050 + 8085, \\
54 \times 210' &= 10 + 210' + 320 + 1782 + 4410 + 4608; \\
120 \times 120 &= 1 + 45 + 54 + 210 + 210 + 770 + 945 + 1050 + 1050 + 4125 + 5940,
\end{aligned}$$

$$\begin{aligned}
120 \times 126 &= 45 + 210 + 945 + 1050 + 5940 + 6930, \\
120 \times 144 &= 16 + \overline{144} + \overline{144} + 560 + 560 + 720 + \overline{1200} + 1440 + \overline{3696_s} + 8800, \\
\{ 120 \times 210 &= 10 + 120 + 120 + 126 + 126 + 320 + 1728 + 1728 + 2970 + 3696 + 3696 + 10560, \\
120 \times 210' &= 945 + 1386 + 8085 + 14784 \\
126 \times 126 &= 1 + 45 + 210 + 770 + 5940 + 8910, \\
\{ 126 \times 126 &= 54 + 945 + 1050 + 2772 + 4125 + 6930, \\
126 \times \overline{144} &= 16 + 144 + 560 + 1200 + 1440 + 3696_s + 11088 \quad ([00002] \times [10010]), \\
126 \times 144 &= 144 + 560 + 720 + 1200 + 1440 + 5280 + 8800 \quad ([00002] \times [10001]), \\
\{ 126 \times 210 &= 10 + 120 + 126 + 320 + 1728 + 2970 + 3696 + 6930 + 10560, \\
126 \times 210' &= 1050 + 8085 + 17325; \\
144 \times \overline{144} &= 1 + 45 + 45 + 54 + 210 + 210 + 770 + 945 + 945 + 1050 + 1050 + 1386 + 5940 + 8085 \\
&\quad ([10001] \times [10010]), \\
144 \times 144 &= 10 + 120 + 120 + 126 + 126 + 210 + 320 + 320 + 1728 + 1728 + 2970 + 4312 + 4950 + 3696 \\
&\quad ([10001] \times [10001]), \\
144 \times 210 &= \overline{16} + 144 + \overline{144} + \overline{560} + \overline{560} + 642 + \overline{720} + 1200 + 1200 + \overline{1440} + 3696_s + \overline{8800} + 11088, \\
144 \times 210' &= \overline{144} + 720 + \overline{2640} + \overline{3696_s} + 7920 + 15120; \\
210 \times 210 &= 1 + 45 + 45 + 54 + 210 + 210 + 770 + 945 + 45 + 1050 + 1050 + 4125 + 5940 + 5940 \\
&\quad + 6930 + 6930 + 8910; \\
210 \times 210' &= 1728 + 4312 + 4950 + 4950 + 28160, \\
210 \times 210' &= 1 + 45 + 54 + 770 + 1386 + 4290 + 7644 + 12870 + 16390.
\end{aligned}$$

SO(12)

$$\begin{aligned}
12 \times 12 &= 1 + 66 + 77, \\
12 \times 32 &= \overline{32} + 352_s, \quad (352_s = [100001]), \\
12 \times 66 &= 12 + 220 + 560, \\
12 \times 77 &= 12 + 352 + 560 \quad (352 = [300000]); \\
32 \times 32 &= 1 + 66 + 462 + 495, \\
32 \times \overline{32} &= 12 + 220 + 792, \\
32 \times 66 &= 32 + \overline{352_s} + [010001] = 1728, \\
32 \times 77 &= \overline{352_s} + [200001] = 2112; \\
66 \times 66 &= 1 + 66 + 77 + 495 + [020000] = 1638 + [101000] = 2079, \\
66 \times 77 &= 66 + 77 + 2079 + [210000] = 2860, \\
77 \times 77 &= 1 + 66 + 77 + [400000] = 1287 + 1638 + 2860.
\end{aligned}$$

SO(14)

$$\begin{aligned}
14 \times 14 &= 1 + 91 + 104, \\
14 \times 64 &= \overline{64} + 832, \\
14 \times 91 &= 14 + 304 + 896; \\
64 \times \overline{64} &= 1 + 91 + [0001000] = 1001 + [0000011] = 3003, \\
64 \times 64 &= 14 + 364 + [0000002] = 1716 + [0000100] = 2002, \\
64 \times 91 &= \overline{64} + 832 + [0100001] = 4928; \\
91 \times 91 &= 1 + 91 + 104 + 1001 + [0200000] = 2240 + [1010000] = 4844.
\end{aligned}$$

SO(2N) N > 7

$$(2N) \oplus (2N) = (1) \quad (2N^2 - N) \quad (2N^2 + N - 1).$$

Sp(6)

$$\begin{aligned}
6 \times 6 &= 1 + 14_a + 21 \quad (14_a = [010], \quad 14_b = [001]), \\
6 \times 14_a &= 14_b + 64 + 6, \\
6 \times 14_b &= 14_a + 70 \\
6 \times 21 &= 6 + 56 + 64, \\
6 \times 56 &= 21 + 126' + 189 \quad (126' = [400] \quad 126 = [011]) \\
6 \times 64 &= 14_a + 21 + 70 + 90 + 189, \\
6 \times 70 &= 14_b + 64 + 126 + 216, \\
6 \times 84 &= 126 + 378, \\
6 \times 90 &= 64 + 126 + 357; \\
14_a \times 14_b &= 6 + 64 + 126, \\
14_a \times 14_a &= 1 + 14_a + 21 + 70 + 90, \\
14_a \times 21 &= 14_a + 21 + 70 + 189, \\
14_a \times 56 &= 56 + 64 + 216 + 448,
\end{aligned}$$

$$\begin{aligned}
14_a \times 64 &= 6 + 14_b + 56 + 64 + 64 + 126 + 216 + 350, \\
14_a \times 70 &= 14_a + 21 + 70 + 84 + 90 + 189 + 512, \\
14_a \times 84 &= 70 + 512 + 594, \\
14_a \times 90 &= 14_a + 70 + 90 + 189 + 385 + 512; \\
14_b \times 14_a &= 1 + 21 + 84 + 90, \\
14_b \times 21 &= 14_b + 64 + 216, \\
14_b \times 56 &= 70 + 189 + 525, \\
14_b \times 64 &= 14_a + 21 + 70 + 90 + 189 + 512, \\
14_b \times 70 &= 6 + 56 + 64 + 126 + 350 + 378, \\
14_b \times 84 &= 14_b + 216 + 330 + 616, \\
14_b \times 90 &= 14_b + 64 + 216 + 350 + 616; \\
21 \times 21 &= 1 + 14_a + 21 + 90 + 126' + 189, \\
21 \times 56 &= 6 + 56 + 64 + 252 + 350 + 448, \\
21 \times 64 &= 6 + 14_b + 56 + 64 + 64 + 126 + 216 + 350 + 448, \\
21 \times 70 &= 14_a + 70 + 70 + 90 + 189 + 512 + 525, \\
21 \times 84 &= 84 + 90 + 512 + [202] = 1078, \\
21 \times 90 &= 21 + 70 + 84 + 90 + 189 + 512 + 924_a \quad (924_a = [220], \quad 924_b = [410]); \\
56 \times 56 &= 1 + 12 + 21 + 90 + 126' + 189 + 385 + 462 + 924_a + 924_b, \\
56 \times 64 &= 14_a + 21 + 70 + 90 + 512 + 126' + 189 + 189 + 525 + 924_a + 924_b, \\
56 \times 70 &= 14_a + 64 + 126 + 216 + 216 + 350 + 448 + [401] = 1100 + [211] = 1386, \\
56 \times 84 &= 126 + 350 + 378 + 1386 + [302] = 2464, \\
56 \times 90 &= 56 + 64 + 126 + 216 + 350 + 378 + 448 + 1386 + [320] = 2016; \\
64 \times 64 &= 1 + 14_a + 14_a + 21 + 21 + 70 + 70 + 70 + 84 + 90 + 90 + 189 + 189 \\
&\quad + 189 + 385 + 512 + 512 + 525 + 924_a \\
64 \times 70 &= 6 + 14_b + 56 + 64 + 64 + 64 + 126 + 126 + 216 + 216 + 350 + 350 \\
&\quad + 378 + 448 + 616 + 1386, \\
64 \times 84 &= 64 + 126 + 216 + 616 + 350 + 378 + 1386 + [112] = 2240, \\
64 \times 90 &= 6 + 14_a + 64 + 64 + 126 + 126 + 216 + 216 + 350 + 350 + 378 \\
&\quad + 448 + 616 + [130] = 1344 + 1386; \\
70 \times 70 &= 1 + 14_a + 21 + 21 + 70 + 84 + 90 + 126' + 189 + 189 + 385 + 512 \\
&\quad + 512 + 594 + 924 + [202] = 1078, \\
70 \times 84 &= 14_a + 70 + 189 + 385 + 512 + 525 + 594 + [103] = 1386' + [121] = 2205, \\
70 \times 90 &= 14_a + 21 + 70 + 70 + 90 + 189 + 189 + 395 + 512 + 512 + 594 \quad + 924_a + [121] = 2205; = [0101]), \\
27 \times 48 &= 8 + 48 + 160 + 288 + 792_a; \\
36 \times 36 &= 1 + 27 + 36 + 308 + 330 + 594, \\
36 \times 42 &= 42 + 315 + [2001] = 1155, \\
36 \times 48 &= 48 + 160 + 288 + [2010] = 1232; \\
42 \times 42 &= 1 + 36 + 308 + 594 + 825, \\
42 \times 48 &= 8 + 160 + 792_a + [0011] = 1056, \\
48 \times 48 &= 1 + 27 + 36 + 308 + 315 + 792_b + 825.
\end{aligned}$$

Sp(10)

$$\begin{aligned}
10 \times 10 &= 1 + 44 + 55, \\
10 \times 44 &= 10 + 110 + 320, \\
10 \times 55 &= 10 + 220 + 320; \\
44 \times 44 &= 1 + 44 + 55 + 65 + 780 + 891, \\
44 \times 55 &= 44 + 55 + 891 + [21000] = 1430, \\
55 \times 55 &= 1 + 44 + 55 + 715 + 780 + 1430.
\end{aligned}$$

Sp(12)

$$\begin{aligned}
12 \times 12 &= 1 + 65 + 78, \\
12 \times 65 &= 12 + 208 + 560, \\
12 \times 78 &= 12 + 364 + 560; \\
65 \times 65 &= 1 + 65 + 78 + 429 + [020000] \\
&= 1650 + [120000] = 2001, \\
65 \times 78 &= 65 + 78 + 2002 + [210000] = 2925, \\
78 \times 78 &= 1 + 65 + 78 + [400000] \\
&= 1365 + 1650 + 2925.
\end{aligned}$$

Sp(14)

$$\begin{aligned}
14 \times 14 &= 1 + 90 + 105, \\
14 \times 90 &= 14 + 350 + 896, \\
90 \times 90 &= 1 + 105 + 90 + 910 + [0200000] \\
&= 3094 + [1010000] = 3900.
\end{aligned}$$

Sp(2N), N > 7

$$(2N) \times (2N) = 1 + [N(2N - 1) - 1] + [N(2N + 1)].$$

¹J. Patera and D. Sankoff, *Tables of Branching Rules for Representations of Simple Lie Algebras* (Les Presses de l'Universite de Montreal, 1973).

²B. Wybourne, *Symmetry Principles and Atomic Spectroscopy* (Wiley-Interscience, New York, 1970).

³H. Weyl, *The Classical Groups* (Princeton U. P., Princeton, N. J., 1939).

⁴S. Dimopoulos and L. Susskind, *Nucl. Phys. B* **155**, 237 (1979).

⁵S. Raby, S. Dimopoulos, and L. Susskind, "Tumbling Gauge Theories," Preprint ITP-635-Stanford, 1979.

Analytical approach to initial-value problems in nonlinear systems

L. Brenig

Service de Chimie-Physique II, Faculté des Sciences, Université Libre de Bruxelles, Brussels, Belgium

V. Fairén

Departamento de Física de Fluidos, Facultad de Ciencias, C-3, Universidad Autónoma de Madrid, Cantoblanco, Madrid, Spain

(Received 16 October 1979; accepted for publication 14 March 1980)

A method based on the equivalence between finite-dimensional nonlinear and infinite-dimensional linear systems of ordinary differential equations is presented in order to calculate the time-dependent solutions of nonlinear physico-chemical systems. The solution is cast in the form of power series whose general term is known analytically.

PACS numbers: 02.30.Jr

I. INTRODUCTION

The time asymptotic properties of nonlinear physico-chemical systems like those in reaction-diffusion, heterogeneous catalysis, hydrodynamics, and lasers have been extensively studied during these last years.¹ On the other hand, in spite of their crucial importance, very little is known about the transient regimes and, in general, time-evolution problems. Indeed, except for very simple and unrealistic exactly solvable models and certain scaling theories,² no general analytic theory concerning transient nonlinear regimes is known. Nowadays, numerical integration is the only general tool for obtaining quantitative information about phenomena like chaos,³ turbulence,⁴ and time behaviour in nonequilibrium phase transition.⁵

In 1931, T. Carleman,⁶ following a suggestion due to Poincaré, showed that any finite-dimensional nonlinear system of ordinary differential equations is equivalent to an infinite-dimensional linear system of ordinary differential equations. This idea remained unexploited until, recently, Montroll *et al.*⁷ have applied it. They suggest, essentially, a technique using Laplace transform and infinite matrix inversion to solve the infinite-dimensional linear system associated with the nonlinear original problem. In this paper, we develop Carleman's idea along a different way and obtain a nonperturbative and easy to handle theory to find the time-dependent solution of initial-value problems. Our main result is that, for nonlinear systems of ordinary first order differential equations with analytical vector function in the right-hand side, the general coefficient of the Taylor series of the solution can be obtained in an explicit and compact form. We show that this general Taylor's coefficient is related to the eigenvalues and eigenvectors of the infinite upper triangular matrix associated to the Carleman's infinite linear system.

In Sec. II, we first introduce Carleman's infinite linear system of equations and present our method to find the solution in terms of eigenvalue and eigenvectors of the associated infinite triangular matrix. We apply, in Sec. III, this method to an exactly solvable example and show that it leads very easily to the exact solution.

II. GENERAL METHOD

A. Extension to an infinite linear system

Let us consider the following general system of nonlinear ordinary differential equations

$$\frac{d}{dt} x_i(t) = F_i(\{x_j(t)\}), \quad i, j = 1, 2, \dots, s, \quad (1)$$

with initial conditions

$$x_i(t=0) = x_i(0), \quad i = 1, \dots, s, \quad (2)$$

and where the $F_i(\{x_j\})$ are in general nonlinear analytical functions on a compact domain \mathcal{D} of \mathbb{R}^s .

As a first step, let us separate linear and nonlinear part of system (1)

$$\frac{d}{dt} x(t) = L x(t) + N(x(t)), \quad (3)$$

in which $x(t)$ is the s -dimensional vector $(x_1(t), x_2(t), \dots, x_s(t))$ and where the $s \times s$ matrix L and the nonlinear operator \mathcal{N} are, respectively, the linear and nonlinear parts of

$$F = \{F_i(\{x_j\}), \quad i = 1, \dots, s\}.$$

We now introduce the linear transformation T :

$$T x = y, \quad (4)$$

which reduces L to its upper-normal Jordan triangular form⁸ (note that the new variable y may be complex). Under T , system (1) is transformed into

$$\frac{d}{dt} y(t) = L_J y(t) + M(y(t)), \quad (5)$$

with initial conditions $y_j(t=0) = \sum_{i=1}^s T_{ji} x_i(0)$, $j, i = 1, \dots, s$, and where $L_J = T L T^{-1}$ is a Jordan triangular $s \times s$ matrix and $M(y) = T N(T^{-1} y)$ is the T -transformed nonlinear operator.

Using the analyticity of $F(x)$ of Eq. (1) it is possible to write the rhs of Eq. (5) as a linear superposition of functions of the form

$$P(m_1, m_2, \dots, m_s) = y_1^{m_1} y_2^{m_2} \dots y_s^{m_s}, \quad \begin{cases} m_1, m_2, \dots, m_s = 0, 1, 2, \dots, \infty \\ \text{and} \\ \sum_{i=1}^s m_i > 0. \end{cases} \quad (6)$$

It is always possible to order the functions $P(m_1, \dots, m_s)$ in a simple sequence $\{P_1, P_2, \dots, P_k, \dots\}$ of increasing power $p \equiv \sum_{i=1}^s m_i$, where the index k is uniquely determined by the sequence (m_1, m_2, \dots, m_s) . Following Carleman's original idea, we next introduce an infinite-dimensional vector \mathcal{L} whose general component \mathcal{L}_k is identified with P_k , $\mathcal{L}_k \equiv P_k$. In this way, the finite-dimensional nonlinear system (5) [which is equivalent to system (1)] is transformed into an equivalent infinite-dimensional linear system which reads

$$\frac{d}{dt} \mathcal{L}(t) = A \mathcal{L}(t), \quad (7)$$

with initial condition for component \mathcal{L}_k of \mathcal{L} given by

$$\begin{aligned} \mathcal{L}_k(t=0) &\equiv P_k(t=0) = y_1^{m_1}(t=0) y_2^{m_2}(t=0) \dots y_s^{m_s}(t=0), \\ & \quad k = 1, 2, \dots, \infty \\ &= \mathcal{L}_k(0). \end{aligned} \quad (7a)$$

An important general feature of (7) is that the infinite matrix A is always triangular, $a_{ij} = \begin{cases} 0, & i > j \\ \neq 0, & i < j \end{cases}$. This is a universal characteristic of infinite matrices originating from nonlinear systems of equations like (1). Triangularity of A is of great convenience for explicit calculations of solutions of system (7) as we show below. For convenience, we use from here on the Dirac bracket notation. System (7) is thus

$$\frac{d}{dt} |\mathcal{L}(t)\rangle = A |\mathcal{L}(t)\rangle. \quad (8)$$

The formal solution of system (8) reads

$$|\mathcal{L}(t)\rangle = \exp(tA) |\mathcal{L}(0)\rangle. \quad (9)$$

It should be stressed that the existence of solution (9) depends on the choice of initial conditions $\{\mathcal{L}_k(0)\}$ and, in general, cannot be ensured for arbitrary value of time t .

B. Explicit calculation of Solution (9)

A first step toward the explicit solution (9) amounts to expanding $\exp(tA)$ in powers of matrix A

$$\varphi_j^{(n)} = \begin{cases} \sum_{\mathcal{P}\{j, j+1, \dots, n\}} (-1)^{q+1} \frac{a_{j_1 j_2} a_{j_2 j_3} \dots a_{j_{q-1} j_q}}{(a_{j_1 j_1} - \lambda_{(n)})(a_{j_2 j_2} - \lambda_{(n)}) \dots (a_{j_{q-1} j_{q-1}} - \lambda_{(n)})}, & 1 \leq j < n \\ 1, & j = n \\ 0, & j > n \end{cases} \quad n = 1, 2, \dots, \infty, \quad (17)$$

where $\mathcal{P}\{j, j+1, \dots, n\}$ is the set of all possible ordered subsets of $\{j, j+1, \dots, n\}$ and is defined as follows:

$$\mathcal{P}\{j, j+1, \dots, n\} = \{(j_1, \dots, j_q) | j = j_1 < j_2 < \dots < j_q = n; (j_1, \dots, j_q) \subset \{j, j+1, \dots, n\}\} \quad (18)$$

the solution of (16) reads

$$\psi_j^{(n)} = \begin{cases} \sum_{\mathcal{P}\{n, n+1, \dots, j\}} (-1)^{q+1} \frac{a_{i_1 i_2} a_{i_2 i_3} \dots a_{i_{q-1} i_q}}{(a_{i_1 i_1} - \lambda_{(n)})(a_{i_2 i_2} - \lambda_{(n)}) \dots (a_{i_{q-1} i_{q-1}} - \lambda_{(n)})}, & j > n \\ 1, & j = n \\ 0, & 1 \leq j < n \end{cases} \quad n = 1, 2, \dots, \infty, \quad (19)$$

where

$$\mathcal{P}\{n, n+1, \dots, j\} = \{(i_1, \dots, i_q) | n = i_1 < i_2 < \dots < i_q = j; (i_1, \dots, i_q) \subset \{n, n+1, \dots, j\}\}. \quad (20)$$

$$|\mathcal{L}(t)\rangle = \left\{ \sum_{n=0}^{\infty} \frac{t^n}{n!} A^n \right\} |\mathcal{L}(0)\rangle, \quad (10)$$

or, in terms of components

$$\mathcal{L}_i(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \sum_{i_1=1}^{\infty} \sum_{i_2=1}^{\infty} \dots \sum_{i_n=1}^{\infty} \times a_{i i_1} a_{i_1 i_2} \dots a_{i_{n-1} i_n} \mathcal{L}_{i_n}(0), \quad i = 1, 2, \dots, \infty. \quad (11)$$

However, in most cases, the explicit calculation of the general term of series (11) is extremely complicated and does not lead to a treatable algorithm. To circumvent this difficulty, we suggest the following scheme. First, let us calculate the eigenvalues and left and right eigenvectors of matrix A

$$A |\varphi^{(n)}\rangle = \lambda_{(n)} |\varphi^{(n)}\rangle, \quad (12)$$

$$n = 1, 2, \dots, \infty.$$

$$\langle \psi^{(n)} | A = \lambda_{(n)} \langle \psi^{(n)} |, \quad (13)$$

By construction of System (7) from System (5), the diagonal elements of A , a_{nn} , are just linear combinations of the diagonal elements of L_j [which obviously, are the eigenvalues of L of System (3)]. Since matrix A is triangular, its eigenvalues are equal to its successive diagonal elements

$$\lambda_{(n)} = a_{nn}, \quad (14)$$

and so are linear combinations of the eigenvalues of L . As for the right and left eigenvectors of A , we have to solve the two systems (12) and (13) or, more explicitly,

$$\sum_{p=m+1}^{\infty} a_{mp} \varphi_p^{(n)} + (a_{mm} - \lambda_{(n)}) \varphi_m^{(n)} = 0, \quad n, m = 1, 2, \dots, \infty, \quad (15)$$

for the right eigenvector $|\varphi^{(n)}\rangle$, and

$$\sum_{p=1}^{m-1} a_{pm} \psi_p^{(n)} + (a_{mm} - \lambda_{(n)}) \psi_m^{(n)} = 0, \quad n, m = 1, 2, \dots, \infty, \quad (16)$$

for the left eigenvector $\langle \psi^{(n)} |$. The triangularity of matrix A and the recurrence relations which, by construction, exist between its elements allow us to solve exactly, in general, Systems (15) and (16). The explicit solution of (15) for the j th component of $|\varphi^{(n)}\rangle$ is

A more compact form can be written:

$$\varphi_j^{(n)} = \left\{ \begin{array}{ll} \sum_{r=1}^{n-j} (-1)^r \left[\left(\frac{1}{PA P - \lambda_{(n)}} PAQ \right) \right]_{jn}^r, & 1 < j < n, \\ 1, & j = n, \\ 0, & j > n, \end{array} \right\} \quad n = 1, 2, \dots, \infty, \quad (21)$$

and

$$\psi_j^{(n)} = \left\{ \begin{array}{ll} \sum_{r=1}^{j-n} (-1)^r \left[\left(PAQ \frac{1}{QAQ - \lambda_{(n)}} \right) \right]_{nj}^r, & j > n, \\ 1, & j = n, \\ 0, & 1 < j < n, \end{array} \right\} \quad n = 1, 2, \dots, \infty, \quad (22)$$

where P is a projector on the diagonal part of matrix A and $Q = I - P$. The orthonormality of the eigenvectors

$$\langle \psi^{(m)} | \varphi^{(n)} \rangle = \delta_{n,m}, \quad n, m = 1, 2, \dots, \infty, \quad (23)$$

and the completeness of the basis of eigenvectors:

$$\sum_{n=1}^{\infty} |\varphi^{(n)} \rangle \langle \psi^{(n)}| = I, \quad (I \text{ is the unity matrix}), \quad (24)$$

follow easily from the above cited peculiar properties of matrix A . Note that in contrast with the eigenvalues of A , which are linear combinations of the eigenvalues of the linear part of System (5), its eigenvectors depend on the full nonlinearity of the original System (5). It is a remarkable features of these infinite matrices that their spectral properties, eigenvalues, and eigenvectors, can be exactly calculated without having recourse to any perturbation technique.

We are now able to construct the explicit solution of System (8); We first project Eq. (10) on the j th component of vector $\mathcal{L}(t)$

$$\mathcal{L}_j(t) \equiv \langle e^{(j)} | \mathcal{L}(t) \rangle = \sum_{k=0}^{\infty} \frac{t^k}{k!} \langle e^{(j)} | A^k | \mathcal{L}(0) \rangle, \quad (25)$$

where the i th component of the unit vector $\langle e^{(j)} |$ is

$$e^{(j)}_i = \delta_{i,j}, \quad i, j = 1, 2, \dots, \infty. \quad (26)$$

Let us now introduce in Eq. (25) the closure relation (24). It leads to

$$\mathcal{L}_j(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \langle e^{(j)} | \varphi^{(n)} \rangle \langle \psi^{(n)} | A^k | \varphi^{(m)} \rangle \langle \psi^{(m)} | \mathcal{L}(0) \rangle \quad (27)$$

or, more explicitly, to

$$\mathcal{L}_j(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \sum_{n=1}^{\infty} \langle e^{(j)} | \varphi^{(n)} \rangle \langle \psi^{(n)} | \mathcal{L}(0) \rangle (\lambda_{(n)})^k \quad (28)$$

where use is made of Eqs. (12), (13), and (23). Expanding the scalar products, Eq. (28) reads

$$\mathcal{L}_j(t) = \sum_{k=0}^{\infty} \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \frac{t^k}{k!} (\lambda_{(n)})^k \varphi_j^{(n)} \psi_m^{(n)} \mathcal{L}_m(0). \quad (29)$$

This is a series of powers of time t :

$\sum_{k=0}^{\infty} \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} t^k C_{knm}^{(j)}$, whose general coefficient is analytically known,

$$C_{knm}^{(j)} = \frac{(\lambda_{(n)})^k}{k!} \varphi_j^{(n)} \psi_m^{(n)} \mathcal{L}_m(0), \quad (30)$$

where $\varphi_j^{(n)}$ and $\psi_j^{(n)}$ are given by expressions (17)–(19) or (21) and (22), and $\lambda_{(n)}$ by (14).

In view of our original problem (5), the only relevant components of vector $|\mathcal{L}(t)\rangle$ are the s first ones

$$\mathcal{L}_j(t) = y_j(t), \quad j = 1, 2, \dots, s. \quad (31)$$

The other components of $|\mathcal{L}(t)\rangle$ are powers of the variables $\{y_j(t)\}$ [see Def. (6)]. Thus, the general time dependent solution of system (5) is given by

$$y_j(t) \equiv \mathcal{L}_j(t) = \sum_{k=0}^{\infty} \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} t^k \left[\frac{(\lambda_{(n)})^k}{k!} \varphi_j^{(n)} \psi_m^{(n)} \mathcal{L}_m(0) \right], \quad (32)$$

for $j = 1, 2, \dots, s$. Solution (32) is made explicit by inserting expressions (17)–(20) of (21) and (22) of the eigenvectors $|\varphi^{(n)}\rangle$ and $\langle \psi^{(n)}|$.

As the rhs of (5) is analytic [because of the analyticity of $F_i(\{x_j(t)\})$ in Eq. (1)] solution $y(t)$ of the initial-conditions problem (5) is also analytic with respect to time and initial values.⁹ Hence, we can state the following result: power series (32) is the Taylor series of the solution $y(t)$ of system (5). The solution $x(t)$ of the original system (1) and (2) can be obtained from (32) by applying the inverse of the linear transformation defined in (4). We stress the fact that, in (32), the general term is known and thus provides an *algorithm* for computing the solution in the whole set of times and initial values for which the Taylor series converges. In contrast with this, in other methods, due to the increasing difficulty of calculating higher order terms step by step, the Taylor series is restricted to the low order ones and, thus its validity is limited to short time-ranges. The knowledge of the general term of series (32) allows for evaluating the error made when truncating it, since the error is known to be of the order of the next term after the truncation. In the cases where the real part of the eigenvalues $\lambda_{(n)}$ are all negative, the rapidity of the convergence of series (32) can be increased by partially-summing it. This yields

$$y_j(t) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} e^{\lambda_{(n)}(t)} \varphi_j^{(n)} \psi_m^{(n)} \mathcal{L}_m(0). \quad (32a)$$

This is always possible when Eq. (5) admits an asymptotically-stable stationary solution.

III. AN EXAMPLE: THE LOGISTIC EQUATION

Here we test our method on the well-known one-dimen-

sional nonlinear equation

$$\frac{d}{dt} x(t) = -x(t)[1 - x(t)], \quad (33)$$

with initial condition $x(t=0) = x_0$. The exact solution of (33) obtained by a simple quadrature is

$$x(t) = \frac{x_0}{x_0 + (1 - x_0)e^{+t}}. \quad (34)$$

We now apply the method of Sec. II to calculate the solution of Eq. (33). Obviously, for a one-dimensional system, there is no need for a transformation of the linear part to the Jordan form.

The infinite-dimensional vector $\mathcal{L}(t)$ of Sec. II, reads here

$$\mathcal{L}(t) = \{\mathcal{L}_k(t)\}, \quad \text{with} \quad (35)$$

$$\mathcal{L}_k(t) = [x(t)]^k, \quad k = 1, 2, \dots, \infty.$$

From Eq. (33), the equations for $\mathcal{L}_k(t)$ are

$$\frac{d}{dt} \mathcal{L}_k(t) = -k\mathcal{L}_k + k\mathcal{L}_{k+1}, \quad k = 1, 2, \dots, \infty, \quad (36)$$

with initial condition $\mathcal{L}_k(t=0) = x_0^k$. The general element of matrix A associated with the infinite linear system (36) is

$$a_{kl} = -k\delta_{k,l} + k\delta_{k+1,l}, \quad k, l = 1, 2, \dots, \infty. \quad (37)$$

From (37), matrix A is upper triangular. As such, its eigenvalues are its diagonal elements

$$\lambda_{(k)} = -k, \quad k = 1, 2, \dots, \infty. \quad (38)$$

Using formulas (17)–(20), the j th component of the right eigenvector $|\varphi^{(n)}\rangle$ of A is:

$$\varphi_j^{(n)} = \begin{cases} (-1)^{n-j} \frac{(n-1)!}{(n-j)!(j-1)!}, & 1 \leq j \leq n \\ 0, & j > n \end{cases} \quad n = 1, 2, \dots, \infty, \quad (39)$$

whereas for the left eigenvectors $\langle \psi^{(n)}|$ one has

$$\psi_j^{(n)} = \begin{cases} \frac{(j-1)!}{(j-n)!(n-1)!}, & j \geq n \\ 0, & 1 \leq j < n \end{cases} \quad n = 1, 2, \dots, \infty, \quad (40)$$

Inserting expressions (38), (39), and (40) into formula (32), we obtain

$$x(t) = \sum_{k=0}^{\infty} \sum_{n=1}^{\infty} \sum_{j=n}^{\infty} t^k \frac{n^k}{k!} (-1)^{n-1+k} \frac{(j-1)!}{(j-n)!(n-1)!} x_0^j. \quad (41)$$

It can easily be seen by direct calculation that (41) coincides with the Taylor expansion of Solution (34) around $(t=0)$. The domain of absolute convergence of (41) corresponds to the values $t > \ln|(1 - x_0)/x_0|$.

IV. CONCLUSION

This method offers a convenient tool for calculating explicit expressions for the solution of the initial-condition problem for a general s -dimensional system of ordinary nonlinear differential equations. The interest of this approach is double. First, it is *nonperturbative*. Second, the solution is obtained in the form of a Taylor series [Eq. (32)] whose general term is known in a compact *analytical* form. This provides a suitable *algorithm* for computer calculations of transient phenomena which usually can only be attained by direct numerical integration. Eventually, the theory could be extended to time- and *space*-dependent systems like, for instance, those found in reaction-diffusion, hydrodynamical and other nonlinear physico-chemical problems.

ACKNOWLEDGMENT

We would like to thank the following people for many interesting discussions: Professor G. Nicolis and Dr. N. Banai, Dr. M. Courbage, Dr. J. Erneux, Dr. M. Hennenberg, Dr. J. Houard, Dr. M. Mareschal, Dr. A. Nazarea, Dr. J. Turner, Dr. J. W. Turner, and Dr. C. Van den Broek. V. F. wishes to thank Professor I. Prigogine and Professor G. Nicolis for the hospitality extended to him. He also acknowledges a fellowship from the Ministère de l'Education Nationale et de la Culture Française of Belgium and financial aid from the Instituto de Estudios Nucleares, Spain. L. B. was supported by the Association Euratom, Etat Belge.

¹G. Nicolis and I. Prigogine, *Self-organization in Nonequilibrium Systems*, (Interscience-Wiley, New York, 1977); H. Haken, *Synergetics, An Introduction*, second enlarged edition (Springer, Berlin, 1978); H. Haken, *Rev. Mod. Phys.* **47**, 67 (1975); I. Procaccia and J. Ross, *J. Chem. Phys.* **67**, 5558 and 5565 (1977).

²M. Suzuki, *Prog. Theor. Phys.* **56**, 77 (1976); K. Kawazaki and S. K. Kim, *J. Chem.* **68**, 319 (1978).

³E. N. Lorentz, *J. Atmos. Sci.* **20**, 130 (1963); O. E. Rössler, *Z. Naturforsch.* **319**, 1168 (1976).

⁴J. B. McLaughlin and P. C. Martin, *Phys. Rev. Lett.* **33**, 1189 (1974); S. A. Orszag, *Lect. Notes Phys.* **59**, 32 (1976); D. Ruelle and F. Takens, *Commun. Math. Phys.* **20**, 167 (1971).

⁵J. C. Legros and J. K. Platten, *Lect. Notes Phys.* **72**, 152 (1978).

⁶T. Carleman, *Ark. Mat., Astron. Fyz.* **228**, 63 (1931).

⁷E. W. Montroll and R. H. G. Helleman, *AIP Conf. Proc.* **27**, 75 (1976); E. W. Montroll, *AIP Conf. Proc.* **46**, 337 (1978).

⁸We suppose that the eigenvalues of L are nonzero and that no linear combination of them with integer coefficients can be zero. This is done in order to avoid subsequent problems of degeneracy.

⁹R. A. Struble, *Nonlinear Differential Equations* (McGraw-Hill, New York, 1962).

Laplace transforms and asymptotic expansions of orthogonal polynomials

M. L. Glasser

Clarkson College, Potsdam, New York 13676

(Received 14 November 1979; accepted for publication 22 February 1980)

New integral representations for orthogonal polynomials which possess a generating function are obtained by considering their Laplace transforms with respect to order. The method is used to derive some uniform asymptotic estimates for the associated Laguerre polynomials and to test an approximation scheme used in electron gas theory.

PACS numbers: 02.30.Qy

In this paper we outline the derivation of a number of new Laplace transform pairs by calculating the transforms of any family of functions, which possesses a generating relation, with respect to their order. This simple device does not appear to have been used before and leads easily to a number of remarkable formulas. The need for these results arises in studying the properties of electron gases in an external magnetic field and it is the Laguerre functions which are of concern. Accordingly, these polynomials are singled out for detailed consideration. The paper is concluded with a brief assessment of an approximation scheme that has frequently been invoked in the aforementioned studies.

The basic procedure is quite simple. Let the sequence of functions $\{\phi_n(x)\}$ be given by the generating formula

$$\sum_{n=0}^{\infty} A_n \phi_n(x) Z^n = F(Z). \quad (1)$$

Let square brackets denote the integer part as usual and consider the Laplace transform

$$I = \int_0^{\infty} e^{-pt} A_{[t]} \phi_{[t]}(x) dt. \quad (2)$$

By decomposing the range of integration into segments of unit length, one sees that all but the exponential function can be pulled out of the integral and we easily obtain

$$I = p^{-1}(1 - e^{-p}) \sum_{n=0}^{\infty} A_n \phi_n(x) e^{-np}. \quad (3)$$

Now by using (1), Eq. (3) reduces to

$$I = p^{-1}(1 - e^{-p}) F(e^{-p}). \quad (4)$$

Finally, if the function on the right-hand side of (4) is analytic in the half-plane $\text{Re } p \geq C$, we have the integral representation

$$A_{[t]} \phi_{[t]}(x) = \int_{C-i\infty}^{C+i\infty} \frac{dp}{2\pi i} \frac{e^{pt}}{p} (1 - e^{-p}) F(e^{-p}), \quad (5)$$

by taking the inverse Laplace transform of both sides of (4). Care must be exercised in letting $t = n$ in (5). This is because $t = n$ is a point of discontinuity of the function whose Laplace transform is being considered and the Bromwich integral representation in (5), like the Fourier representation, yields the average of the right and left limits at such a point. There are a number of ways of circumventing this problem, one of which is described in an example below.

For the associated Laguerre functions one easily derives the relation

$$(1 - e^{-2p})^{-\alpha} \exp(-x \coth p) = e^{-x}(1 - e^{-2p}) \sum_{n=0}^{\infty} L_n^{(\alpha)}(2x) e^{-2np} \quad (6)$$

from the generating function given in Ref. 1. By following the procedure outlined above we find

$$\int_{C-i\infty}^{C+i\infty} \frac{dp}{2\pi i} \frac{e^{p(\alpha+t)}}{p} (\text{csch } p)^{\alpha} e^{-x \coth p} = 2^{\alpha} e^{-x} L_{[(1/2)t]}^{(\alpha)}(2x), \quad C > 0. \quad (7)$$

Similarly, from the identity

$$(xy)^{-\alpha/2} \exp\left[\frac{p\alpha}{2} + \frac{1}{2}(x+y)(1 - \coth p)\right] \times I_{\alpha}(\sqrt{xy} \text{csch } \frac{1}{2}p) = \sum_{n=0}^{\infty} \frac{n!}{\Gamma(n + \alpha + 1)} L_n^{(\alpha)}(x) L_n^{(\alpha)}(y) (e^{-np} - e^{-(n+1)p}), \quad \text{Re } p > 0, \quad (8)$$

we find

$$\int_{C-i\infty}^{C+i\infty} \frac{dp}{2\pi i} \frac{e^{p(\alpha+t)}}{p} I_{\alpha}(\sqrt{xy} \text{csch } p) \times \exp\left\{\frac{1}{2}(x+y)(1 - \coth p)\right\} = (xy)^{\alpha/2} \frac{\Gamma([(1/2)t] + 1)}{\Gamma([(1/2)t] + \alpha + 1)} L_{[(1/2)t]}^{(\alpha)}(x) L_{[(1/2)t]}^{(\alpha)}(y), \quad C > 0. \quad (9)$$

Formulas of the type (8) and (9) can be produced in great profusion with little effort.

As a demonstration of the utility and manipulation of these formulas, we examine the correction to the uniform asymptotic limit²

$$\lim_{n \rightarrow \infty} n^{-\alpha} L_n^{(\alpha)}(x/n) = x^{-\alpha/2} J_{\alpha}(2\sqrt{x}). \quad (10)$$

We begin by replacing t by $2t$ in (7) and, as allowed by Watson's lemma, closing the contour by a large semicircle in the left half-plane which we then decompose into small counterclockwise circles C_k about the points $p = k\pi i$, $k = 0, \pm 1, \pm 2, \dots$, which are isolated essential singularities of the integrand. By making use of the periodicity of the hyperbolic functions and translating each of these circles to the origin we come to

$$2^{\alpha} e^{-x} L_{[t]}^{(\alpha)}(2x) = \oint \frac{ds}{2\pi i} \frac{e^{(2t+\alpha)s}}{s} (\text{csch } s)^{\alpha} e^{-x \coth s} + 2 \sum_{k=1}^{\infty} \cos(2\pi kt) \oint \frac{s ds}{2\pi i} \frac{e^{(2t+\alpha)s}}{s^2 + k^2 \pi^2} (\text{csch } s)^{\alpha} e^{-x \coth s}$$

$$+ 2\pi \sum_{k=1}^{\infty} k \sin(2\pi kt) \oint \frac{ds}{2\pi i} \frac{e^{(2t+\alpha)s}}{s^2 + k^2\pi^2} (\operatorname{csch} s)^\alpha e^{-x \coth s}. \quad (11)$$

Here it is assumed that $t = n + t_0$, $0 < \epsilon \leq t_0 \leq 1 - \epsilon < 1$, where $\epsilon > 0$ is small; we have also combined the integrals originally about C_k and C_{-k} , $k = 1, 2, \dots$. The trigonometric series are uniformly convergent with respect to s (and t_0) and can be summed inside the integrals:

$$\sum_{k=1}^{\infty} \frac{k \sin 2\pi kt}{s^2 + k^2\pi^2} = \frac{1}{2\pi} \frac{\sinh(1 - 2t_0)s}{\sinh s}, \quad (12)$$

$$\sum_{k=1}^{\infty} \frac{\cos 2\pi kt}{s^2 + k^2\pi^2} = \frac{1}{2s} \left[\frac{\cosh(1 - 2t_0)s}{\sinh s} - \frac{1}{s} \right].$$

After some simplification both sides of (11) are seen to depend only on $[t] = n$ and there is no difficulty in taking the limit $t \rightarrow n^+$, whence we have

$$L_n^{(\alpha)}(2x) = 2^{-\alpha} e^x \oint \frac{ds}{2\pi i} \frac{e^{(2n+\alpha+1)s - x \coth s}}{(\sinh s)^{\alpha+1}}. \quad (13)$$

In particular

$$L_n^{(\alpha)}(x/n) = 2^{-\alpha} e^{x/2n} \oint \frac{ds}{2\pi i} \frac{e^{2ns - x/2ns}}{s^{\alpha+1}} \phi(s), \quad (14)$$

where

$$\phi(s) = \exp\left\{(\alpha+1)s - \frac{x}{2n} \left(\coth s - \frac{1}{s}\right)\right\} (s \operatorname{csch} s)^{\alpha+1} \quad (15)$$

is analytic inside the contour and has the series expansion

$$\phi(s) = \sum_{k=0}^{\infty} a_k s^k, \quad a_0 = 1, \quad a_1 = \left(\alpha + 1 - \frac{x}{6n}\right), \dots \quad (16)$$

Following some slight simplification after inserting (16) into (14) we find the representation

$$L_n^{(\alpha)}(x/n) = (n/x)^\alpha e^{x/2n} \sum_{k=0}^{\infty} a_k (x/2n)^k \frac{\partial^k}{\partial x^k} \{x^{\alpha/2} J_\alpha(2\sqrt{x})\}, \quad (17)$$

where we have made use of the integral representation³

$$\frac{1}{2\pi i} \oint \exp\left[\frac{1}{2}Z(s - a^2 s^{-1})\right] \frac{ds}{s^{\alpha+1}} = a^{-\alpha} J_\alpha(aZ). \quad (18)$$

Equation (17) is easily sorted out into descending powers of n , whence we obtain the desired correction

$$\begin{aligned} n^{-\alpha} L_n^{(\alpha)}(x/n) &= x^{-\alpha/2} \left\{ J_\alpha(2\sqrt{x}) \right. \\ &\quad \left. + \frac{x}{2n} \left[\frac{\alpha(\alpha+1)}{x} J_\alpha(2\sqrt{x}) - J_{\alpha+2}(2\sqrt{x}) \right] \right\} \\ &\quad + O(n^{-2}). \end{aligned} \quad (19)$$

The only prior published results of this sort are due to Moecklin⁴ who treated the case $\alpha = 0$ by a complicated steepest descents calculation. Equation (19) reduces to his result in this case. Results in the same spirit have been obtained for the Hahn polynomials by Wilson.⁵ Another interesting asymptotic relation is

$$\lim_{\alpha \rightarrow \infty} \alpha^{-n/2} L_n^{(\alpha)}(\alpha^{1/2}x + \alpha) = (-1)^n 2^{-n/2} (n!)^{-1} H_n(2^{-1/2}x), \quad (20)$$

which is problem 81 in Ref. 6. This has been established by Calogero⁷ by considering the asymptotic behavior of the zeros of these functions.

We conclude by examining an approximation scheme that has frequently been used in studies concerning electron gases in an external magnetic field.⁸⁻¹⁰ For this purpose we consider the integral

$$H(a) = \int_{C-i\infty}^{C+i\infty} \frac{dp}{2\pi i} \frac{e^{ip}}{p} e^{-x \coth ap}, \quad C > 0, \quad (21)$$

where x , t , and a are positive parameters. This integral is typical of those which arise naturally in these investigations. As we have seen, the exact value is

$$H(a) = e^{-x} L_{[t/2a]}(2x). \quad (22)$$

The approximation scheme follows the procedure used above in closing and decomposing the contour into small circles surrounding the essential singularities $p_n = n\pi i/a$, $n = 0, \pm 1, \dots$. Now we allow the radii of these circles to shrink and argue that the hyperbolic cotangent can be replaced by the most singular term in its Laurent expansion about the pertinent singularity. This leads to the approximation

$$\begin{aligned} H(a) &\cong J_0(2\sqrt{xt/a}) - (2/\pi)(ax/t)^{1/2} J_1(2\sqrt{xt/a}) \\ &\quad \times \sum_{n=1}^{\infty} \sin(t\pi n/a)/n \\ &= J_0(2\sqrt{xt/a}) - (ax/t)^{1/2} J_1(2\sqrt{xt/a}) \\ &\quad \times (1 - 2\{t/2a\}). \end{aligned} \quad (23)$$

The curly brackets denote the fractional part (with the proviso that the second term of (23) vanishes if t/a is an integer). In practice it happens that x is proportional to a which is a small quantity and we can compare (23) with (19) for $\alpha = 0$. We see that for $a \ll t$ the approximation scheme gives the leading term correctly; this part is related to behavior in very weak magnetic fields. The second term in (23), which is oscillatory, relates to the so-called de Haas-van Alphen behavior; here, t/a is bounded and not less than unity. We set $t = 2ay$ and take the values $x = t = 1$; thus we wish to examine the accuracy of the approximation

$$e^{-1/2y} L_{[y]}(1/y) - J_0(2) \cong -(2y)^{-1} J_1(2)(1 - 2\{y\}). \quad (24)$$

Some comparative values are

y	LHS (24)	RHS (24)
1.2	-0.4345	-0.1442
4.8	0.0382	0.0360
10.1	-0.0222	-0.0228
20.3	-0.0055	-0.0057
50.6	0.0012	0.0011

It therefore appears that the approximation scheme is quite good in describing the oscillatory behavior of $H(a)$ up to very high field strengths.

ACKNOWLEDGMENT

The author gratefully acknowledges conversations with Professor O. Ruehr. This material is based on work supported by the National Science Foundation under Grant No. MCS78-04005.

¹A. Erdelyi *et al.*, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Vol. II, p. 189.

²Reference 1, p. 191.

³Reference 1, p. 7.

⁴E. Moecklin, *Comment. Math. Helv.* **7**, 24 (1934); O. Ruehr (unpublished) has calculated further terms in (19) directly from the defining series.

⁵M. W. Wilson, *SIAM J. Math. Anal.* **1**, 131 (1970).

⁶O. Szegő, *Orthogonal Polynomials* (American Mathematics Society, Providence, R. I., 1959).

⁷F. Calogero, *Lett. Nuovo Cimento* **23**, 100 (1978).

⁸M. L. Glasser, *Phys. Rev. A* **134**, 1296 (1964).

⁹N. J. Horing, *Ann. Phys. (N. Y.)* **31**, 1 (1965).

¹⁰M. L. Glasser, *Theoretical Chemistry: Advances and Perspectives*, edited by H. Eyring and D. Henderson (Academic, New York, 1976), Vol. II.

Perimeter expansion in the n -bug system and its relationship to stability

F. Behroozi

University of Wisconsin-Parkside, Kenosha, Wisconsin 53141

R. Gagnon

Hughes Aircraft Company, Fullerton, California 92634

(Received 17 June 1980; accepted for publication 21 November 1980)

We consider a system of n bugs located at $(x_1, y_1), \dots, (x_n, y_n)$, where bug i runs away from bug $i + 1$ with common speed v along the instantaneous line of sight. To close the cycle of flight, bug n runs away from bug 1. The computer simulation of this system indicates that random initial configurations evolve into stable regular center-symmetric patterns—all of which have a vertex angle of less than $\pi/2$. By utilizing the Lagrange multiplier method, we show that for these stable configurations the perimeter expansion rate \dot{P} is a local maximum. The most stable configuration has the smallest possible vertex angle and is associated with an absolute maximum for \dot{P} . The regular center-symmetric patterns with vertex angles greater than $\pi/2$ also have a stationary perimeter expansion rate. These are local minima rather than maxima, however, and belong to configurations which are unstable.

PACS numbers: 02.40. + m

I. INTRODUCTION

The chase problem, or cyclic pursuit, where a number of point bugs pursue one another in a cyclic pattern, has been of interest to mathematicians and “natural philosophers” for over a century.¹⁻³ The problem has appeared in the form of games, puzzles, research topics, and in a variety of other forms.⁴⁻¹¹ Here we address the problem of cyclic flight in an n -bug system confined to a plane. We define the problem as follows:

Given n bugs located at $(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)$, it is arranged for bug i to run away from bug $i + 1$ with a common speed v along the instantaneous line joining them. To close the cycle, bug n runs away from bug 1. Starting with n bugs placed at random initial positions, we wish to investigate the time evolution of this system.

In an earlier paper⁹ dealing with a computer assisted analysis of this system, we have reported that random initial configurations evolve into stable regular center-symmetric patterns, all of which have a vertex angle of less than $\pi/2$. In this paper we show that for these stable configurations the perimeter expansion rate \dot{P} is a local maximum, where the perimeter P of a bug configuration is the sum total of distances between bug i and bug $i + 1$. The regular center-symmetric configurations with vertex angles greater than $\pi/2$ are characterized by local minima rather than maxima for \dot{P} , and belong to configurations which are unstable.

We will be mainly concerned with configurations which maximize perimeter expansion for cyclic flight. However, we note that a configuration which maximizes perimeter expansion for cyclic flight will also maximize perimeter contraction for cyclic pursuit. For the sake of clarity and because there is a relationship between perimeter expansion and stability for cyclic flight, we shall restrict the discussion to cyclic flight. The relevance to cyclic pursuit, however, is obvious.

II. SYMMETRIC CONFIGURATIONS

By a configuration we mean the closed plane figure which results at a given time by connecting bug i to bug

$i + 1$. The bug positions and consequently the resulting configurations change with time. A “regular configuration” is a bug configuration with n equal sides and n equal vertex angles. The line segments joining bug i to bug $i + 1$, may cross or coincide with other lines as long as the resulting figure is equiangular and equilateral.

For n points, the first regular configuration is the familiar regular polygon of n sides with vertex angle $\phi = (n - 2)\pi/n$. The other possible regular configurations depend to a certain extent on whether n is odd or even.

When n is even, the number of possible regular configurations equals $n/2$, where the associated vertex angles ϕ , in the order of decreasing value, are given by $\phi = (n - 2)\pi/n, (n - 4)\pi/n, (n - 6)\pi/n, \dots, 2\pi/n, 0$. As already stated, the first such configuration is the regular convex polygon for which $\phi = (n - 2)\pi/n$; the others are star-shaped figures of progressively smaller vertex angles. The case when $\phi = 0$ needs special attention. Here the regular configuration consists of n superimposed sides resulting in a line configuration where the odd-numbered bugs are located at one end of the line and their mates (the even-numbered bugs) are at the other end. This regular line configuration occurs for all cases in which n is even when $\phi = 0$.

When n is odd, the possible number of regular configurations equals $(n - 1)/2$ and the vertex angles ϕ , in the order of decreasing value, are given by $\phi = (n - 2)\pi/n, (n - 4)\pi/n, (n - 6)\pi/n, \dots, 3\pi/n, \pi/n$. As before, $\phi = (n - 2)\pi/n$ corresponds to the vertex angle of the regular convex polygon, while the slimmest star-shaped figure is represented by $\phi = \pi/n$. Figures 1 and 2 illustrate the possible regular configurations for $n = 8$ and $n = 9$. In Fig. 1, note that among the four regular configurations for $n = 8$, one finds the square where each side is repeated twice, as well as the line configuration where one side is repeated eight times. In Fig. 2, one of the four configurations for $n = 9$ is an equilateral triangle with each side repeated three times.

The time evolution of regular configurations, apart from the question of stability, has been studied by many

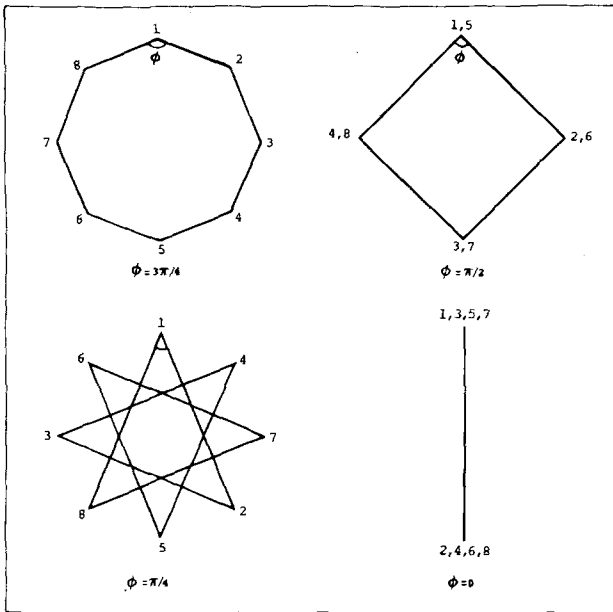


FIG. 1. The four regular configurations for $n = 8$. In the square configuration each side is repeated twice while in the line configuration one side is repeated eight times.

workers.^{4,7,11} Due to the inherent symmetry of the regular configurations, it is apparent that as the bugs recede from one another, these figures stay regular while the configurations expand and rotate in time.

We treat briefly the case of a regular n -gon¹¹ with the vertex angle ϕ as shown in Fig. 3. Using plane polar coordinates r and θ to represent the position of a representative bug, we note that

$$\dot{r} = v \cos(\phi/2), \quad (1)$$

$$\dot{\theta} = (v/r)\sin(\phi/2). \quad (2)$$

Here the dots represent time differentiation, and v is the

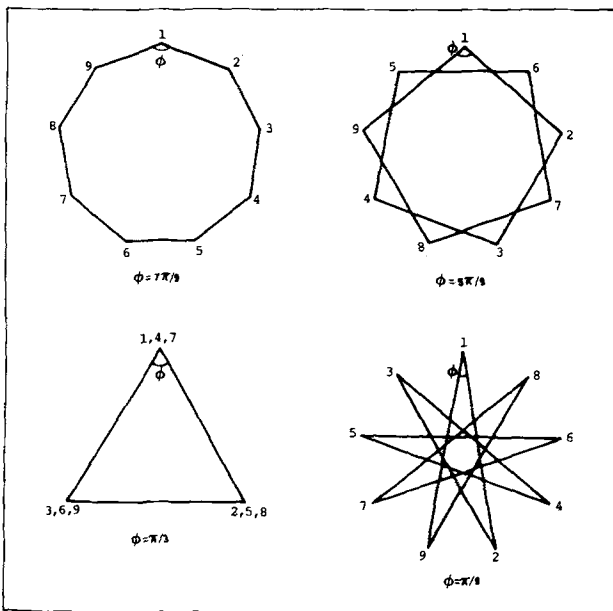


FIG. 2. The four regular configurations for $n = 9$. In the triangular configuration each side is repeated three times.

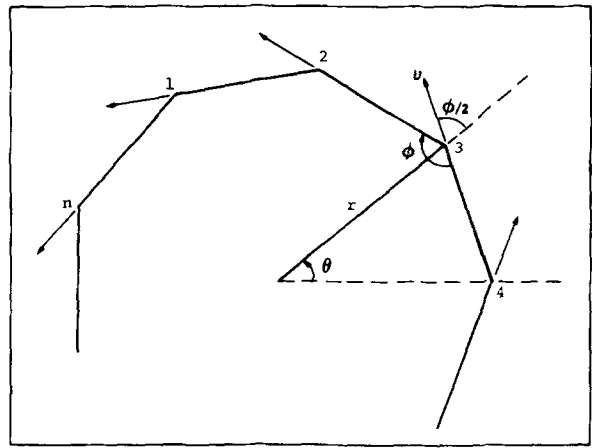


FIG. 3. A regular convex n -gon with a vertex angle $\phi = (n - 2)\pi/n$.

common speed of the bugs. Equation (1) indicates that r grows uniformly, whereas according to Eq. (2), as r increases, θ increases at an ever slower pace. Equations (1) and (2) can be combined to give $dr/d\theta = r \cot(\phi/2)$, whence

$$r = r_0 \exp[(\theta - \theta_0)\cot(\phi/2)]. \quad (3)$$

Substitution of Eq. (3) into Eq. (2) results in

$$\dot{\theta} = (v/r_0) \exp[-(\theta - \theta_0) \cos(\phi/2)].$$

Thus, apart from the question of stability, the evolution of regular configurations is relatively simple: the sides of the n -gon expand linearly with time while the configuration as a whole rotates with respect to the centroid at an ever slower pace. We note here that an observer located at the centroid of the system sees the bugs receding from him radially with speed $\dot{r} = v \cos(\phi/2)$. Consequently, in a regular configuration the smaller the vertex angle ϕ , the faster is the radial rate of expansion. As a matter of fact, the radial expansion rate is greatest when ϕ equals the minimum value. Consequently, when n is even, the line configuration with $\phi = 0$ is the most efficient way for the bugs to get away from one another. In this case, the relative distance between each bug pair, R , equals $2r$ and increases at the maximum rate of $\dot{R} = 2v$. When n is odd, we note that the slimmest star configuration with the vertex angle $\phi = \pi/n$ gives the largest expansion rate, where in this case, $\dot{R} = 2v \cos(\pi/n)$ with $\lim_{n \rightarrow \infty} \dot{R} = 2v$.

III. NONSYMMETRIC CONFIGURATIONS

Figure 4 shows bug i running away from bug $i + 1$ with speed v . For convenience we denote the instantaneous distance between the bug pair by l_i , and the vertex angle at bug i by ϕ_i . The equations of motion are given by

$$\dot{l}_i = v(1 + \cos\phi_{i+1}), \quad (4)$$

$$\dot{\phi}_i = (\sin\phi_{i+1})/l_i - (\sin\phi_i)/l_{i-1}. \quad (5)$$

Thus, there are $2n$ coupled first-order nonlinear differential equations to solve. Analytic solutions for such systems are hopelessly difficult; however, a computer can be used to investigate the general behavior of the equations and help point the direction of a suitable analytical approach.

The computer simulation of the problem is straightforward.⁹ A summary of the computer results for various initial

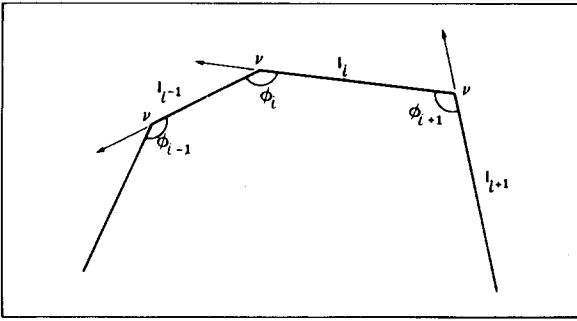


FIG. 4. A general bug configuration where bug i is running away from bug $i + 1$ with speed v .

positions and different numbers of bugs follow.

When $n = 3$, the final configuration is always an equilateral triangle. For $n = 4$, the computer takes a very long time before finally settling for a line. However, if the input configuration is nearly a perfect square or a perfect rhombus, the computer settles for a perfect square. For the case of $n = 5$, the final configuration is always a regular star with $\phi = \pi/5$. The regular pentagon with $\phi = 3\pi/5$ is not stable and a small perturbation is enough to send it to the star configuration.

For n greater than 5, when n is even, the final figures are the possible regular configurations with the vertex angle $\phi < \pi/2$ depending on initial conditions. In general, regular configurations with the smaller vertex angles are more frequently obtained as the final result, with the regular line configuration having $\phi = 0$ being the most frequent result. However, randomly selected input data always results in the line configuration.

When n is odd, one obtains different regular configurations depending on the initial conditions, but most often the final configuration is the slimmest star, corresponding to a vertex angle $\phi = \pi/n$. As mentioned before, randomly selected input data always produces the slimmest star as the final configuration. However, in all cases whether n is even or odd, the final configurations are regular and have a vertex angle $\phi < \pi/2$. For example, when $n = 9$, Fig. 2 shows four possible regular configurations starting with the polygon having the vertex angle $\phi = 7\pi/9$ followed by a first star with $\phi = 5\pi/9 > \pi/2$. The regular polygon and this first star are both unstable in the sense that a small perturbation sends them to the $\phi = \pi/9$ star configuration. In general we observe that for a given n , all regular configurations for which $\phi > \pi/2$ are unstable. Thus the regular polygons (except for $n = 2, 3$, and possibly 4) are all unstable. Furthermore, there is a preference for those regular configurations which have the smallest possible vertex angle.

IV. ANALYSIS

Figure 4 shows a general irregular n -bug system where bug i runs away from bug $i + 1$ with speed v . As stated before, the equations governing the evolution of the sides and angles are (4) and (5). To these can be added the constraint equation

$$\sum_{i=1}^n \phi_i = (n - 2)\pi, (n - 4)\pi, \dots$$

Whenever the summation sign appears henceforth, the index i will be understood to run from 1 to n .

We denote $\sum_i l_i = P$, where P is interpreted as the "perimeter" of the configuration and is a function of time. Hence,

$$\dot{P} = \sum_i \dot{l}_i = v(n + \sum_i \cos \phi_i). \quad (6)$$

It is apparent that \dot{P} is the expansion rate of the perimeter and that it achieves its maximum value when $\sum_i \cos \phi_i$ is a maximum. Our computer results indicate that random initial positions evolve into the regular configuration with the smallest possible vertex angle, for which $\dot{P} = 2vn$ when n is even, and $\dot{P} = [n + n \cos(\pi/n)]v$ when n is odd. Thus, in the most stable final configurations, the system attains the symmetric structure in which the distance between each bug pair increases at the maximum possible rate. In other words, the most stable configuration for a given bug system is associated with the maximum escape efficiency. We therefore enquire as to whether other stable configurations are characterized by local maxima for the function \dot{P} .

To identify the configurations for which \dot{P} is a maximum, we first note that in view of Eq. (6), to maximize \dot{P} one needs to maximize the function $f = \sum_i \cos \phi_i$ subject to the constraint condition

$\sum_i \phi_i = (n - 2)\pi, (n - 4)\pi, (n - 6)\pi, \dots$, which accommodates the various possible configurations for a given n .

To simplify our treatment, we will make the following observation. For a given n , it is possible to use only one constraint equation

$$\sum_i \phi_i = (n - 2)\pi \quad (7)$$

for all possible configurations, but with the understanding that some of the vertex angles may be counted as $\phi_i + 2\pi$ to satisfy Eq. (7). It is easy to see how this can be done if we consider the gradual evolution of a pentagon into a star as illustrated in Fig. 5. As shown in the illustration, the resulting star configuration of $n = 5$ with vertex angle $\phi = \pi/5$

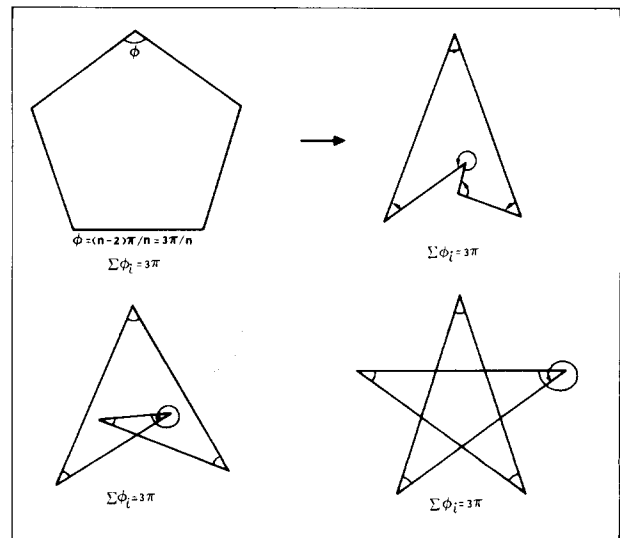


FIG. 5. The gradual evolution of a pentagon into a star.

can be considered to have four angles equal to $\pi/5$ and one equal to $\pi/5 + 2\pi$, thereby satisfying Eq. (7).

Our task is then to identify the configurations for which the function $f = \sum_i \cos\phi_i$ is a maximum under the constraint condition $\sum_i \phi_i = (n - 2)\pi$. To achieve this we use the method of Lagrange's undetermined multipliers.

Following the standard procedure, we let

$$f(\phi_1, \dots, \phi_n) = \sum_i \cos\phi_i + \lambda \left[\sum_i \phi_i - (n - 2)\pi \right],$$

where λ is the undetermined multiplier. The condition on ϕ_i 's so that f attains its extremum values is given by

$$\partial f / \partial \phi_i = -\sin\phi_i + \lambda = 0,$$

whence

$$\sin\phi_i = \lambda. \quad (7a)$$

Since $\phi_i = \sin^{-1}\lambda$ is a multiple-valued function, we cannot as yet conclude that condition (7a) specifies the regular configurations as the only ones associated with the extremum values of \dot{P} . Let us assume then that a general configuration satisfying condition (7a) will have

$$\begin{aligned} n_1 \text{ angles equal to the principal value } \phi_1 &= \sin^{-1}\lambda, \\ n_2 \text{ angles equal to } \phi_2 &= \pi - \phi_1 = \pi - \sin^{-1}\lambda, \\ n_3 \text{ angles equal to } \phi_3 &= 2\pi + \phi_1 = 2\pi + \sin^{-1}\lambda. \end{aligned}$$

Then the constraint condition as given by Eq. (7) requires that

$$n_1(\sin^{-1}\lambda) + n_2(\pi - \sin^{-1}\lambda) + n_3(2\pi + \sin^{-1}\lambda) = (n - 2)\pi,$$

resulting in

$$\lambda = \sin(n_1 - n_3 - 2)\pi / (n_1 - n_2 + n_3). \quad (8)$$

Assuming that $-\pi/2 \leq (\sin^{-1}\lambda) \leq \pi/2$, the solution to the extremum problem must be sought amongst triplets n_1, n_2 , and n_3 such that

$$n_1 + n_2 + n_3 = n \quad (8a)$$

and

$$-\frac{1}{2} \leq (n_1 - n_3 - 2) / (n_1 - n_2 + n_3) \leq \frac{1}{2},$$

with the understanding that $(n_1 + n_3)$ vertex angles have positive cosine values equal to $(1 - \lambda^2)^{1/2}$ and n_2 angles have negative cosine values equal to $-(1 - \lambda^2)^{1/2}$.

Therefore,

$$\begin{aligned} f &= \sum_i \cos\phi_i = (n_1 + n_3 - n_2)(1 - \lambda^2)^{1/2} \\ &= (n - 2n_2)(1 - \lambda^2)^{1/2}, \end{aligned}$$

which when combined with Eq. (6) results in

$$\dot{P} = nv + v(n - 2n_2)(1 - \lambda^2)^{1/2}. \quad (9)$$

To maximize \dot{P} for a given n , in view of Eq. (9), we must choose the smallest possible values for n_2 and λ . Clearly the absolute maximum for \dot{P} in Eq. (9) would be achieved by letting $\lambda = 0$ and $n_2 = 0$. This choice for λ and n_2 substituted into Eq. (8) leads to

$$n_1 - n_3 = 2,$$

and through condition (8a) we also have,

$$n_1 + n_3 = n,$$

which results in

$$\begin{aligned} n_1 &= n/2 + 1, & n_2 &= 0, & n_3 &= n/2 - 1, \\ \phi_{n_1} &= 0, & \phi_{n_2} &= \pi, & \phi_{n_3} &= 2\pi, \\ \dot{P} &= 2vn. \end{aligned}$$

This solution is admissible only when n is even and corresponds to a configuration in which $n/2 + 1$ angles equal zero and $n/2 - 1$ angles equal 2π . This is at once recognized as the line configuration where half of the bugs are on one end and their mates are at the other end. In this case, the expansion rate $\dot{P} = 2vn$ is an absolute maximum.

We now ask if there are other possible choices of λ and n_2 such that \dot{P} attains the next highest value. Equation (9) suggests that the next highest value for \dot{P} may be achieved either when

$$(a) \lambda = 0, \quad n_2 = 1$$

or

$$(b) \lambda = \sin(\pi/n), \quad n_2 = 0.$$

These choices for λ and n_2 are dictated by the fact that the smallest possible value next to zero for n_2 is unity, and that for λ is $\sin(\pi/n)$ [see Eq. (8)]. We proceed to examine each case in turn to determine which one gives the largest \dot{P} .

Choice (a), when substituted in Eq. (8) in view of condition (8a), leads to

$$\begin{aligned} n_1 - n_3 &= 2, \\ n_1 + n_3 &= n - 1, \end{aligned}$$

resulting in

$$\begin{aligned} n_1 &= (n + 1)/2, & n_2 &= 1, & n_3 &= (n - 3)/2, \\ \phi_{n_1} &= 0, & \phi_{n_2} &= \pi, & \phi_{n_3} &= 2\pi, \\ \dot{P} &= 2(n - 1)v. \end{aligned}$$

For even n , this solution is not admissible as it gives half integer values for n_1 and n_3 . However, for odd n , this solution gives a most interesting configuration in which all the bugs lie along the same line where one vertex angle is π , and the other angles are either 0 or 2π to satisfy the constraint condition $\sum_i \phi_i = (n - 2)\pi$. As shown in Fig. 6, these odd- n configurations are mimicking the behavior of the most stable even- n configurations! However, note that here according to Eq. (9), $\dot{P} = 2(n - 1)v$, which is due to the fact that always one bug remains at the same distance from its mate.

For choice (b), since $\lambda = \sin\pi/n$, Eq. (8) leads to

$$n_1 - n_3 - 2 = 1,$$

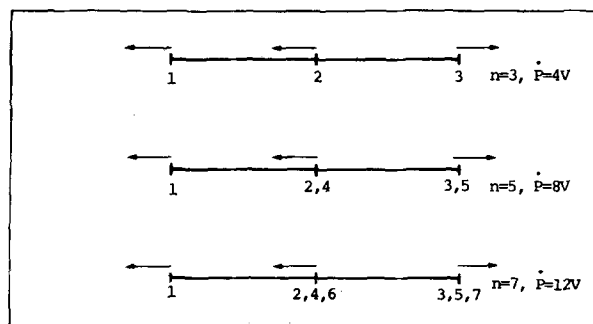


FIG. 6. Linear configurations for $n = 3, 5, 7$ which mimic the regular line configuration for even n .

and

$$n_1 + n_3 = n,$$

whence

$$\begin{aligned} n_1 &= (n + 3)/2, & n_2 &= 0, & n_3 &= (n - 3)/2, \\ \phi_{n_1} &= \pi/2, & \phi_{n_2} &= \pi - \pi/n, & \phi_{n_3} &= \pi/n + 2\pi, \\ \dot{P} &= nv[1 + \cos(\pi/n)]. \end{aligned}$$

These values for n_1 and n_3 are permissible for odd values of n only, where in a configuration we have $(n + 3)/2 = n_1$ of vertex angles equal to π/n , and $(n - 3)/2 = n_3$ of vertex angles equal to $2\pi + \pi/n$. This is precisely the slimmest star configuration for odd n . Let us examine a few representative cases:

$$\begin{aligned} n = 3: & \quad n_1 = 3, \quad n_2 = n_3 = 0, \\ & \quad \phi_{n_1} = \pi/3, \\ & \quad \dot{P} = 3[1 + \cos(\pi/3)]v = 4.5v. \end{aligned}$$

$$\begin{aligned} n = 5: & \quad n_1 = 4, \quad n_2 = 0, \quad n_3 = 1, \\ & \quad \phi_{n_1} = \pi/5, \quad \phi_{n_3} = 2\pi + \pi/5, \\ & \quad \dot{P} = 5[1 + \cos(\pi/5)]v = 9v. \end{aligned}$$

$$\begin{aligned} n = 7: & \quad n_1 = 5, \quad n_2 = 0, \quad n_3 = 2, \\ & \quad \phi_{n_1} = \pi/7, \quad \phi_{n_3} = 2\pi + \pi/7, \\ & \quad \dot{P} = 7[1 + \cos(\pi/7)]v = 13.3v. \end{aligned}$$

As is clearly evident, for odd n case (b), namely $\lambda = \sin \pi/n, n_2 = 0$ is the winner, a fact which corresponds completely with the computer results when n is odd. The general expression for \dot{P} in this case is $\dot{P} = nv(1 + \cos \pi/n) > 2v(n - 1)$; i.e., case (b) results in the absolute maximum for \dot{P} when n is odd.

We now show that for even n , next to the line configuration, the slimmest star with $\phi = 2\pi/n$ has the largest \dot{P} and therefore presumably is the next most stable. We identify three cases where λ and n_2 are, excepting the cases considered so far, the smallest possible.

$$(a) \lambda = \sin[\pi/(n - 2)] \quad \text{and} \quad n_2 = 1.$$

This choice leads to

$$n_1 - n_3 - 2 = 1$$

and

$$n_1 + n_3 + 1 = n,$$

whence

$$\begin{aligned} n_1 &= (n + 2)/2, \quad n_2 = 1, \quad n_3 = (n - 4)/2, \\ \phi_{n_1} &= \pi/(n - 2), \quad \phi_{n_2} = \pi - \pi/(n - 2), \quad \phi_{n_3} = 2\pi + \pi/(n - 2), \\ \dot{P} &= nv + v(n - 2)\cos[\pi/(n - 2)]. \end{aligned}$$

These solutions are permissible for even n only. A few representative examples follow:

$n = 4:$	square	$n_1 = 4,$	$n_3 = 0,$	$\phi_{n_1} = \pi/2,$	$\dot{P} = 4v.$	
$n = 6:$	star	$n_1 = 5,$	$n_3 = 1,$	$\phi_{n_1} = \pi/3,$	$\phi_{n_3} = \pi/3 + 2\pi,$	$\dot{P} = 9v.$
$n = 8:$	star	$n_1 = 6,$	$n_3 = 2,$	$\phi_{n_1} = \pi/4,$	$\phi_{n_3} = \pi/4 + 2\pi,$	$\dot{P} = 8(1 + \sqrt{2}/2)v.$
$n = 10:$	star	$n_1 = 7,$	$n_3 = 3,$	$\phi_{n_1} = \pi/5,$	$\phi_{n_3} = \pi/5 + 2\pi,$	$\dot{P} = 18v.$

These are the stars with the smallest vertex angles for even n .

$$(c) \lambda = 0 \quad \text{and} \quad n_2 = 2.$$

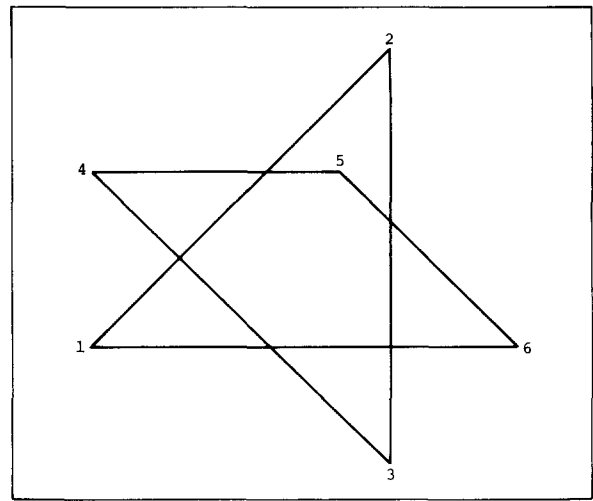


FIG. 7. An example of a configuration for $n = 6$ which resembles a regular star with one side askew.

$$\begin{aligned} n = 4: & \quad n_1 = 3, \quad n_2 = 1, \quad n_3 = 0, \\ & \quad \phi_{n_1} = \pi/2, \quad \phi_{n_2} = \pi/2, \\ & \quad \dot{P} = 4v. \end{aligned}$$

$$\begin{aligned} n = 6: & \quad n_1 = 4, \quad n_2 = 1, \quad n_3 = 1, \\ & \quad \phi_{n_1} = \pi/4, \quad \phi_{n_2} = 3\pi/4, \quad \phi_{n_3} = 2\pi + \pi/4, \\ & \quad \dot{P} = (6 + 2\sqrt{2})v. \end{aligned}$$

$$\begin{aligned} n = 8: & \quad n_1 = 5, \quad n_2 = 1, \quad n_3 = 2, \\ & \quad \phi_{n_1} = \pi/6, \quad \phi_{n_2} = 5\pi/6, \quad \phi_{n_3} = 2\pi + \pi/6, \\ & \quad \dot{P} = 11v. \end{aligned}$$

As is evident, these configurations are mimicking a star with one side askew! An example for $n = 6$ is shown in Fig. 7.

$$(b) \lambda = \sin 2\pi/n \quad \text{and} \quad n_2 = 0$$

This choice leads to

$$n_1 - n_3 = 4$$

and

$$n_1 + n_3 = n,$$

whence

$$\begin{aligned} n_1 &= (n + 4)/2, \quad n_2 = 0, \quad n_3 = (n - 4)/2, \\ \phi_{n_1} &= 2\pi/n, \quad \phi_{n_3} = 2\pi/n + 2\pi, \\ \dot{P} &= nv + v[n\cos(2\pi/n)]. \end{aligned}$$

Some representative solutions are:

This choice lead to

$$n_1 = n_3 = 2 \text{ and } n_1 + n_3 = n - 2,$$

whence

$$\begin{aligned} n_1 &= n/2, & n_2 &= 2, & n_3 &= (n - 4)/2, \\ \phi_{n_1} &= 0, & \phi_{n_2} &= \pi, & \phi_{n_3} &= 2\pi, \\ \dot{P} &= nv + v(n - 4). \end{aligned}$$

Some representative solutions are:

$$\begin{aligned} n = 4: & \text{ line } n_1 = 2, n_2 = 2, n_3 = 0, \phi_{n_1} = 0, \phi_{n_2} = \pi, \dot{P} = 4v. \\ n = 6: & \text{ line } n_1 = 3, n_2 = 2, n_3 = 1, \phi_{n_1} = 0, \phi_{n_2} = \pi, \phi_{n_3} = 2\pi, \dot{P} = 8v. \\ n = 8: & \text{ line } n_1 = 4, n_2 = 2, n_3 = 2, \phi_{n_1} = 0, \phi_{n_2} = \pi, \phi_{n_3} = 2\pi, \dot{P} = 12v. \end{aligned}$$

As shown in Fig. 8, these configurations mimic the regular line configuration.

A comparison of \dot{P} for the three cases considered shows that case (b) provides the next highest perimeter expansion rate for even n after the regular line configuration. That is, for even n , \dot{P} attains its second highest value for the star configuration of vertex angle $2\pi/n$. A similar analysis indicates that among regular configurations \dot{P} is the third highest when the vertex angle $\phi = 4\pi/n$; \dot{P} is fourth highest when $\phi = 6\pi/n$, etc. Similarly, when n is odd, the second highest value of \dot{P} occurs for the star configuration of vertex angle $\phi = 3\pi/n$; the third highest \dot{P} occurs for the configuration of vertex angle $\phi = 5\pi/n$, and so on. As mentioned earlier our computer simulation of this system indicates that stable configurations are all regular, and therefore are also associated with the local maxima of the perimeter expansion rate \dot{P} . However, we must still explain the observed computer result that for $\phi_i > \pi/2$; the regular configurations are not stable.

To investigate this question, we allow the angles ϕ_i to be a function of some parameter x . Differentiating Eq. (6) with respect to x , we get

$$d\dot{P}/dx = -v \sum_i \sin\phi_i \, d\phi_i/dx.$$

A second differentiation results in

$$d^2\dot{P}/dx^2 = -v \left(\sum_i \sin\phi_i \, d^2\phi_i/dx^2 + \sum_i \cos\phi_i (d\phi_i/dx)^2 \right). \quad (10)$$

However, when the configuration is in the extremum state, according to condition (7a) $\sin\phi_i = \lambda$, and since $\sum_i \phi_i = (n - 2)\pi$, the first term of Eq. (10) is zero and the equation yields:

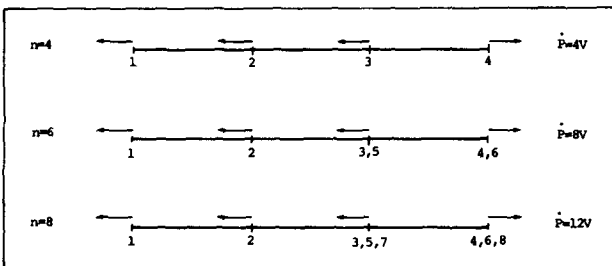


FIG. 8: Linear configurations for $n = 4, 6, 8$ which mimic the regular line configuration for even n .

$$d^2\dot{P}/dx^2 = -v |\cos\phi_i| \left(\sum_{n_1, n_3} (d\phi_i/dx)^2 - \sum_{n_2} (d\phi_i/dx)^2 \right). \quad (11)$$

The perimeter expansion rate will be a maximum if Eq. (11) is negative and a minimum if Eq. (11) is positive. In the case of a regular configuration with the vertex angles less than $\pi/2$, $n_2 = 0$ and Eq. (11) is negative. Hence the perimeter expansion rate is a local maximum. On the other hand, a regular configuration with the vertex angle greater than $\pi/2$ has $n_1 = n_3 = 0$, so that Eq. (11) is positive and the perimeter expansion rate is a local minimum. These regular configurations we have shown to be unstable.⁹ We may therefore assert that every stable configuration is a regular configuration for which the perimeter expansion rate is a local maximum, and every regular configuration for which the perimeter expansion rate is a local minimum is unstable.

An interesting situation arises when neither term in the brackets of Eq. (11) vanishes. In such a case, the configuration is not regular, and it may be possible to vary the angles indexed n_1 and n_3 so the expansion appears to be a maximum. At the same time (by a different choice of the parametric dependence of the vertex angles on the variable x), it may be possible to vary the angles indexed n_2 so that the expansion appears to be a minimum. In such a case, the stationary condition for \dot{P} corresponds to a saddle point rather than a true maximum or minimum.

An example for which this unusual state of affairs holds true is a rhombus or a parallelogram. As shown in Fig. 9, for a rhombus $n_1 = n_2 = 2$, and $\phi_{n_1} = \theta$, $\phi_{n_2} = \pi - \theta$. Suppose that the dependence of the vertex angles on the parameter x

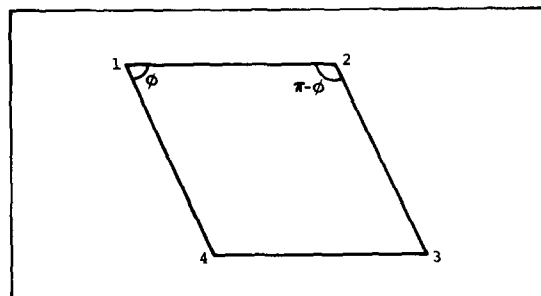


FIG. 9. The perimeter expansion rate is neither a maximum nor a minimum for this rhombus.

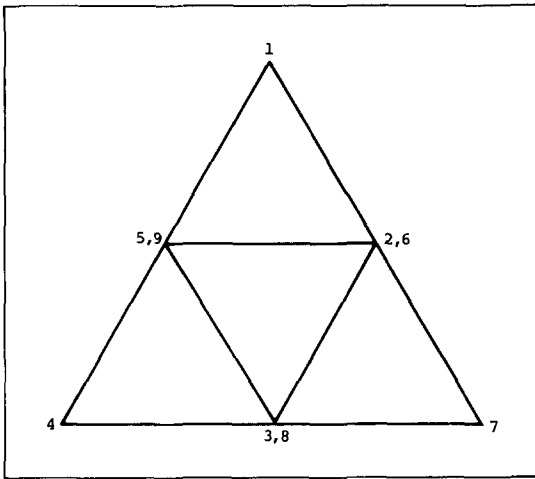


FIG. 10. The perimeter expansion rate is stationary but is neither a maximum nor a minimum for this regular-looking configuration.

is such that $\phi_1 = \theta - x$, $\phi_3 = \theta + x$, $\phi_2 = \phi_4 = \pi - \theta$. Then at $x = 0$, $d^2P/dx^2 = -2\nu\cos\theta$, so that the perimeter expansion is maximized. On the other hand, if we put $\phi_1 = \phi_3 = \theta$, $\phi_2 = \pi - \theta - x$, $\phi_4 = \pi - \theta + x$, we get $d^2P/dx^2 = 2\nu\cos\theta$,

and the perimeter expansion is minimized. Hence the rhombus is a saddle point rather than a true maximum or minimum for perimeter expansion. A more complicated example of a saddle-point configuration is the nonagon shown in Fig. 10. Three of the angles are equal to 60° , and six of the angles are equal to 120° . The sides are all equal in length.

¹E. Lucas, *Nouvo Corresp. Math.* **3**, 175 (1877).

²H. Brocard, *Nouvo Corresp. Math.* **3**, 280 (1877).

³H. Brocard, *Nouvo Corresp. Math.* **6**, 211 (1880).

⁴In a series of four papers in *Scripta Math.*, Arthur Bernhart gives an excellent and extensively referenced review of pursuit problems. These articles are listed here in chronological order: (a) *Scripta Math.* **20**, 125 (1954); (b) **23**, 49 (1957); (c) **24**, 23 (1959); (d) **24**, 189 (1959). The third paper in the list is of particular interest to readers of the present article.

⁵M. Gardner, *Sci. Am.* **213**, 101 (1965).

⁶H. Steinhaus, *Mathematical Snapshots* (Oxford University, New York, 1969), p. 136.

⁷A. Watton and D. W. Kydon, *Am. J. Phys.* **37**, 220 (1969).

⁸M. S. Klamkin and D. J. Newman, *Amer. Math. Monthly* **78**, 631 (1971).

⁹F. Behroozi and R. Gagnon, *Amer. Math. Monthly* **82**, 804 (1975).

¹⁰F. Behroozi and R. Gagnon, *Am. J. Phys.* **43**, 237 (1979).

¹¹F. Behroozi and R. Gagnon, *J. Math. Phys.* **20**, 2212 (1979).

On the Killing surface–event horizon relation

J. P. Krisch

Department of Physics, University of Michigan, Ann Arbor, Michigan 48109
(Received 27 August 1979; accepted for publication 3 September 1980)

A projective transformation on the scalar norm and twist of a timelike Killing vector can be used to generate new space-times. The effect of the transformation on the new Killing surface and its relation to the local event horizon is discussed. It is shown that the Geroch transformation will only connect spaces where this relation is the same.

PACS numbers: 02.40. + m, 04.20. – q

I. INTRODUCTION

The problem of finding single solutions to the Einstein field equations has received much attention. Recently the problem has broadened in scope with the introduction and development of methods for generating new families of vacuum solutions from a single known solution.^{1–5} Given a metric with a timelike Killing vector ξ^a , the technique gives a new metric g'_{ab} with the same Killing vector. As described by Geroch,^{4,5} the new metric is generated from the base metric by projective transformations on the scalar norm, λ , and scalar twist, ω , of the Killing vector, where

$$\begin{aligned} \lambda &= \xi^a \xi_a, \\ \omega_a &= \epsilon_{abcd} \xi^b \nabla^c \xi^d = D_a(\omega). \end{aligned} \quad (1)$$

The transformations are performed in the three-dimensional manifold defined by the Killing trajectories. D_a is the covariant derivative in this space.

While the solution sets are relatively easy to generate, their interpretation is more difficult. Many applications^{6–9} have concentrated on the scalars associated with a given metric, for example the multipole structure of potentials defined on the transformed space-time. In some cases, comparing this multipole structure at infinity with the structure of Newtonian potentials can provide insight into the new metric. The use of scalar potentials to interpret the new space-time is a clear first step since the transformation itself is a simple rotation of potential functions in the three-dimensional trajectory space, the potentials acting as homogeneous coordinates for the norm and twist.

The vector norm and twist provide another approach to gaining information about the nature of the transformed space-times. The vector twist is given by (1). The vector norm is defined by

$$n_a = \frac{1}{2}(\lambda)_{;a}. \quad (2)$$

Since the structure of the Killing surface $\lambda = \text{const}$ is determined by these vectors, the difference between the transformed and base surfaces can be studied. The Killing surface $\lambda = 0$, in some spaces, will coincide with the event horizon. By examining the changes in the Killing surface, one can see if this relation is maintained under the transformation. If it is, this will provide a strong limitation on the spaces that are bridg-able by the Geroch transformation.

In this note we discuss the effects of Geroch's transformation on the vector norm and twist and, through

them, on the Killing surface-horizon relationship. The discussion is carried out in terms of the Frenet–Serret formalism for both null and non-null Killing vectors. We briefly review the single Killing vector Geroch transformation in the next section. In the third part the transformation is applied to the norm and twist in a Frenet basis. In the last section we discuss the structure of the Killing surface. We show the Geroch transformation takes null-geodesic Killing surfaces into similar surfaces in the transformed space-time but that it will not produce coincident event-Killing surfaces from more general spaces.

II. THE TRANSFORMATION

Start with a vacuum solution g_{ab} possessing a single timelike Killing vector ξ^a . The norm λ and twist ω_a of the Killing vector are given by (1). The solution g_{ab} is described by a set of equations on a four-dimensional space G : g_{ab} . Geroch⁴ has shown that g_{ab} is also described by a set of equations written on the three-dimensional manifold, H : h_{ab} of Killing trajectories

$$\begin{aligned} \tilde{h}_{ab} &= -2(\tau - \tau^*)^{-2} = -2(\tau - \tau^*)^{-2}(\tilde{D}_a \tau \tilde{D}_b \tau), \\ \tilde{D}^2 \tau &= 2(\tau - \tau^*)^{-1}(\tilde{D}_a \tau)(\tilde{D}_b \tau) \tilde{h}^{ab}, \end{aligned} \quad (3)$$

where $\tau = \omega + i\lambda$ and $h_{ab} = \tilde{h}_{ab}/\lambda$ is given by

$$h_{ab} = g_{ab} - \xi_a \xi_b / \lambda. \quad (4)$$

\tilde{D} is the covariant derivation with respect to \tilde{h}_{ab} .

To generate a new metric g'_{ab} from g_{ab} , one goes to \tilde{H} and looks for a new solution, τ' , of (2) subject to the condition $\tilde{h}'_{ab} = \tilde{h}_{ab}$. The only such solution is

$$\tau' = (a\tau + b)/(c\tau + d), \quad (5)$$

which Geroch writes as

$$\tau' = (\cos(\gamma)\tau + \sin(\gamma))/(-\sin(\gamma)\tau + \cos(\gamma)). \quad (6)$$

One may show that the transformation is equivalent to the potential rotation.⁶

$$\begin{aligned} \phi'_J &= \phi_J \cos(2\gamma) - \phi_M \sin(2\gamma), \\ \phi'_M &= \phi_J \sin(2\gamma) + \phi_M \cos(2\gamma), \end{aligned} \quad (7)$$

where

$$\begin{aligned} \phi_J &= \omega/2\lambda, \\ \phi_M &= (\lambda^2 + \omega^2 - 1)/4\lambda \end{aligned} \quad (8)$$

are related to the Newtonian mass and angular momentum potentials.

The metric g'_{ab} corresponding to the new norm and twist is found by inverting (1) for ω' to find $\nabla_{[a}\xi'_{b]}$.

One finds

$$\xi'_a = \xi_a + \lambda' \alpha_a \sin 2\gamma - \lambda' \beta_a \sin^2 \gamma, \quad (9)$$

with α_a and β_a given by

$$\begin{aligned} \nabla_{[a}\alpha_{b]} &= \frac{1}{2}\epsilon_{abcd}\nabla^c\xi^d, \quad \xi^a\alpha_a = \omega, \\ \nabla_{[a}\beta_{b]} &= 2\lambda\nabla_a\xi_b + \omega\epsilon_{abcd}, \quad \xi^a\beta_a = \omega^2 + \lambda^2 - 1; \end{aligned} \quad (10)$$

using $\tilde{h}'_{ab} = \tilde{h}_{ab}$ one obtains

$$\lambda'g'_{ab} - \xi'_a\xi'_b = \lambda g_{ab} - \xi_a\xi_b. \quad (11)$$

The transformation to the space H : h_{ab} is, strictly speaking, defined only for nonzero λ .⁴ In order to extend the Geroch transformation to the case $\lambda = 0$, we perform the transformation for nonzero λ and then take (6) to define the transformation in the limit $\lambda, \lambda' \rightarrow 0, \lambda'/\lambda \neq 0$. The new null Killing vector becomes

$$\xi'_a = \xi_a + \tilde{\alpha}_a \sin 2\gamma - \tilde{\beta}_a \sin^2 \gamma, \quad (12)$$

with

$$\xi'_a\xi'^a = \xi^a\xi_a = \xi^a\tilde{\alpha}_a = \xi^a\tilde{\beta}_a = 0; \quad (13)$$

$\tilde{\alpha}_a$ and $\tilde{\beta}_a$ are either zero or null. If they are zero then $\xi'_a = \xi_a$. If they are null then by (13) they can be written $\tilde{\alpha}_a = h_1\xi_a, \tilde{\beta}_a = h_2\xi_a, h_1, h_2$ scalar functions. In the $\lambda, \lambda' = 0$ limit one can write

$$\nabla_a\tilde{\alpha}_b = 2n_a\tilde{\alpha}_b, \quad (14)$$

and similarly for $\tilde{\beta}_a$. If $\tilde{\alpha}_b$ is null this becomes

$$\xi_b\nabla_a h_1 - h_1\nabla_b\xi_a = 2n_a\tilde{\alpha}_b, \quad (15)$$

by Killing's equation. Multiplying by ξ^a we have

$$\begin{aligned} \xi_b\xi^a\nabla_a h_1 &= h_1\xi^a\nabla_b\xi_a, \\ \xi_b\xi^a\nabla_a h_1 &= 2h_1n_b. \end{aligned} \quad (16)$$

If n_b is not null this gives $h_1 = 0$ and similarly for h_2 so again $\xi'_a = \xi_a$. If n_b is null we can only say $\xi'_a = h_3\xi_a, h_3$, a scalar function.

III. THE EFFECT OF THE TRANSFORMATION

A. Frenet-Serret formalism

Before finding the explicit effect of the transformation it is useful to write down the Frenet formalism that will be needed. There are three separate Frenet tetrads to consider. The first is the ordinary Frenet tetrad,^{10,11} valid for non-null tangents. This tetrad is not useful in discussing a Killing surface-horizon coincidence. We find the effect of the transformation on the Frenet parameters of this tetrad in order to demonstrate the similarity of the Geroch transformation and duality rotations.¹¹ The second tetrad^{12,13} is used to discuss null, nongeodesic Killing vectors. This tetrad will be needed to discuss spaces where the surface $\lambda = 0$ is not coincident with the local event horizon. The last tetrad is valid for null-geodesic tangents and will be used to discuss space-times, where the $\lambda = 0$ surface coincides with the local event horizon. It is necessary to consider three separate tetrads since in each case the trajectory parameter is different.

The first tetrad consists of the standard set of Frenet

vectors $e^a_{(0)}$, the timelike unit tangent, $e^a_{(1)}$, the spacelike normal and $e^a_{(2)}, e^a_{(3)}$ the spacelike binormals. The tetrad satisfies the Frenet-Serret equations

$$\begin{bmatrix} \dot{e}^a_{(0)} \\ \dot{e}^a_{(1)} \\ \dot{e}^a_{(2)} \\ \dot{e}^a_{(3)} \end{bmatrix} = \begin{bmatrix} 0 & \kappa & 0 & 0 \\ \kappa & 0 & \tau_1 & 0 \\ 0 & -\tau_1 & 0 & \tau_2 \\ 0 & 0 & -\tau_2 & 0 \end{bmatrix} \begin{bmatrix} e^a_{(0)} \\ e^a_{(1)} \\ e^a_{(2)} \\ e^a_{(3)} \end{bmatrix}. \quad (17)$$

Dot denotes absolute differentiation. κ is the curvature and τ_1, τ_2 the first and second torsions, respectively. In terms of these vectors we have¹¹

$$\begin{aligned} \xi^a &= \sqrt{\lambda} e^a_{(0)}, \quad \lambda \neq 0, \\ n^a &= -\lambda \kappa e^a_{(1)}, \quad n^a n_a \neq 0, \\ \omega^a &= \lambda(\tau_1 e^a_{(3)} + \tau_2 e^a_{(1)}). \end{aligned} \quad (18)$$

The second tetrad consists of two null vectors ξ^a and B^a and two spacelike vectors A^a and C^a , orthogonal to ξ^a and B^a . ξ^a is identified with the tangent. We have $\xi^a B_a = 1 = -A^a A_a = -C^a C_a$. Defining $\kappa_1 = (-\xi^a \xi_a)^{1/2}$, $\kappa_2 = (1/\kappa_1^3)[\kappa_1^2 + \xi^a \xi_a]$, $\kappa_3 = (1/\kappa_1^5)\epsilon^{abcd}\xi_a \xi_b \xi_c \xi_d$ we can write the Frenet-Serret equations for this tetrad as

$$\begin{bmatrix} \dot{\xi}^a \\ \dot{A}^a \\ \dot{B}^a \\ \dot{C}^a \end{bmatrix} = \begin{bmatrix} 0 & \kappa_1 & 0 & 0 \\ \kappa_2 & 0 & \kappa_1 & 0 \\ 0 & \kappa_2 & 0 & \kappa_3 \\ \kappa_3 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \xi^a \\ A^a \\ B^a \\ C^a \end{bmatrix}. \quad (19)$$

In terms of these vectors we have¹⁴

$$\begin{aligned} \xi^a &= \xi^a, \quad \lambda = 0, \\ n^a &= \kappa_1 A^a, \quad n^a n_a \neq 0, \\ \omega^a &= \kappa_1 C^a. \end{aligned} \quad (20)$$

The Frenet parametrization makes it obvious there is no overlap between the norm and twist on $\lambda = 0$. For $\lambda = 0, n^a n_a = 0$, the Killing surface is null and geodesic so a single null vector L^a suffices to parametrize all the vectors. We have^{14,15}

$$\begin{aligned} \xi^a &= L^a, \quad \lambda = 0, \\ n^a &= \epsilon L^a, \quad n^a n_a = 0, \\ \omega^a &= \delta L^a \text{ or } 0, \end{aligned} \quad (21)$$

with ϵ, δ scalar functions.

B. The transformation

1. $\lambda \neq 0, n^a n_a \neq 0$

Consider the hypersurface $\xi^a \xi_a = \lambda = \text{const}$ in G . The normal to this surface is $n_a = \frac{1}{2}(\lambda)_{,a}$ and the vector twist is $\omega_a = (\omega)_{,a}$. These vectors can be transformed⁴ to the three-dimensional trajectory space H giving $n_a = \frac{1}{2}D_a(\lambda)$ and $\omega_a = D_a(\omega)$. In the space \tilde{H} we have

$$\omega' + i\lambda' = (\cos\gamma(\omega + i\lambda) + \sin\gamma)/(-\sin\gamma(\omega + i\lambda) + \cos\gamma), \quad (22)$$

with $\tilde{h}'_{ab} = \tilde{h}_{ab}$. Taking covariant derivatives in \tilde{H} and transforming to H using $\tilde{D}_a = \lambda D_a$ one obtains

$$\begin{pmatrix} \omega'_a \\ n'_a \end{pmatrix} = \begin{pmatrix} \cos\alpha & +\sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} \omega_a \\ n_a \end{pmatrix} \quad (23)$$

with

$$\begin{aligned} \cos(\alpha) &= -\omega \sin 2\theta + \cos 2\theta + \omega' \sin 2\theta - \omega \omega' (1 - \cos 2\theta) \\ &= 1 - \lambda \lambda' (1 - \cos 2\theta), \\ \sin(\alpha) &= \lambda' \omega (1 - \cos 2\theta) - \lambda' \sin 2\theta \\ &= -\lambda \sin 2\theta - \omega' \lambda (1 - \cos 2\theta), \end{aligned} \quad (24)$$

for $\lambda, \lambda' \neq 0$. Substituting, Eq. (20) becomes

$$\begin{aligned} \lambda' \kappa' e'_{a(1)} &= \lambda e_{a(1)} (\kappa \cos \alpha + \tau_2 \sin \alpha) + \lambda \tau_1 \sin \alpha e_{a(3)}, \\ \lambda' \tau_1' e'_{a(3)} + \lambda' \tau_2' e'_{a(1)} &= \lambda e_{a(3)} \tau_1 \cos \alpha + \lambda e_{a(1)} (\tau_2 \cos \alpha - \kappa \sin \alpha), \end{aligned} \quad (25)$$

multiplying, using $h'^{ab} = (\lambda'/\lambda) h^{ab}$ we obtain

$$\begin{aligned} \lambda' (\kappa'^2 + \tau_1'^2 + \tau_2'^2) &= \lambda (\kappa^2 + \tau_1^2 + \tau_2^2), \\ \lambda' \begin{bmatrix} \tau_1'^2 + \tau_2'^2 - \kappa'^2 \\ 2\tau_2' \kappa' \end{bmatrix} &= \lambda \begin{bmatrix} \cos 2\alpha - \sin 2\alpha \\ \sin 2\alpha \cos 2\alpha \end{bmatrix} \begin{bmatrix} \tau_1^2 + \tau_2^2 - \kappa^2 \\ 2\tau_2 \kappa \end{bmatrix}. \end{aligned} \quad (26)$$

2. $\lambda = 0, n^a n_a \neq 0$

In the limit $\lambda, \lambda' \rightarrow 0$ we obtain from (6) in G'

$$\omega'_a = \omega_a. \quad (27)$$

By (20) we have

$$\kappa_1' C'_a = C_a \kappa_1. \quad (28)$$

By the discussion following (16) we know $\xi'_a = \xi_a$. For this case we can generate the normals by absolute derivatives. From (19) we get

$$n'_a = n_a. \quad (29)$$

This result also follows simply, but less rigorously, from (24) in the $\lambda', \lambda = 0$ limit.

For the curvatures we obtain by absolute differentiation and squaring

$$\begin{aligned} (\kappa_1')^2 &= (\kappa_1)^2, \\ (\kappa_2')^2 &= (\kappa_2)^2, \\ (\kappa_3')^2 &= (\kappa_3)^2. \end{aligned} \quad (30)$$

3. $\lambda = 0, n^a n_a = 0$

The discussion following (16) and Eqs. (21) indicate all the vectors remain null, or zero.

IV. DISCUSSION

The formalism we have developed allows a discussion of both null and non-null Killing vectors and general, null and null-geodesic Killing surfaces. The non-null Killing vector formalism can be used to demonstrate the similarity between the Geroch transformation and duality rotation.

Honig and Schucking¹¹ have shown that electromagnetic fields can be described by

$$\begin{aligned} (e/mc^2)H^a &= \tau_1 e^a_{(3)} + \tau_2 e^a_{(1)}, \\ (e/mc^2)E^a &= \kappa e^a_{(1)}. \end{aligned}$$

Duality rotations of these fields have invariants

$$\begin{aligned} \kappa^2 + \tau_1^2 + \tau_2^2 &= (\kappa')^2 + (\tau_2')^2 + (\tau_1')^2, \\ \kappa' \tau_1' &= \kappa \tau_1, \end{aligned} \quad (31)$$

while the Lorentz invariants $\kappa \tau_2, \kappa^2 - \tau_1^2 - \tau_2^2$ transform as

$$\begin{bmatrix} \tau_1'^2 + \tau_2'^2 - \kappa'^2 \\ \kappa' \tau_2' \end{bmatrix} = \begin{bmatrix} \cos 2\alpha & -\sin 2\alpha \\ \sin 2\alpha & \cos 2\alpha \end{bmatrix} \begin{bmatrix} \tau_1^2 + \tau_2^2 - \kappa^2 \\ \kappa \tau_2 \end{bmatrix}. \quad (32)$$

Comparing (31) and (32) to (26) we see that the Geroch transformation is analogous to a duality rotation with a scaling due to the change in metric. This has been previously noticed by Hansen⁶ in the context of potential rotations.

The null formalism allows one to discuss the horizon structure. The event horizon in static space-times coincides with the static limit of the Killing surface $\lambda = \text{const}$. In static spaces the Killing surface $\lambda = 0$ is a null-geodesic hypersurface. If the Killing congruence is rotational, this no longer must be true since the normal vector n^a and ξ^a may become null at different points, as in the Kerr space-time. This result is expressed by the well-known equation¹⁴⁻¹⁷

$$\omega^a \omega_a - n^a n_a = \frac{1}{2} \xi^a \xi_a (\xi^b_{;c})(\xi^c_{;b}). \quad (33)$$

$\lambda = 0$ doesn't imply a null normal unless the twist is also zero. If the twist is zero on $\lambda = 0$ then the Killing surface will locally constitute part of the event horizon. Since one effect of the Geroch transformation can be to change the rotation of a space-time, it is of interest to study the behavior of the Killing surface with regard to the event horizon. There are two cases to consider. (1) The Killing surface is a null-geodesic hypersurface. (2) The Killing vector is null.

A. $\lambda = 0, n^a n_a = 0$

The base space for this case has vectors zero or null. The zero Killing surface is part of the local event horizon. Under the Geroch transformation the vectors remain zero or null. The transformed space will also have the zero killing surface as part of the event horizon.

A good example of this occurs in the Schwarzschild space-time. For $\lambda \neq 0$ we have

$$dS^2 = -(1 - 2M/r)dt^2 + dr^2/(1 - 2M/r) + r^2(d\theta^2 + \sin^2\theta d\phi^2). \quad (34)$$

The vectors are

$$\hat{n} = -\frac{2M}{r^2} \frac{\hat{r}}{(g_{rr})^{1/2}}, \quad \hat{\omega} = 0.$$

Under the Geroch transformation the metric becomes¹⁸

$$\begin{aligned} dS^2 &= -\frac{[1 - 2M/r]}{F(r)} dt^2 + \frac{F(r)}{(1 - 2M/r)} dr^2 + r^2 F(r) d\theta^2 \\ &\quad + F(r) r^2 \sin^2\theta d\phi^2 + d\phi^2 \left(\frac{g'_{\phi\phi}}{\lambda} \right)^2 + 2g'_{\phi t} d\phi dt, \end{aligned} \quad (35)$$

with $F(r) = 1 + [(1 - 2M/r)^2 - 1] \sin^2\gamma$, $\lambda' = -(1 - 2M/r)/F(r)$.

$$g'_{\phi\phi} = 2Mr \frac{(r - 2M)(\cos\theta \sin 2\gamma - 2M \sin^2\gamma)}{r^2 + 4M(M - r) \sin^2\gamma}.$$

The new space has normal

$$\hat{n}' = -\frac{M \cos \alpha}{r^2} \frac{\hat{r}'}{(g_{rr})^{1/2}}$$

and twist

$$\hat{\omega}' = (-M/r^2) \sin \alpha [\hat{r}' / (g_{rr})^{1/2}]$$

for $\lambda' \neq 0$. On $\lambda' = 0$ we have $\hat{\omega}' = 0$ and $\hat{n}' = \hat{n}$ so the horizon structure is maintained.

B. $\lambda = 0, n^a n_a = 0$

In the base space the Killing surface $\lambda = 0$ is not coincident with the event horizon. The Geroch transformation will not connect this space with any space where the surfaces are coincident. To see this, note $\lambda = 0, n^a n_a \neq 0$ implies $n^a n_a = \omega^a \omega_a = \kappa_1^2$ by (20). Since by (27) and (29) we have $n'_a = n_a$ and $\omega'_a = \omega_a$, we find the same behavior in the transformed space-time at $\lambda' = 0$. λ' cannot be part of the local event horizon if λ is not. The Kerr metric is an example of this kind of behavior. We have for this metric

$$\lambda = +[\rho^2 - 2Mr]/\rho^2,$$

with $\rho^2 = r^2 + a^2 \cos^2 \theta$.

The norm and twist are

$$\begin{aligned} \hat{n} &= (m/\rho^4) \left[(-r^2 + a^2 \cos^2 \theta) \hat{r} \sqrt{\frac{\Delta}{\rho^2}} + (2ra \cos \theta \sin \theta) \hat{\theta} \sqrt{\frac{1}{\rho^2}} \right], \\ \hat{\omega} &= (m/\rho^4) \left[-2ar \cos \theta \hat{r} \sqrt{\frac{\Delta}{\rho^2}} + a \sin \theta (-r^2 + a^2 \cos^2 \theta) \hat{\theta} \sqrt{\frac{1}{\rho^2}} \right]; \end{aligned} \quad (36)$$

on the zero Killing surface these become

$$\begin{aligned} \hat{n}'_0 &= \hat{n}_0 = \frac{m}{(2Mr)^2} \sqrt{\frac{\Delta}{2Mr}} [(-r^2 + a^2 \cos^2 \theta) \hat{r} + (2ra \cos \theta) \hat{\theta}], \\ \hat{\omega}'_0 &= \hat{\omega}_0 = \frac{m}{(2Mr)^2} \sqrt{\frac{\Delta}{2Mr}} [-2ra \cos \theta \hat{r} + (-r^2 + a^2 \cos^2 \theta) \hat{\theta}]. \end{aligned} \quad (37)$$

Equation (33) is satisfied in both the base and transformed space.

Consider a space, exhibiting the above behavior, which has an axial Killing vector ξ^a_ϕ in addition to the time-translational Killing vector ξ^a_t . In such a space it is possible to define a mixed Killing vector

$$\xi^a_M = \rho \sin(\psi) \xi^a_t + \cos(\psi) \xi^a_\phi, \quad (38)$$

where the coefficients are chosen so that $\lambda_M = \xi^a_M \xi_{Ma}$ is zero on the horizon Σ . One may either explicitly set $\lambda_M = 0^{16}$ on Σ or equivalently require $\xi^a_M \xi_{Ma} = 0^{19}$ on Σ . The latter course gives $\rho \sin \psi = -\cos \psi (g_{\phi\phi}/g_{00})$ and gives ξ^a_M as

$$\xi^a_M = \rho \sin(\psi) \left[\xi^a_t - \frac{g_{0\phi}}{g_{\phi\phi}} \xi^a_\phi \right] \quad (39)$$

giving

$$\lambda_M = \rho^2 \sin^2(\psi) [\lambda - g_{0\phi}^2/g_{\phi\phi}], \quad (40)$$

which is zero on Σ . This procedure changes the Killing surface-horizon relation to that considered in subsec. A above, $\lambda_M = 0, n^a_M n_{Ma} = 0$. To see what happens to the mixed Killing surface under the transformation, form a new mixed Killing vector

$$\xi^a'_M = \rho \sin(\psi) [\xi^a_t - (g'_{0\phi}/g'_{\phi\phi}) \xi^a_\phi] \quad (41)$$

and

$$\lambda'_M = \rho^2 \sin^2(\psi) [\lambda' - g'^2_{0\phi}/g'_{\phi\phi}]$$

under the transformation we have from (11) and (12)

$$\begin{aligned} \lambda' &= \lambda/F, \\ g'_{0\phi} &= g_{0\phi} + \lambda A, \\ g'_{\phi\phi} &= F g_{\phi\phi} + 2g_{0\phi} A + \lambda'^2 A^2, \end{aligned} \quad (42)$$

with

$$\begin{aligned} A &= \alpha \sin 2\gamma - \beta \sin^2 \gamma, \\ F &= 1 + \omega \sin 2\gamma + (\lambda^2 + \omega^2 - 1) \sin^2 \gamma. \end{aligned}$$

Substituting into (41) we find $\lambda'_M = 0$ on Σ . The transformation also maintains the relation between the mixed Killing surface and the horizon.

In conclusion, we have shown that the Geroch transformation will only connect spaces with an identical Killing surface-horizon relation. This is a strong limitation on the kinds of spaces one may reach for a given base space.

¹H. Buchdahl, *Quart. J. Math.* 5, (1954).

²J. Ehlers in *Les theories relativistes de la gravitation* (CNRS, Paris, 1959).

³B. K. Harrison, *J. Math. Phys.* 9, 1744 (1968).

⁴R. Geroch, *J. Math. Phys.* 12, 918 (1971).

⁵R. Geroch, *J. Math. Phys.* 13, 394 (1972).

⁶R. O. Hansen, *J. Math. Phys.* 15, 46 (1974).

⁷W. Kinnersley, *J. Math. Phys.* 18, 1529 (1977).

⁸W. Kinnersley and D. M. Chitre, *J. Math. Phys.* 18, 1538 (1977); 19, 1926 (1978); 19, 2037 (1978).

⁹W. Kinnersley and D. M. Chitre, *Phys. Rev. Lett.* 40, 1608 (1978).

¹⁰J. L. Synge, *Relativity: The General Theory* (North-Holland, Amsterdam, 1965), 2nd ed., p. 11.

¹¹E. Honig and E. L. Schucking, *J. Math. Phys.* 15, 775 (1974).

¹²J. L. Synge, *Tensor* 24, 69 (1972).

¹³W. B. Bonner, *Tensor* 20, 229 (1969).

¹⁴C. C. Dyer and E. Honig, *J. Math. Phys.* 20, 1 (1979).

¹⁵C. C. Dyer and E. Honig, *J. Math. Phys.* 20, 409 (1979).

¹⁶C. V. Vishveshwara, *J. Math. Phys.* 9, 1319 (1968).

¹⁷B. Carter, *J. Math. Phys.* 10, 70 (1969).

¹⁸The transformations $r' = r - 2M \sin^2 \gamma, t' = t + 4M^2 (\sin^2 \gamma) \phi$ will put this metric into the TAUB-NUT form.

¹⁹J. M. Bardeen, *Ap. J.* 162, 71 (1970).

All algebraically degenerate \mathcal{H} spaces, via $\mathcal{H}\mathcal{H}$ spaces

J. D. Finley, III ^{a)}

Department of Physics and Astronomy, University of New Mexico, Albuquerque, New Mexico, 87131

J. F. Plebański ^{b)}

Departamento de Física, Centro de Investigación y de Estudios Avanzados, Instituto Politécnico Nacional, México 14, D. F., Mexico

(Received 16 July 1980; accepted 7 November 1980)

Starting from a canonical $\mathcal{H}\mathcal{H}$ space, which by virtue of the self-dual part of its conformal curvature tensor being algebraically degenerate, has a single congruence of totally null, extremal two-surfaces, along which the coordinates are built, we specialize to the case where the anti-self-dual conformal curvature vanishes, giving us an algebraically degenerate \mathcal{H} space with coordinates especially adapted to its degeneration. We then solve the $\mathcal{H}\mathcal{H}$ equation in this case and obtain, at one stroke, all algebraically degenerate \mathcal{H} spaces in a simple, compact form, useful for later applications. The generic solution depends on four arbitrary holomorphic functions of two variables, and any special Petrov type desired is easily distinguished. Also, a special comparison is made for all such spaces of type D, showing the relation of their parameters to the usual real, type-D parameters of mass, Newman-Unti-Tamburino (NUT) parameter, rotation and acceleration. Lastly, a contraction to the special cases in which the leaves of the congruence are relatively plane is performed explicitly.

PACS numbers: 02.40.Re

I. INTRODUCTION

Within the theory of \mathcal{H} spaces, approached along different lines by several groups, the problem of an explicit determination of the class of all algebraically degenerate heavens was one of the first to be systematically pursued. Already¹ in 1975, simple examples were given of all algebraically degenerate types. Further progress was made in Ref. 2, particularly with respect to their symmetries. Then, Fette, Janis, and Newman³ determined all algebraically degenerate \mathcal{H} spaces, separated according to whether they were diverging or not. The Sachs–Goldberg theorem⁴ assures us that every algebraically degenerate vacuum spacetime has a distinguished congruence of geodesic, shearfree (real) null directions. In Ref. 3 the case in which the divergence of this congruence was nonzero was investigated, while the case of vanishing divergence was completed in Ref. 4. Their results establish, in explicit although somewhat complicated form, the metric and the nonzero components of the conformal curvature. In particular they give a characterization, by complex Petrov type, of the number of arbitrary functions in the most general case. In Ref. 2 the current authors also gave a complete description of heavens of type N, only—in terms of two functions of two variables—in a very simple and compact form. However, both of the above approaches use traditional techniques of integration inherited from the theory of *real* algebraically degenerate spaces. In particular the use of the behavior of the real congruence of null geodesics is not necessarily what would be motivated by the Sachs–Goldberg theorem on a complex manifold. Plebański and Hacyan⁵

have given a complex form of the Sachs–Goldberg theorem which shows that there are separate geometric consequences of the algebraic degeneration of the self-dual and anti-self-dual components of the conformal curvature tensor. The geometric object given because of the degeneration of only one side of the conformal curvature is a congruence of totally geodesic null two-surfaces—null strings. Although an algebraically degenerate heaven is degenerate from both sides and therefore has two such congruences, whose intersection determines the geodesic, shearfree, null congruence mentioned earlier, it is not the most natural object in the space on which to build the character of explicit solutions. Therefore, in this approach we consider separate cases determined by the complex expansion of the fundamental congruence of null two-surfaces required by the complex Sachs–Goldberg theorem because of the algebraic degeneration.

Since these null two-surfaces play a prominent role in the discussion which follows, we give here a brief geometrical picture of their properties and refer the reader to the literature⁶ for more details. A (two-parameter) congruence of totally null two-surfaces *foliates* the given four-dimensional manifold. A basis for the (reduced) tangent space of any point on such a surface contains two *orthogonal* null vectors. In general we can study the curvature of one of our two-surfaces by looking at the changes in both the (two-dimensional set of) tangent vectors and the (two-dimensional set of) normal vectors as one moves along the surface. Referring to Ref. 7 for the details of the calculations, we first consider the covariant derivative, along a particular leaf of our congruence, of the tangent vectors. In general this would be expected to have both tangential components and normal components. However, for these two-surfaces, the normal components vanish, thereby justifying the statement that they are totally geodesic two-surfaces—they are extrinsic-

^{a)}Work supported in part by the Fomento Educacional, A. C., México 5, D. F., México, and by the C.I.E.A. del I.P.N., México 14, D. F., México.

^{b)}On leave of absence from the University of Warsaw, Warsaw, Poland.

ly "flat".

Also, the basis set may always be chosen in such a way that the tangential components also vanish—they have a vanishing intrinsic curvature. What, then, may one say about the structure of these two-surfaces. The desired information is in the covariant derivative, along a particular leaf, of the normal vectors. Particularly important are the normal components of these covariant derivatives. These components simply measure how the normal direction changes as one moves (by parallel transport) along a given null string. In particular, if these covariant derivatives have vanishing normal components, then the normal directions may all be chosen parallel as one moves around. The one-form which has the coefficients of these normal components as components is then referred to as the *expansion* of the leaves of the congruence. An explicit (coordinate-based) expression for this expansion is given in Sec. II. Here we have simply tried to give an intuitive geometric description of this quantity. We also note that, while the general case is one where the expansion of the leaves is nonzero, it is certainly permitted to consider the case of zero expansion—referred to as plane. Since the coordinatization to be used to describe the general case depends in an explicit way upon the nonzero value of the expansion, it is necessary, in this approach, to perform a limiting procedure in order to go to a reasonable coordinatization in the plane case.

We note that every $\mathcal{H}\mathcal{H}$ space automatically has two distinct congruences of null strings because of the fact that it is half conformally flat, e.g., the anti-self-dual part of the conformal curvature vanishes. However, the existence of neither of these congruences is directly addressed by the extra consideration that the \mathcal{H} space in question is algebraically degenerate. This degeneration guarantees us *another* congruence of null strings which are a direct consequence of the special case which we wish to consider. Therefore, we want to build our characterization of the solutions around this particular congruence. In order to do that we note that this special congruence is the only one possessed by a more complicated family of spacetimes, namely $\mathcal{H}\mathcal{H}$ spaces with, for example, algebraic degeneration of the self-dual conformal curvature but generality of the anti-self-dual part.⁸ The integration procedure for such $\mathcal{H}\mathcal{H}$ spaces has already been carried out in detail and allows all $\mathcal{H}\mathcal{H}$ spaces to be completely described in terms of a potential function W which must satisfy a certain nonlinear differential equation and certain functions μ , ν , ξ , and γ of two variables only—constant on each leaf of the distinguished congruence of null strings. By working, therefore, in this more general formalism we ensure that the approach is based on the particular null string desired. Then, by putting the differential constraint on the system that the previously general part of the conformal curvature should now vanish, we descend to the desired level of an algebraically degenerate \mathcal{H} space and obtain constraints on the form of W which are sufficient to allow us to proceed to a complete integration of the problem. In Ref. 7 this alternate approach was pushed forward in a simple subcase—nonexpanding \mathcal{H} spaces of type N. In this paper we proceed with the program in complete generality.

In Sec. II, we give a brief description of the basic spinor-

ial formalism⁹ for $\mathcal{H}\mathcal{H}$ spaces and then proceed immediately to the problem of integration. We show that the system of differential equations which results when the right side of an $\mathcal{H}\mathcal{H}$ space is made flat amounts to a natural structure of an ideal of two forms in two dimensions. That system is integrated in all generality. Then, realizing that a number of the functions in the solution are only there because of coordinate gauge freedom, in Sec. III we discuss the general (local) group of transformations which leave the structural equations form invariant. This group is utilized, leaving only those arbitrary functions which have geometric significance. In particular, the relation of these functions to the Petrov type is detailed.

In Sec. IV, we discuss in more detail the special case of type D \mathcal{H} spaces and the relation to the type D $\mathcal{H}\mathcal{H}$ spaces already known. In addition, we show how these relate to some common gravitational instantons. In Sec. V, we discuss the limiting procedure in those degenerate cases where the distinguished congruence of two-surfaces is plane rather than expanding. Only \mathcal{H} spaces of type III or N have this possibility, and they are all simple contractions of the more general ones already discussed.

II. GENERAL INTEGRATION OF THE EQUATIONS

We recall that a general (expanding) $\mathcal{H}\mathcal{H}$ space can be described by a pair of spinor coordinates, p^A , which are coordinates along any given leaf of the congruence, and q_B , which are parameters which label the various members of the congruence. The two-surfaces in the congruence are then the surfaces of the (closed) two-form $\Sigma = \frac{1}{2}dq_A \wedge dq^A$. The (nonzero) expansion of the congruence picks out a special direction on any given leaf. In order to specify this we pick a constant spinor basis J_A, K^B (such that $K^A J_A = \tau$) and set

$$\phi = J_A p^A, \quad \eta = K^A p_A, \quad (2.1)$$

as coordinates on the leaf, distinguished relative to the direction of the expansion. (In Ref. 9 there is an additional constant κ in the definition of ϕ , which, although useful in some applications, will be dropped in this paper.) The expansion form¹⁰ is then proportional to the components of $d\phi$ along the distinguished congruence Σ . Then the potential function $\bar{W} = \bar{W}(p^A, q_B)$ must satisfy the hyperheavenly equation

$$\frac{1}{2}\phi^A(\partial^A\phi - 2\partial^B\bar{W})(\partial_A\phi - 2\partial_B\bar{W}) + \phi^{-1}\partial^A\bar{W}_{,A} - \mu\phi^A\partial_\phi\phi^{-1}\partial_\phi\phi^{-1}\bar{W} + (\eta/\tau^2)p_A K^{(A}J^{B)}\mu_{,B} = N_A p^A + \gamma, \quad (2.2)$$

for some functions μ , γ , and N_A , functions of the q_B only, i.e., constant on any given leaf of the congruence. We use the symbols

$$\partial_A \bar{W} \equiv \partial \bar{W} / \partial p^A, \quad \bar{W}_{,A} \equiv \partial \bar{W} / \partial q^A, \quad \partial_\phi \bar{W} \equiv \partial \bar{W} / \partial \phi. \quad (2.3)$$

The metric is determined by the tetrad

$$ds^2 = 2e^2 \otimes_s e^2 + 2e^3 \otimes_s e^4 = 2E^A \otimes_s e_A, \quad (2.4)$$

$$e_A = \phi^{-2} dq_A = \begin{pmatrix} e^3 \\ e^1 \end{pmatrix},$$

$$E^A = -dp^A + Q^{AB}dq_B = \begin{pmatrix} e^A \\ e^2 \end{pmatrix},$$

$$Q^{AB} = -\phi^3 \partial^A \phi^{-2} \partial^B \bar{W} + (\mu/\tau^2) \phi^3 K^A K^B. \quad (2.5)$$

The spinorial expressions for the connections and the conformal curvature⁹ will also be needed:

$$\begin{aligned} \Gamma_{11} &= -\phi^{-1} J_A e^A, \\ \Gamma_{12} &= -\frac{1}{2} \phi (\partial^A \phi Q_{AB}) e^B - \frac{3}{2} \phi^{-1} J_B E^B \\ \Gamma_{22} &= -\phi^6 (N_B + M_{BA} P^A + \phi^{-2} J^A \partial_B \bar{W}_{,A}) e^B \\ &\quad + \phi^3 [\partial_B \phi^{-1} J^A \partial_A \bar{W} + (\mu/\tau) \phi K_B] E^B, \end{aligned} \quad (2.6)$$

with

$$\begin{aligned} M_{BA} &= [(1/2\tau) K^C \epsilon_{BA} - (1/\tau^2) K_B K_A J^C] \mu_{,C}, \\ \Gamma_{AB} &= \phi (\phi \partial_{(A} Q_{B)C} + \epsilon_{C(B} J^D Q_{A)D}) e^C + \phi^{-1} J_{(A} E_{B)}, \end{aligned} \quad (2.7)$$

$$C_{ABCD} = \phi^3 \partial_A \partial_B \partial_C \partial_D [\bar{W} - (\mu/4\tau^2) \phi^2 \eta^2]. \quad (2.8)$$

(We note that the expansion one-form is just $\Gamma_{11} = -[(\partial_A (\ln \phi))] e^A$, in this choice of coordinates.¹⁰)

$$\begin{aligned} C^{(5)} &= 0 = C^{(4)}, \quad C^{(3)} = -2\mu \phi^3, \\ C^{(2)} &= 2\phi^5 (N_A J^A - p^A \mu_{,A}), \\ C^{(1)} &= 2\phi^7 [\phi^3 (\mu/\tau) K^A N_A + J^A (\gamma + 3\mu \bar{W})_{,A} \\ &\quad + 2(N_A J^A J_B - p_{(B} J_{A)} \mu_{,A}) \partial^B \bar{W} \\ &\quad + J^B N_{B,A} P^A + (\mu/2\tau^2) \phi^3 \eta K^A \mu_{,A} - \frac{1}{2} p^A p^B \mu_{,AB}]. \end{aligned} \quad (2.9)$$

The obvious way to specialize such a solution to a general \mathcal{H} space is by setting μ , $N^A J_A$, and γ equal to zero. However, this would give us a space parametrized over the pair of null strings possessed by all \mathcal{H} spaces. (Actually an \mathcal{H} space in this case.) On the other hand, by requiring that \bar{W} be such as to make C_{ABCD} vanish, we obtain an arbitrary algebraically degenerate \mathcal{H} space parametrized over the special null string the space possesses because of its degeneration. This requirement tells us that \bar{W} is a fourth-order polynomial in ϕ and η ,

$$\bar{W} = (\mu/4\tau^2) \phi^2 \eta^2 + \frac{1}{6} A_{ABC} p^A p^B p^C + \frac{1}{2} B_{AB} p^A p^B + C_A p^A + \bar{D}, \quad (2.10)$$

where A_{ABC} , B_{AB} , C_A , and \bar{D} are, as yet, arbitrary functions of the q_A only.

However, \bar{W} must still satisfy the hyperheavenly equation (2.2). Since the unknown functions depend only on q_A , this takes the form of a coupled system of six equations for the ten unknown quantities, as well as the four quantities μ , γ , N_A

$$\begin{aligned} \frac{1}{2} A_{ABC}{}^C + J^C A_{CD(A} B^D{}_{B)} + (N_{(A} + \mu C_{(A} J_{B)}) &= 0, \\ B_{AB}{}^{,B} + (\frac{1}{2} B_{BC} B^{\dot{B}C} + \gamma + 3\mu \bar{D}) J_A + 2J^B C^C A_{ABC} &= 0, \\ C_A{}^{,A} + 2J^A C^B B_{AB} &= 0. \end{aligned} \quad (2.11)$$

These equations can be made much more tractable by reinterpreting them as equations involving some appropriate one-forms.

$$\begin{aligned} A &\equiv 2A_{ABC} J^A J^B dq^C \equiv \bar{\alpha} dt + \bar{\beta} dw, \\ D &\equiv 2A_{ABC} J^A K^B dq^C \equiv \bar{\beta} dt + \bar{\gamma} dw, \\ E &\equiv 2A_{ABC} K^A K^B dq^C \equiv \bar{\gamma} dt + \bar{\delta} dw, \end{aligned} \quad (2.12)$$

$$\begin{aligned} B &\equiv B_{AB} J^A dq^B \equiv \bar{\epsilon} dt + \bar{\zeta} dw, \\ F &\equiv B_{AB} K^A dq^B \equiv \bar{\zeta} dt + \bar{\eta} dw, \\ C &\equiv C_A dq^A \equiv \bar{\theta} dt + \bar{\kappa} dw, \end{aligned}$$

where the combinations of the q_A have been chosen so as to be adapted to the expansion direction, with

$$t \equiv K^A q_A, \quad w \equiv J^A q_A. \quad (2.13)$$

We note that not all the components of these forms are independent. Rather, the obvious constraints are

$$\begin{aligned} A \wedge dt + D \wedge dw &= 0 \\ &= D \wedge dt + E \wedge dw \\ &= B \wedge dt + F \wedge dw. \end{aligned} \quad (2.14)$$

The complete contraction of Eqs. (2.11) with J^A then leads to part of the equations involving only a closed subideal of our forms

$$dA = 2B \wedge A, \quad dB = C \wedge A, \quad dC = 2C \wedge B, \quad (2.15)$$

while the remaining equations become

$$dD = F \wedge A + B \wedge D + \tau(N_C + C_C) J^C dq^A \wedge dq_A, \quad (2.16)$$

$$dF = F \wedge B + C \wedge D + \frac{1}{2} \tau (\gamma + 3\mu \bar{D}) dq^A \wedge dq_A,$$

$$dE = 2F \wedge D + 2\tau(N_C + \mu C_C) K^C dq^A \wedge dq_A. \quad (2.17)$$

Equations (2.15) form a very simple set of equations which are, in fact, simply the Maurer–Cartan equations for the connection over $SU(2)$. A very convenient form in which to display their solutions is obtained by considering an orthonormal pair of spinor functions M^A , L_B such that $M^A L_A = 1$. In terms of the three independent functions remaining, the solution is then calculated to be

$$\begin{aligned} A &= L^A dL_A, \quad B = M^A dL_A = L^A dM_A, \\ C &= M^A dM_A. \end{aligned} \quad (2.18)$$

By exhibiting the solution in this form an explicit invariance under constant $SL(2, C)$ transformations is retained, allowing degenerate cases to be handled more easily.

For the inhomogeneous equations we set

$$\begin{aligned} Q^B &= M^B D - L^B F, \\ \frac{1}{2} R^B &= \tau M^B (N_C + \mu C_C) J^C - \frac{1}{2} \tau L^B (\gamma + 3\mu \bar{D}), \\ T &= 4\tau(N_C + \mu C_C) K^C, \end{aligned} \quad (2.19a)$$

which allows the equations to be written more simply:

$$dQ^B = \frac{1}{2} R^B dq^A \wedge dq_A, \quad dE = Q^A \wedge Q_A + \frac{1}{2} T dq^A \wedge dq_A. \quad (2.19b)$$

Since the entries in these equations are all two-forms in two dimensions, they are (locally) closed and therefore (locally) integrable. To demonstrate this explicitly we introduce three potentials

$$S^{\dot{B}C}{}_{,C} \equiv R^B, \quad H^A{}_{,A} \equiv T, \quad X^A{}_{,A} \equiv S^{\dot{B}C} S_{\dot{B}C}. \quad (2.20)$$

It is then immediate that there should exist some functions

P^A such that $Q^A = S^{AB}dq_B + dP^A$. However, S^{AB} has considerable freedom of choice since any solution of the defining equation (2.20) serves as well as any other. In particular, if S_0^{AB} is a suitable such potential, then $S_{AB} = S_0^{AB} + P^{A,B}$ will also satisfy the potential equation. Therefore, without loss of generality, the dP^A term may always be absorbed in the other, giving us the general solution

$$Q^A = S^{AB}dq_B, \quad E = (X^A + H^A)dq_A, \quad (2.21)$$

where an arbitrary dp term in the integration for E has, similarly, been absorbed in the $X^A dq_A$ term, while no particular choice of the potential H^A has yet been made.

We have acquired a general solution of Eqs. (2.15)–(2.17) in a rather simple form. However, the constraints given by Eqs. (2.14) must still be imposed in order to guarantee that we have a solution of the hyperheavenly equation. In this notation they take the form $S^{AB}J_B + \tau L_w^A = 0$ and $X^A J_A = S^{AB}L_A K_B$, where we have chosen $H^A = (h - f)J^A$ and the subscript w indicates partial differentiation with respect to w . This implies that there is some spinor Z^A such that $S^{AB} = Z^A J^B - L_w^A K^B$. Writing $X^A = fJ^A + bK^A$, the solution to all the constraint equations is then given by

$$\begin{aligned} Z^A &= cL^A - bM^A, & f_t &= 2cL^A L_{Aw} - 2bM^A L_{Aw} - b_w, \\ T &= \tau(h - f)_t, & R^A &= \tau(cL^A - bM^A)_t - \tau L_{ww}^A, \end{aligned} \quad (2.22)$$

where only f_t (as opposed to f) appears in the final result and the equations for T and R^A are to be interpreted as determining the previously unknown functions N_A and $\gamma + 3\mu\bar{D}$ in terms of the arbitrary functions c , h , and b . At this point, we summarize the solution by giving, in terms of the three independent components of L^A and M^B , as well as b , c , and h .

$$\begin{aligned} A &= L^A dL_A, & B &= M^A dL_A, & C &= M^A dM_A, \\ D &= bdw - L_B L_w^B dt, & E &= -hdw + bdt, \\ F &= cdw - M_B L_w^B dt, \\ 2N^A J_A &\equiv \tau v = b_t + L_D L_{ww}^D + bM_A L_t^A - cL_A L_t^A \\ &\quad - 2\tau\mu M_A M_t^A, \\ 2K^A N_A &\equiv \tau\xi = \frac{1}{2}h_t + \frac{1}{2}b_w + cL_A L^A w - bM_A L_w^A \\ &\quad + 2\tau\mu M_A M_w^A \\ \gamma + 3\mu\bar{D} &= -c_t + cM_A L_t^A - bM_A M_t^A - M_A L_{ww}^A. \end{aligned} \quad (2.23)$$

Having the one-forms A through F and the functions $v, \xi, \gamma + 3\mu\bar{D}$ in terms of L^A, M^B (where $M^A L_A = 1$) and b, c, h , and \bar{D} , we could insert these in \bar{W} and have the general solution in terms of seven arbitrary functions of two variables. However, some of these functions simply express the gauge freedom still available. Therefore, in the next section we look at the gauge freedom.

III. GAUGE TRANSFORMATIONS IN $\mathcal{H}\mathcal{H}$ SPACES

There is no reason to believe that there is any essential physics in a particular parameterization of the leaves of our special null congruence. Therefore we consider an invertible transformation to new parameters $q'_R = q'_R(q_A)$. Since the specific form of the tetrad given in Eq. (2.4) has been used to

acquire the particular forms given above, we determine new longitudinal coordinates p'^S so as to preserve the form of the tetrad in terms of the new variables:

$$\begin{aligned} dq'_R &= D_R^A dq_A, \\ p'^R &= \lambda^{-1} D^{-1}{}^R{}_A p^A + \sigma^R, \\ \phi' &= \lambda^{-1/2} \phi, \\ E'^R &= \lambda^{-1} (D^{-1}{}^R{}_A E^A - \tilde{\eta} e'^R), \end{aligned} \quad (3.1)$$

where D_R^A, λ, σ^R , and $\tilde{\eta}$ are arbitrary functions of the q_B .

In Refs. 7 and 9 a number of elementary consequences of this transformation are worked out. However, the restrictions which must be maintained in order to maintain the expansion direction as special are not given there. In order to do this we may specify a new, *constant* spinor basis J'^S, K'^S such that $K'^S J'_S = \tau' \equiv a_0 \tau$. These restrictions are sufficient to ensure the existence of a completely analogous potential description in terms of $\bar{W}'(p'^R, q'^S), \mu', \nu'$, etc. It is straightforward (if somewhat lengthy) to show that the most general such transformation is determined by four arbitrary functions ξ, s, θ, v of q_A , one arbitrary function $g = g(w)$, only, and four arbitrary constants, J'^R, K'^S , and $a_0 = \tau'/\tau$. Then one finds the explicit form of the transformation to be given by

$$\begin{aligned} q'_R &= (1/\tau)(\xi J'_R + gK'_R/a_0), \\ D_R^A &= (1/\tau)(J'_R \xi^A - yK'_R J^A/a), \end{aligned}$$

where $y = dg/dw$, or

$$\begin{aligned} t' &= a_0 \xi(t, w), & w' &= g(w), & \lambda^{-1/2} &= \xi_t, \\ \sigma^R &= sJ'^R, & \det D_R^A &= \lambda^{-1/2} y \neq 0, \\ 2\tilde{\eta}\phi^{-2} &= \partial\sigma^R/\partial q'^R + (\lambda^{1/2}/y)p^A (\lambda^{-1})_{,A} \\ &= \lambda^{1/2} \tau s_t - (\lambda^{-3/2}/y)(\phi\partial_w + \eta\partial_t)\lambda. \end{aligned} \quad (3.2)$$

Having new coordinates, the transformations of the potential functions which determine the metric are shown to be

$$\begin{aligned} \bar{W} &= \lambda^{3/2} (y)^2 \bar{W}' + \frac{1}{2\tau} \lambda^{1/2} y p_A \left(\frac{1}{y} p_B \xi^{,B} \right)^A \\ &\quad - \frac{\mu \xi_w}{2\tau^2 \xi_t} \phi^3 \left(\eta + \frac{1}{2} \frac{\xi_w}{\xi_t} \phi \right) + \frac{1}{2} \lambda y p^A s_{,A} \\ &\quad + \frac{1}{3} \theta \phi^3 + v, \\ \mu &= \lambda^{-3/2} \mu', & \nu &= a_0 y \lambda^{-1} \nu' + 2\lambda y s \mu_t + 3\mu y (\lambda s)_t, \\ \xi &= \lambda^{-1/2} (y)^2 \xi' - \lambda^{1/2} \xi_w \nu - 2\tau \theta_t \\ &\quad - \lambda^{1/2} y s [\lambda (\xi_w \mu_t - \mu_w \xi_t) + \frac{3}{2} \mu (\xi_w \lambda_t - \lambda_w \xi_t)], \\ \gamma &= y^2 \gamma' - 3\mu \nu + (y)^{+1/2} [(y)^{-1/2}]'' - \frac{1}{2} \lambda y \tau s v \\ &\quad + (\lambda y s)^2 \tau \mu_t + \frac{1}{4} \tau \lambda (y s)^2 (3\mu \lambda_t - 2\lambda \mu_t). \end{aligned} \quad (3.3)$$

By insisting that \bar{W}' should be expressed in terms of the components of some A', B' , etc., we then easily find the desired transformation equations for them

$$\begin{aligned} A' &= \lambda y A, & B' &= B - \frac{1}{2} \lambda y s A + \frac{1}{2} d(\ln \lambda y), \\ C' &= (\lambda y)^{-1} [C - \lambda y s B + \frac{1}{4} (\lambda y s)^2 A - \frac{1}{2} d(\lambda y s)], \\ D' &= a_0 \lambda^{1/2} [D - \lambda^{1/2} \xi_w A - 2\tau \mu \lambda y s d w], \\ E' &= a_0^2 (y)^{-1} [E - 2\lambda^{1/2} \xi_w D + \lambda (\xi_w)^2 A \\ &\quad - 2\mu \lambda^{3/2} \tau s y d \xi - 4\tau^2 \theta d w], \end{aligned} \quad (3.4)$$

$$\begin{aligned}
F' &= a_0(\lambda^{1/2}y)^{-1}[F - \lambda^{1/2}\xi_w B - \frac{1}{2}\lambda y s(D - \lambda^{1/2}\xi_w A) \\
&\quad - d\lambda^{1/2}\xi_w + \frac{1}{2}\lambda^{1/2}(\ln y)'d\xi + \frac{1}{2}\tau\mu(\lambda y s)^2 dw], \\
\bar{D}' &= \lambda^{-3/2}(y)^{-2}\{\bar{D} - v - \tau\lambda y s[C_A - \frac{1}{2}\lambda y s B_{A\bar{B}} J^{\bar{B}} \\
&\quad + \frac{1}{6}(\lambda y s)^2 A_{A\bar{B}C} J^{\bar{B}} J^{\bar{C}}] J^A - \frac{1}{4}\tau\lambda(y)^2(\lambda s^2)_i\}.
\end{aligned}$$

As a first step, we note that, in Eqs. (3.4), v appears only in the transformation equation for \bar{D} and, moreover, that $\gamma + 3\mu\bar{D}$ is invariant under v . We may therefore always choose to eliminate \bar{D} via the freedom in v , which we assume done from now on. Again, θ appears only in the transformation for $\bar{\delta}$ (the dw -part of E) and so we may always eliminate it from the scene, making [via Eq. (2.23)] h vanish as well.

In order to proceed further we must have the transformation behavior of M^A and $L^{\bar{B}}$. Working from the behavior of A , B , and C , we find that

$$L'^{\bar{R}} = (\lambda y)^{1/2} G^{\bar{R}}_A L^A, \quad (3.5)$$

$$M'^{\bar{R}} = (\lambda y)^{-1/2} G^{\bar{R}}_A (M^A - \frac{1}{2}\lambda y s L^A),$$

where $G^{\bar{R}}_A$ is an arbitrary, constant matrix from $SL(2, C)$. Similarly, by working with D , E , and F we find, consistently, that

$$\begin{aligned}
c' &= a_0 \lambda^{-1/2}(y)^{-2} [c - \frac{1}{2}\lambda y s b \\
&\quad - (\partial_w - \lambda^{1/2}\xi_w \partial_i) \lambda^{1/2}\xi_w \\
&\quad + \lambda^{1/2}\xi_w (M^A - \frac{1}{2}\lambda y s L^A) \\
&\quad \times (2\partial_w - \lambda^{1/2}\xi_w \partial_i) L^A + \frac{1}{2}\tau\mu(\lambda y s)^2], \\
b' &= a_0 \lambda^{1/2}(y)^{-1} \\
&\quad \times [b + \lambda^{1/2}\xi_w L^A (2\partial_w - \lambda^{1/2}\xi_w \partial_i) L^A - 2\tau\mu\lambda y s].
\end{aligned} \quad (3.6)$$

Since g is a function of only one variable it is reasonable to try to gauge either L^A or $M^{\bar{B}}$ to be a function of only one variable so as to use y to affect it further. In general the only way this can be done is to cause $(M^{\bar{R}})_{i'}$ = 0 (vanishing of $\bar{\epsilon}$ and $\bar{\theta}$). By consulting Eqs. (3.4) we find that choosing a transformation with $s_i = \frac{1}{2}\alpha s^2 - 2\bar{\epsilon}s + 2\bar{\theta}$ causes $\bar{\theta}'$ to vanish and then a transformation with $\lambda_i = -2\bar{\epsilon}\lambda$ will cause $\bar{\epsilon}'$ to vanish (and maintain $\bar{\theta}'$ as zero). So using up s and λ_i will cause $(M^{\bar{R}})_{i'}$ = 0. Dropping primes, we assume this has been done. This implies that M^A depends only on w and that there is also a spinor function $B_{\bar{A}}(w)$, only, such that $L_{\bar{A}} = aM_{\bar{A}} + B_{\bar{A}}$, where a is an arbitrary function of q_A . We then find that $a' = \lambda y(a + l)$ where $l = l(w)$ is a new gauge freedom, and $B'^{\bar{A}} = (\lambda y)G^{\bar{A}}_A (B^{\bar{A}} - lM^{\bar{A}})$. By proper choice of l and λ_w we can arrange for $B'^{\bar{R}}$ to be constant. Lastly, using $y(w)$ to gauge to one a proportionality constant that appears, we can finally write L^A and $M^{\bar{B}}$ in the form

$$M_{\bar{A}} = wB_{\bar{A}} + A_{\bar{A}}, \quad L_{\bar{A}} = aM_{\bar{A}} + B_{\bar{A}}, \quad (3.7)$$

where $A^{\bar{A}}B_{\bar{A}} = 1$ and are both constant, which allows us to determine explicit forms for A , B , etc. having used up essentially all the gauge freedom. The results are most easily quoted by giving \bar{W} , v , ξ , and γ in terms of the four arbitrary functions of two variables, μ, a, b, c

$$\begin{aligned}
\bar{W} &= (1/\tau^2) [\frac{1}{4}\mu^2\eta^2 - \frac{1}{12}a_i\eta^3 + \frac{1}{4}(a^2 - a_w)\eta^2\phi + \frac{1}{4}b\eta\phi^2] \\
&\quad + \frac{1}{\tau} [a\eta\phi + \frac{1}{2}c\phi^2] + \phi,
\end{aligned} \quad (3.8)$$

$$\begin{aligned}
\tau v &= b_i - (a^2 - a_w)_w - ca_i, \\
\tau\xi &= -2\tau\mu + \frac{1}{2}b_w + ba - c(a^2 - a_w), \\
\gamma &= -c_i + 2a_w.
\end{aligned}$$

Also, for convenience we note that

$$\begin{aligned}
ds^2 &= (\phi\tau)^{-2} \{ 2\tau(d\eta \otimes_s dw - d\phi \otimes_s dt) \\
&\quad + [\mu\phi^2 + a_i\eta\phi - (a^2 - a_w)\phi^2] dt \otimes_s dt \\
&\quad - (2\mu\eta\phi^2 + b\phi^2 + a_i\eta^2) dw \otimes_s dt \\
&\quad + [\mu\eta^2\phi + (a^2 - a_w)\eta^2 + b\eta\phi + 4\tau a\eta \\
&\quad + 2\tau c\phi + 4\tau^2] dw \otimes_s dw \}.
\end{aligned} \quad (3.9)$$

The curvature and other quantities are determined by μ , v , ξ , γ and \bar{W} via Eqs. (2.6)–(2.9).

IV. PETROV TYPES, ESPECIALLY D

The solution given contains all possible algebraically degenerate \mathcal{H} spaces. Therefore, it includes complex Petrov types II, D, III, N, and flat. It is interesting to point out the specializations obtained by insisting that the solution be of a particular type. To have Petrov type III it is only necessary to insist that $\mu \equiv 0$, leaving us with three independent functions of the q_A . Petrov type N is obtained by putting both μ and v to zero, which gives the simple constraint between the three remaining functions

$$b_i - (a^2 - a_w)_w - ca_i = 0, \quad (4.1)$$

which we may think of, for example, as determining b , reducing the number of independent functions to two. It is worth noting that the most general type N \mathcal{H} spaces, with two arbitrary functions of two variables, were, in fact, already published in Ref. 2 sometime ago. However, as already mentioned, they were in a form more suited for their study as if they were real manifolds rather than as the complex manifolds they are (without real Minkowski sections). The \mathcal{H} space would be flat if, in addition to $\mu = 0$ and $v = 0$ [Eq. (4.1)], we also would insist that

$$\gamma_i = 0 = (2a_w - c_i)_i. \quad (4.2)$$

Clearly, at this point the extra arbitrary function left over simply represents some residual gauge freedom which was not available in general.

In the above list we have omitted Petrov type D \mathcal{H} spaces. In general, a type II space can be constrained to be type D (when a tetrad is used in which $C^{(5)}$ vanishes) by insisting on the constraint

$$2[C^{(2)}]^2 = 3C^{(1)}C^{(3)}, \quad C^{(3)} \neq 0. \quad (4.3)$$

This is clearly [from Eqs. (2.9) and (3.8)] a very messy set of coupled equations to solve for the coefficients of the various powers of ϕ and η . However, there is a more useful gauge to demonstrate the solutions of type D, which also makes a bridge to the work of García and Plebański¹¹ which demonstrates the usual real Petrov type D seven-parameter solution of Plebański and Demiański¹² in the canonical form of an $\mathcal{H}\mathcal{H}$ space.

Starting with the general form for \bar{W} of Eq. (2.10) and all gauge freedom, we note that $C^{(3)} \neq 0$ assures us that $\mu \neq 0$. However, Eqs. (3.3) allows us to use the gauge freedom of λ to make $\mu = \mu_0$, a (nonzero) constant. Then the freedom of s , may be used to make ν' zero, from which we see that $C^{(2)}$ is now gauged to zero. The requirement to be type D [Eq. (4.3)] now tells us that $C^{(1)}$ must vanish. However, again from the transformation equations in Eqs. (3.3), we may choose θ , to cause ξ' to vanish and ν , to cause γ , to vanish, leaving $C^{(1)} = 6\mu_0\phi^2\tau\bar{W}$, which indicates that there is a choice of gauge for type D in which μ is a constant ($=\mu_0$), and \bar{W} is independent of t . Therefore, all the coefficients [in Eq. (2.10)] which define \bar{W} must be functions of w only. We also note that the gauge freedom indicated in Eqs. (3.1)–(3.3) is retained with the restriction that s , θ , and ν must now be functions of w , while $\xi = z_0t + z(w)$, where z_0 is a constant. Since the functions that determine W also depend, in this case, on w only, this restriction causes no trouble at all. As before, we use $\theta = \theta(w)$ and $\nu = \nu(w)$ to eliminate δ and \bar{D} .

To proceed further we first consider the general case in which $\bar{\alpha} \neq 0$. In this case one may choose $z'(w)$ to make β' vanish, and then a transformation with $y(w)$ may be performed so as to make ξ' vanish, still allowing constant y transformations. With these transformations, the appropriate differential equations (2.11) may be rewritten in terms of the variables in Eq. (2.12) as

$$\begin{aligned}\bar{\alpha}_w &= 0, & 0 &= \bar{\alpha}\bar{\eta} - \bar{\gamma}\bar{\epsilon} + 2\tau\mu_0\bar{\theta}, \\ \bar{\epsilon}_w &= \bar{\alpha}\bar{\kappa}, & \bar{\gamma}_w &= 4\tau\mu_0\bar{\kappa}, \\ \bar{\theta}_w &= 2\bar{\epsilon}\bar{\kappa}, & 0 &= \bar{\epsilon}\bar{\eta} - \bar{\gamma}\bar{\theta} + \tau\gamma.\end{aligned}\quad (4.4)$$

These equations can already be completely integrated. The solution is given explicitly by

$$\begin{aligned}\bar{\alpha} &= \bar{\alpha}_0, & \bar{\eta} &= -2\tau\mu_0\bar{\theta}_0/\bar{\alpha}_0 + \bar{\gamma}_0\bar{\epsilon}/\bar{\alpha}_0 + 2\tau\mu_0\bar{\epsilon}^2/\bar{\alpha}_0^2 \\ \bar{\beta} &= 0, & \bar{\theta} &= \bar{\theta}_0 + \bar{\epsilon}^2/\bar{\alpha}_0, \\ \bar{\gamma} &= \bar{\gamma}_0 + 4\tau\mu_0\bar{\epsilon}/\bar{\alpha}_0, & \bar{\kappa} &= \bar{\epsilon}_w/\bar{\alpha}_0, \\ \bar{\delta} &= 0, & \nu &= 0, \\ \bar{\epsilon} &\text{arbitrary}, & \xi &= 0, \\ \bar{\zeta} &= 0, & \gamma &= 2\mu_0\bar{\epsilon}^3/\bar{\alpha}_0^2 + 6\mu_0\bar{\theta}_0\bar{\epsilon}/\bar{\alpha}_0 + \bar{\gamma}_0\bar{\theta}_0/\tau,\end{aligned}\quad (4.5)$$

where $\bar{\alpha}_0$, $\bar{\gamma}_0$, and $\bar{\theta}_0$ are arbitrary constants while ϵ is an arbitrary function of w . However, it is reasonable to perform a transformation with $s = 2\bar{\epsilon}/\bar{\alpha}$, which ensures that $\bar{\epsilon}' \equiv 0$. Also, since $\bar{\alpha}$ is a nonzero constant, we use the constant value of g' to gauge it to one. Similarly we can use a transformation with the constant z_0 (which changes the constant value of μ_0) to arrange for $\bar{\theta}$ to become one. The solution (4.5) is then much simpler and gives (with the numerical constant τ set equal to one for simplicity)

$$\bar{W} = \frac{1}{4}(\mu_0'\phi^2\eta^2 + \frac{1}{3}\eta^3 + \gamma_0'\eta\phi^2) - \mu_0'\phi^2 + \eta, \quad (4.6a)$$

a solution depending on only two, *complex*, constant parameters, with

$$\nu = 0, \quad \gamma = \gamma_0', \quad \mu = \mu_0'. \quad (4.6b)$$

The remaining case of $\alpha \equiv 0$ can be reached by contractions of the above. However, for simplicity we note them here. There are two subcases. If $\beta \neq 0$, then *it* may be gauged

to one and the solution is quickly found to be

$$\bar{W} = \frac{1}{4}\phi\eta^2(\mu_0\phi + 1), \quad \mu = \mu_0, \quad \nu = 0, \quad \gamma = 0. \quad (4.7a)$$

On the other hand, if, originally, $\beta \equiv 0$, then the solution is found to be

$$\bar{W} = \frac{1}{4}\eta^2(\mu_0\phi^2 + 1), \quad \mu = \mu_0, \quad \nu = 0, \quad \gamma = 0. \quad (4.7b)$$

Referring back to the general form of this way of writing the type D \mathcal{H} spaces, we want to show the relation of these to the general solution given in Eqs. (2.23), and to the solution in Ref. 11. It is very straightforward to see that Eqs. (4.6) are simply the \bar{W} generated by Eqs. (2.23) with the choices

$$\begin{aligned}L^A &= -X_0^A \sin t + Y_0^A \cos t, & h &= 0, & b &= \gamma_0, \\ X_0^A Y_{0A} &= 1, \\ M^A &= +X_0^A \cos t + Y_0^A \sin t, & c &= -2\mu_0,\end{aligned}\quad (4.8)$$

where X_0^A, Y_0^A are still constant $SL(2, C)$ gauge relics.

More interestingly, we recall that Ref. 11 gives a form for \bar{W} for the real seven-parameter Petrov type D solutions of Plebański and Demiański (considered as complex $\mathcal{H}\mathcal{H}$ spaces). By restricting these spaces so that $C_{ABCD} = 0$ ($m = in$, m the mass, and n the NUT parameter), we acquire (at least) a large class of type D \mathcal{H} spaces ($\bar{W} = -\phi^3\Pi$, where Π is the key function of Ref. 11 and

$$\begin{aligned}J_A &= \begin{pmatrix} -1 \\ -1 \end{pmatrix}, & K_A &= \begin{pmatrix} 1 \\ -1 \end{pmatrix}, & \tau &= 2 \\ \bar{W} &= -m\phi^2\eta^2/16 + (\bar{\epsilon} + 2i\tilde{\gamma})\eta^3/48 - (\bar{\epsilon} - 2i\tilde{\gamma})\phi^2\eta/16 \\ &\quad + im(\phi^2 - \eta^2)/16 - \tilde{\gamma}\eta/4,\end{aligned}\quad (4.9)$$

where we have put tildes over the ϵ and γ of Ref. 11—parameters related¹² to the acceleration and rotation—to distinguish them from the similar symbols being used elsewhere in this article. This would appear to have three independent complex parameters. It would be exactly the form of solution given in Eqs. (4.5) if $\bar{\epsilon}$ had not been gauged to zero. The correspondence

$$\begin{aligned}\mu_0 &= -m, & \bar{\alpha} &= \bar{\epsilon} + 2i\tilde{\gamma}, & \bar{\gamma} &= -\bar{\epsilon} + 2i\tilde{\gamma}, \\ \bar{\epsilon} &= -im/4, \\ \bar{\eta} &= +im/4, & \bar{\theta} &= -\tilde{\gamma}/4, & \bar{\beta} = 0 = \bar{\delta} = \bar{\zeta} = \bar{D},\end{aligned}\quad (4.10)$$

matches this \bar{W} to the general form and is clearly a solution of Eqs. (4.4), so it is clear that *all* type D \mathcal{H} spaces are, in fact, included in this restriction of the $\mathcal{H}\mathcal{H}$ spaces of Ref. 11. We may also note that the allowed gauge transformation

$$\begin{aligned}\mu_0' &= -mq_0^{-3}, & \gamma_0' &= -(\bar{\epsilon}^2 + 4\tilde{\gamma}^2 + 2im^2)q_0^{-2}, \\ q_0 &\equiv [m^2/4 - \tilde{\gamma}(\bar{\epsilon} + 2i\tilde{\gamma})]^{1/2}, \\ q_0\eta' &= (\bar{\epsilon} + 2i\tilde{\gamma})\eta - im, & \phi' &= q_0\phi, \\ t' &= \frac{1}{2}q_0t, & w' &= \frac{1}{2}(\bar{\epsilon} + 2i\tilde{\gamma})^{-1}q_0^3w,\end{aligned}\quad (4.11)$$

shows the \bar{W} in Eq. (4.9) to be equivalent to the general form for type D \mathcal{H} spaces in Eqs. (4.6). [It is to be noted that this transformation does *not* reduce the number of parameters in the more general $\mathcal{H}\mathcal{H}$ space (which have all the real cross sections) from which these spaces were obtained.]

Since the solutions of Ref. 11 originate from the complexification of the usual “black hole” type real solutions

with parameters which are at least partially understood physically, we note that this association of the two sets of parameters is of some interest. Also, the contractions to, for example, the pure Schwarzschild-NUT \mathcal{H} space is of interest. Since they are previously unreported in the literature, we give first the contraction of the general Plebański-De-miański seven-parameter solution down to the Kerr-NUT case, which requires a contraction, as explained in detail in Ref. 12. The $\mathcal{H}\mathcal{H}$ canonical form for this solution is then

$$\begin{aligned} \bar{W} &= \frac{1}{4}[\phi\eta^2 - \bar{\gamma}\phi(1 - 2i\phi\eta) + i\mu\phi\eta(1 - i\phi\eta) \\ &\quad + \frac{1}{2}\bar{\mu}(1 - 4i\phi\eta)^{3/2}], \\ \nu = 0 = \xi, \quad \gamma = \frac{1}{4}, \quad J_A &= \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad K^A = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \\ \tau &= 1 \text{—Kerr-NUT.} \end{aligned} \quad (4.12)$$

The identification with the usual Boyer-Lindquist coordinates is obtained by setting

$$\begin{aligned} -1/\phi &= q + ip, & +i/\eta &= 1/q - i/p, \\ -\mu &= m + in, & -\bar{\mu} &= m - in, \end{aligned} \quad (4.13)$$

and then making the identification given in Ref. 12. This is of course not yet an \mathcal{H} space. We arrange for that by setting $\bar{\mu} = 0$. Then we also perform the contraction on the usual Schwarzschild-NUT self-dual solution (with $m = in$), obtaining

$$\bar{W} = -\frac{1}{4}\phi(1 + 2m\phi) + \frac{1}{4}\phi\eta^2(1 - 2m\phi), \quad (4.14a)$$

$\mu = -2m, \nu = 0 = \xi, \gamma = \frac{1}{4}, J_A = \begin{pmatrix} 0 \\ -1 \end{pmatrix} = K^A, \tau = 1$ —self-dual Schwarzschild-NUT. By performing the gauge transformation

$$\begin{aligned} \phi' &= \phi, & \eta' &= e^v(\eta + 2m\phi + 1), & u' &= u + 2mv, \\ v' &= -e^{-v}, \end{aligned}$$

this becomes simply

$$\bar{W}' = \frac{1}{4}\phi'\eta'^2(\mu\phi' + 1), \quad \nu' = 0 = \xi' = \gamma', \quad (4.14b)$$

which is in the canonical form given as a degenerate contraction of our general solution by Eq. (4.7a). The identification with the usual Schwarzschild variables is

$$\begin{aligned} \phi &= \frac{1}{(r+m)}, & \eta &= -[(r-m)/(r+m)] \cos\theta, \\ u &= r + t + m \ln(\sin^2\theta), & v &= -i\varphi - \ln \tan\theta/2. \end{aligned}$$

V. THE LIMIT TO PLANE \mathcal{H} SPACES

All the algebraically degenerate \mathcal{H} spaces given in Secs. III and IV were based on a congruence of null strings which had nonzero complex expansion—the general case. However, it is possible for an $\mathcal{H}\mathcal{H}$ space of complex Petrov type [III] \otimes [Any] or [N] \otimes [Any] to admit a nonexpanding (plane) congruence of null strings. This is a particularly degenerate case which can be obtained from the more general case by a limiting contraction. Following the suggestions in Ref. 9, this contraction may be explicitly performed by setting

$$\begin{aligned} J_A &= \epsilon\bar{J}_A, & K^A &= \epsilon\bar{K}^A, & \tau &= \epsilon^2\bar{\tau}, & \mu &= \epsilon\bar{\mu}, \\ \bar{W} &= \bar{\Theta} + \frac{\bar{\mu}\bar{\eta}^2(1 + \epsilon\bar{\phi})}{8\epsilon} [\frac{1}{3}\epsilon\bar{\phi} + 1], & \kappa &= 1, \end{aligned} \quad (5.1)$$

where

$$\bar{\phi} \equiv \bar{J}_A p^A, \quad \bar{\eta} \equiv K^A p_A, \quad \bar{J}_A \bar{K}^A = \bar{\tau},$$

and taking the limit as $\epsilon \rightarrow 0$. In order to actually calculate the limit one must determine the behavior of $a, b,$ and c as ϵ goes to zero. Insisting that $\bar{\Theta}$ be finite gives the behavior

$$\begin{aligned} a &= f_0 + (f_0^2 \bar{w} + g_0)\epsilon + h\epsilon^2 + j\epsilon^3 + O(\epsilon^4), \\ b &= -4\bar{\tau}f_0\epsilon^2 + \bar{b}\epsilon^3 + O(\epsilon^4), \\ c &= 2\bar{\tau}\bar{c}\epsilon^2 + O(\epsilon^3), \end{aligned} \quad (5.2a)$$

where f_0 and g_0 are constants, $h, j, \bar{b},$ and \bar{c} are arbitrary functions of the q_A and $\bar{\mu}$ is now a shorthand for

$$\bar{\mu} = 2f_0(f_0^2 \bar{w} + g_0) - h_{\bar{w}}. \quad (5.2b)$$

With a gauge transformation of $\bar{\Theta}$ linear in $\bar{\eta}$ and a rescaling of the variables to gauge f_0 to one and g_0 to zero, the most general form for $\bar{\Theta}$ easily becomes

$$\begin{aligned} \bar{\Theta} &= (1/\bar{\tau}^2)(p^A l_{,A} + a)\bar{\eta}^2 - (1/\tau)\bar{\eta}(\bar{\phi} - \bar{w}), \\ F^A &= (l_{\bar{w}}/\bar{\tau})K^A, & N^A &= -(a_{\bar{w}}/2\bar{\tau})\bar{K}^A, & \gamma &= 0, \end{aligned} \quad (5.3)$$

where l and a are arbitrary functions of the two q_A formed as combinations of h and j , while \bar{b} and \bar{c} have been gauged away. We note that

$$\begin{aligned} C^{(2)} &= -l_{\bar{w}\bar{w}}, \\ C^{(1)} &= \bar{\eta}[(l_{\bar{w}})^2/2\bar{\tau} - l_{\bar{w}\bar{t}}]_{\bar{w}} \\ &\quad - l_{\bar{w}\bar{w}\bar{w}}\bar{\phi} + a_{\bar{w}\bar{w}}. \end{aligned} \quad (5.4)$$

We note that the very special case of plane type N \mathcal{H} spaces—those which have $l_{\bar{w}\bar{w}} = 0$ —have already been treated by this approach.⁷ However, they do not appear particularly similar. The reason is that although $C_{ABCD} = 0$, as is essential for an \mathcal{H} space, $\Gamma_{AB} \neq 0$, whereas the solution given in Ref. 7 is, in fact, in a gauge in which Γ_{AB} vanishes. It turns out, however, that such a gauge is possible, within the \bar{W} or $\bar{\theta}$ formalism, only for these plane N spaces. Although there is, of course, an $SL(2, \mathbb{C})$ gauge transformation which transforms Γ_{AB} away, since the anti-self-dual curvature two-form Ω_{AB} vanishes, this transformation cannot generally be written in the form of an element of the group of automorphisms which preserves the canonical form of an $\mathcal{H}\mathcal{H}$ space.

VI. CONCLUSION

In summary we point out that we have given simple explicit expressions for the potential function \bar{W} (or $\bar{\Theta}$) which generates the metric for all possible algebraically degenerate \mathcal{H} spaces [Eqs. (3.8) and (5.3)]. The solution is, of course, generically of Petrov type II, and depends on four arbitrary holomorphic functions of two variables. We point out the simple constraints that must be imposed in order to restrict attention to solutions of more degenerate Petrov type. In the case of type III one need only set $\mu = 0$, but there are two cases, depending on the complex expansion of the leaves of the distinguished congruence. In general, \mathcal{H} spaces of type III depend on three arbitrary functions of two variables with one of these disappearing when the leaves of the congruence are relatively plane.

For \mathcal{H} spaces of type N both μ and ν must vanish [Eq. (4.1)] and again there are two possibilities for the expansion of the congruence. The general case depends on two arbitrary functions of two variables, while the plane case depends on only one. For type D we have shown that the set of \mathcal{H} spaces obtained by the obvious specialization of the four-parameter solutions of Plebański and Demiański¹² do in fact exhaust all type D \mathcal{H} spaces, and have given explicit correlation with some known gravitational instantons.¹³ Since all real Euclidean algebraically degenerate manifolds are necessarily of type D or conformally flat, this does exhibit all the self-dual algebraically degenerate instantons. We note that, in this form, the invariant $C^{ABCD}C_{\dot{A}\dot{B}\dot{C}\dot{D}}$ which one needs for calculation of the topological invariant of such a space is simply proportional to $(\mu\phi^3)^2$.

Although these \mathcal{H} spaces have already been determined by Fette, Janis, and Newman,^{3,4} we believe for several reasons that it is of considerable value to give them again in this approach. The form which we have is quite simple and allows one to consider easily the entire set of these \mathcal{H} spaces as a single entity. For instance, work is now proceeding on the question of Killing vectors admitted by these spaces. This, of course, uses the formalism already set up for Killing vectors in arbitrary \mathcal{H} and $\mathcal{H}\mathcal{H}$ spaces.

Another reason for this study is the extensions the approach has to more general situations. In particular all *real*, algebraically degenerate spacetimes are contained in the (complex) $\mathcal{H}\mathcal{H}$ spaces. In this study we set $C_{\dot{A}\dot{B}\dot{C}\dot{D}} = 0$ and found the corresponding algebraically degenerate \mathcal{H} spaces. However, by looking, instead, for those $C_{\dot{A}\dot{B}\dot{C}\dot{D}}$ which would be, for example, of type [Any] \otimes [N], or [Any] \otimes [III], and

then inserting this information into the $\mathcal{H}\mathcal{H}$ equation, completely analogously to this study, we will be able to determine, for example, all complex spaces of type [N] \otimes [N]. Having this explicitly one may begin looking for the real cross sections. Spacetimes of type N are particularly amenable to such a process and work on this approach is under way.

ACKNOWLEDGMENT

We are particularly appreciative for some useful discussions with C. P. Boyer on these subjects.

¹J. F. Plebański, *J. Math. Phys.* **16**, 2396 (1975).

²J. D. Finley, III, and J. F. Plebański, *J. Math. Phys.* **17**, 585 (1976).

³C. W. Fette, A. I. Janis, and E. T. Newman, *J. Math. Phys.* **17**, 660 (1976), and C. W. Fette, A. I. Janis, and E. T. Newman, *Gen. Relativ. Gravit.* **8**, 29 (1977).

⁴J. Goldberg and R. Sachs, *Acta Phys. Pol.*, Suppl. **22**, 13 (1962).

⁵J. F. Plebański and S. Hacyan, *J. Math. Phys.* **16**, 2403 (1975).

⁶See J. F. Plebański and I. Robinson, *Phys. Rev. Lett.* **37**, 493 (1976) and Refs. 5 or 9 for early discussions. A more complete discussion is to be found in Sec. 5.1 of Ref. 7 or in C. P. Boyer and J. F. Plebański, *Rep. Math. Phys.* **14**, 111 (1978).

⁷C. P. Boyer, J. D. Finley, III, and J. F. Plebański, *General Relativity and Gravitation*, Vol. 2, edited by A. Held (Plenum, New York, 1980).

⁸J. F. Plebański and I. Robinson in *Asymptotic Structure of Space-time*, edited by F. P. Esposito and L. Witten (Plenum, New York, 1977).

⁹J. D. Finley, III, and J. F. Plebański, *J. Math. Phys.* **17**, 2207 (1976).

¹⁰This particular form for the expansion is first worked out in Ref. 9.

¹¹A. García and J. F. Plebański, *Nuovo Cimento B* **40**, 224 (1977).

¹²J. F. Plebański and M. Demiański, *Ann. Phys.* **98**, 98 (1976).

¹³See also A. S. Lapedes, "Type D Gravitational Instantons," Princeton preprint.

The wave equation in asymptotically retarded time coordinates: Waves as simple, regular functions on a compact manifold ^{a)}

Robert H. Gowdy

Department of Physics, Virginia Commonwealth University, Richmond, Virginia 23284

(Received 18 September 1980; accepted for publication 21 November 1980)

The Minkowski-space scalar wave equation is represented on a spatially compact manifold with an asymptotically retarded time. In this coordinate system, with hyperbolic space slices and space coordinates generated by a conformal map, the wave equation takes a simple time-independent form which may form a model for numerical integration calculations with nonlinear wave equations.

PACS numbers: 02.60.Lj, 04.30. + x, 02.40. + m

1. INTRODUCTION

The use of asymptotically retarded time (ART) coordinates has been suggested as a way to simplify numerical calculations of gravitational radiation from strong-field sources.¹ In such coordinates, waves could be followed all the way to future null infinity in a finite number of steps on a finite grid. Numerical calculations could then incorporate rigorous retarded boundary conditions and use the exact definitions of emitted radiation, stated at future null infinity, to interpret the results.

This simple and attractive idea gives a surprising amount of trouble when one tries to construct a *simple* example of its use. The essential problem is the need to combine two distinct features: (1) the asymptotically retarded time coordinate itself and, (2) the introduction of a compact space coordinate system. The compact space coordinate system is needed in order to realize the primary benefit of asymptotically retarded time—the description of waves by regular functions on a compact manifold. The example presented here does combine these two features.

York and Smarr have used an asymptotically retarded time representation of Minkowski space as an example of their minimal distortion shift vector condition.² The resulting spacetime metric and therefore the resulting form of the scalar wave equation remain quite simple in that case, but the spatial coordinate patch is not compact.

There is no difficulty in imagining the introduction of compact space coordinates on each surface of asymptotically retarded time. However, one must choose these space coordinates carefully if a truly simple example is to result.

Here, I present a spatially compact, asymptotically retarded coordinate system in which the Minkowski-space scalar wave equation takes a simple form. The resulting form of the scalar wave equation on Minkowski is so simple and easy to work with that it cannot be entirely new. However, I am unable to find any evidence of it in the current literature of general relativity and feel that it should be more widely known. All of the expected formal advantages of such a coordinate system are realized explicitly in this example. Also,

the method of constructing this coordinate system and the properties of the resulting form of the spacetime metric may give clues to the use of ART coordinate systems in curved spacetimes.

The key to this particular asymptotically retarded (ART) coordinate system is the conformal transformation that one ultimately hopes to make. This conformal transformation turns out to contain the structure needed to perform the time coordinate transformation. Section 2 reviews the familiar conformal map of spheres in order to establish the notation that will be used in this paper. The transformation to asymptotically retarded time is performed in Sec. 3.

Section 4 performs the ART transformation of the wave equation in one space dimension. This example shows the essential features of the transformation. In order to establish that the simplicity of the transformation is not a peculiarity of $(1 + 1)$ -dimensional Minkowski space, Sec. 5 performs the ART transformation on the $3 + 1$ wave equation.

The problem for general relativistic applications is to decide which features of this example are peculiar to simple wave equations in Minkowski space and which can be generalized to curved spacetimes. Section 6 discusses this problem and lists a few properties which might be used to define ART transformations in more general situations.

2. NOTATION FOR CONFORMAL MAPS ONTO SPHERES

A Euclidean R^n space-coordinate manifold M is to be mapped onto an n -sphere S^n . Denote the space coordinates by X^j . It is convenient to represent the n -sphere as the surface

$$(x^1)^2 + (x^2)^2 + \dots + (x^n)^2 + w^2 = e^2,$$

in another Euclidean manifold m which I will call the "host manifold." The host manifold m is $(n + 1)$ -dimensional with coordinates x^j and w . The virtue of working in the host manifold is that its coordinate functions are smooth everywhere on the n -sphere. Thus, it provides a simple way to decide questions of differentiability.

Radius functions are defined in M and m according to

$$R = [(X^1)^2 + (X^2)^2 + \dots + (X^n)^2]^{1/2},$$

$$r = [(x^1)^2 + (x^2)^2 + \dots + (x^n)^2]^{1/2},$$

^{a)}Supported by the University Grants-in-aid Program for Faculty of Virginia Commonwealth University.

so that the n -sphere definition can be written as

$$r^2 + w^2 = e^2. \quad (2.1)$$

The point on the n -sphere with $w = -e$ will be called the *source-point* while the point with $w = e$ will be called the *far-field boundary*. The $n - 1$ subsphere with $w = 0$ is the *equator*.

In spherical polar coordinates, a point in M can be represented by the value of R and a set of polar angles Ω . Similarly, a point in m can be represented by values of r , w , and a set ω of polar angles which locate the point's projection in the $w = 0$ plane.

A conformal projection which takes the origin of the space-coordinate manifold into the source-point, infinity into the far-field boundary, and the sphere of radius $R = A$ into the equator of the n -sphere is given by the relations

$$\begin{aligned} r &= 2e \frac{R/A}{R^2/A^2 + 1}, \\ w &= e \frac{R^2/A^2 - 1}{R^2/A^2 + 1}, \\ \omega &= \Omega. \end{aligned} \quad (2.2)$$

The following consequences of Eq. (2.2) will be needed:

$$R/A = r/(e - w), \quad (2.3)$$

$$(R/A)^2 = (e + w)/(e - w), \quad (2.4)$$

$$(R/A)^2 + 1 = 2e/(e - w), \quad (2.5)$$

$$(R/A)^2 - 1 = 2w/(e - w), \quad (2.6)$$

$$eA\partial/\partial R = (e - w)(r\partial/\partial w - w\partial/\partial r). \quad (2.7)$$

A far-field polar angle θ will be defined by the relations

$$\begin{aligned} w &= e \cos \theta, \\ r &= e \sin \theta, \end{aligned} \quad (2.8)$$

so that the far-field boundary is at $\theta = 0$ while the source point is at $\theta = \pi$.

3. ASYMPTOTICALLY RETARDED TIME

Now add a time coordinate T to the space coordinates to form the spacetime coordinate manifold $M \times R$. The wave equation on $M \times R$ is

$$\partial^2 \Psi / \partial T^2 - \partial^2 \Psi / \partial^2 X^1 - \dots - \partial^2 \Psi / \partial^2 X^n = j, \quad (3.1)$$

where j is a source with compact support within a distance A of the origin in M .

A direct conformal mapping of the level surfaces of T into the sub-manifold $S^n \times R$ of $m \times R$ achieves a compact manifold on which to give initial conditions and evolve solutions. This type of transformation has been used with some success in the study of Einstein's equations.³ However, the solutions of the wave equation typically have zeros which collect arbitrarily close to the far-field point in this picture. Such solutions cannot be extended to regular functions on the compact n -ball bounded by the far-field boundary. Thus, conformal mapping alone does not secure the full advantages of dealing with functions on a compact manifold.

To obtain a picture in which retarded solutions are regular, take advantage of the fact that retarded wave solutions

have the general form

$$\Psi(X, T) = F(X, T - R),$$

where all of the rapid spatial dependence is in the second argument of F . Change to a new time coordinate t which is regular at the source point and goes asymptotically to the retarded time $T - R$ at great distance from the source. Once a suitable asymptotically retarded time coordinate t has been found, apply the conformal transformation to its level surfaces.

The choice of a new time coordinate must be made carefully if the extreme simplicity of Eq. (3.1) is not to be lost. To avoid obscuring the time-translation invariance of the equation, the new coordinate t must be linear in the old one. Thus, the candidates can be taken to have the form

$$At = T - Af(X), \quad (3.2)$$

where the constant A makes t and f dimensionless and the function f must go asymptotically to R/A at large values of R . To avoid conical space slices, the function f must have a gradient which vanishes at $R = 0$.

Equation (2.5) reveals that the desired function is already implicit in the conformal transformation. Thus, choose

$$f = [2e/(e - w)]^{1/2} = [(R/A)^2 + 1]^{1/2}. \quad (3.3)$$

With this choice, the level surfaces of the new time coordinate t are just the future hyperbolas given by

$$(T - At)^2 - R^2 = A^2. \quad (3.4)$$

The essential feature of this choice is that the hyperbolic radius A is the same as the radius parameter which enters into the conformal transformation. *This meshing of the time transformation with the conformal space transformation is essential if simple structures are to result when the transformations are combined.*

Now apply the conformal map to the level surfaces of t and drag the essential structures of $M \times R$ into $S^n \times R$. I will adopt the usual ambiguous notation in which the same symbol is used for an object and its image under a diffeomorphism. The $S^n \times R$ image of the derivative with respect to Minkowski coordinate time is evidently given by

$$eA\partial/\partial T = \partial/\partial t, \quad (3.5)$$

while the image of the derivative with respect to the Minkowski radius coordinate is found to be

$$eA\partial/\partial R = (e - w)J - (1/2)rf\partial/\partial t, \quad (3.6)$$

where

$$J = -\partial/\partial \theta = r\partial/\partial w - w\partial/\partial r. \quad (3.7)$$

The retarded time $T - R$ takes an especially simple form in the new coordinates. Using Eq. (2.4) for R and Eqs. (3.2) and (3.3) for T , one finds

$$T - R = t + \tan(\theta/4), \quad (3.8)$$

where θ is the far-field polar angle in S^n .

The expected form of a retarded wave can also be dragged over to the compactified picture. One expects these solutions to have the form

$$\Psi = F(\omega, \theta, t + \tan(\theta/4)). \quad (3.9)$$

This form shows the expected behavior: the waves move from the source point to the far-field boundary of the sphere with a nearly constant speed. Each wave departs through the far-field boundary a finite time after its emission from the source.

The function

$$\sigma = \tan(\theta/4), \quad (3.10)$$

plays a central role in the compactified picture of retarded waves. It ranges from a value of zero at the far-field point to a value of one at the source point. Because waves propagate at constant speed in this coordinate, I will call it the "wave-distance function."

The wave-distance function simplifies the compactified forms of several useful functions and differential operators. A bit of algebra and repeated use of the half-angle formulas yields the following relations:

$$2R/A = \sigma^{-1} - \sigma, \quad (3.11)$$

$$A(\partial/\partial T - \partial/\partial R) = 2(1 + \sigma^2)^{-1}(\partial/\partial t + \sigma^2\partial/\partial\sigma), \quad (3.12)$$

$$A(\partial/\partial T + \partial/\partial R) = 2\sigma^2(1 + \sigma^2)^{-1}(\partial/\partial t - \partial/\partial\sigma). \quad (3.13)$$

4. THE VIBRATING STRING COMPACTIFIED

The wave equation in one space dimension,

$$\partial^2\Psi/\partial T^2 - \partial^2\Psi/\partial X^2 = j(X,T),$$

can be split into two equations on the positive real line, one for the positive parity part and the other for the negative parity part:

$$j = j_+ + j_-, \quad \Psi = \Psi_+ + \Psi_-, \\ \partial^2\Psi_s/\partial T^2 - \partial^2\Psi_s/\partial R^2 = j_s, \quad s = + \text{ or } -. \quad (4.1)$$

In this form the equation can immediately be put into ART coordinates by using Eqs. (3.12) and (3.13). The resulting equation on $S^1 \times R$ is

$$4(1 + \sigma^2)^{-1}(\partial/\partial t + \sigma^2\partial/\partial\sigma)\sigma^2(1 + \sigma^2)^{-1}(\partial/\partial t - \partial/\partial\sigma)\Psi_s = A^2 j_s. \quad (4.2)$$

Here the space manifold has been mapped into half of the circle. The space coordinate σ ranges from zero at the far-field point to a value of one at the source point.

Although the ART form of the wave equation appears to be complicated at first glance, it turns out to be simple enough to be solved by standard techniques. From the factored form of Eq. (4.2) and the fact that the factors commute, the general solution outside the source is just

$$\Psi_s(\sigma, t) = F(t + \sigma) + G(t + 1/\sigma).$$

The retarded solution F remains regular near the far-field point while the advanced solution G is singular there. In these coordinates, the retarded characteristic surfaces are at 45° to the time axis while the characteristic surfaces for the advanced solutions accumulate at the far-field point.

For a periodic source, one's first idea of how to proceed is to seek solutions of the form

$$\Psi = a(\sigma)e^{i\omega(t + \sigma)}.$$

This idea works perfectly. Make this substitution into Eq. (4.2) and obtain a first order linear equation for the derivative a' . This equation reduces to quadrature by the standard for-

mula. Be careful to perform the integrals so that the boundary condition $a'(0) = 0$ is built into them. Otherwise, the advanced solutions will appear as essential singularities. Make use of the boundary condition at the source point (imposed by the parity of the wavefunction) in order to perform the final integration. I have not included the details of this procedure because it does not generalize to nonlinear wave equations such as the ones encountered in general relativity.

The procedure which is being tried in general relativity is direct numerical integration.⁴ To get Eq. (4.2) into a form which is suitable for this procedure, define the advanced field momentum

$$\Phi = (\partial\Psi/\partial t - \partial\Psi/\partial\sigma)/(1 + \sigma^2), \quad (4.3)$$

and replace the somewhat intimidating second order form of the wave equation which appears in Eq. (4.2) by a simple first order system:

$$\partial\Phi/\partial t + \sigma^2\partial\Phi/\partial\sigma + 2\sigma\Phi = A^2[(1 + \sigma^2)/(4\sigma^2)]j, \quad (4.4)$$

$$\partial\Psi/\partial t - \partial\Psi/\partial\sigma - (1 + \sigma^2)\Phi = 0. \quad (4.5)$$

Equations (4.4) and (4.5) can now be converted into difference equations by one of the standard schemes and evolved numerically, subject to the boundary conditions:

$$\Phi|_{\sigma=0} = 0, \quad \partial\Psi/\partial t - 2\Phi|_{\sigma=1} = 0, \text{ for even parity}, \quad (4.6)$$

and the same outgoing boundary condition with

$$\Psi|_{\sigma=1} = 0, \text{ for odd parity}. \quad (4.7)$$

Equations (4.4)–(4.7) offer several advantages for numerical evolution. A transient source which oscillates just a few times will produce wavefunctions with just a few zeros at any given ART. These waves will all disappear into the far-field boundary in a finite amount of coordinate time. Thus, one can achieve accurate results with relatively few grid points and the effects of artificial viscosity terms (required to suppress numerical instabilities) are limited by the finite evolution time.

The ability to follow waves all the way to future null infinity in a finite number of steps becomes extremely important when one applies the compactification procedure to a nonlinear wave equation. Because of the self-scattering of such waves, rigorous outgoing wave boundary conditions and precise definitions of the amount of radiation that is finally emitted can only be stated at future null infinity.

5. THE THREE-DIMENSIONAL WAVE EQUATION COMPACTIFIED

The three-dimensional wave equation in spherical coordinates,

$$[\partial^2/\partial T^2 - R^2(\partial/\partial R)R^2\partial/\partial R - R^{-2}L^2]\Psi = j,$$

can be rewritten in the form

$$(\partial/\partial T - \partial/\partial R)(\partial\psi/\partial T + \partial\psi/\partial R) - R^{-2}L^2\psi = Rj,$$

where

$$\psi = R\Psi.$$

Equations (3.11)–(3.13) may be used to convert this equation into the first-order system

$$4\partial\phi/\partial t - 4\sigma^2\partial\phi/\partial\sigma + 8\sigma\phi - (1 - \sigma^2)^{-1}L^2\psi = A^3\sigma^{-3}(1 - \sigma^2)^2j, \quad (5.1)$$

$$\partial\psi/\partial t = \partial\psi/\partial\sigma + (1 + \sigma^2)\phi, \quad (5.2)$$

with the boundary conditions

$$\phi|_{\sigma=0} = 0, \quad \partial\psi/\partial t - 2\phi|_{\sigma=1} = 0 \quad (5.3)$$

This system is nearly identical to the system that describes the (1 + 1)-dimensional wave equation. The only difference is the angular momentum term which introduces some coupling between the two first-order evolution equations. A numerical evolution requires one to lay out a grid of evaluation points on the 2-sphere in order to evaluate this term. All of the comments of the previous section apply to this form of the (3 + 1)-dimensional wave equation.

6. COORDINATE CONDITIONS FOR GENERAL RELATIVITY

In general relativity, where the coordinates evolve along with the field variables, a choice of coordinates means imposing restrictions or coordinate conditions on some of the variables. In order to use asymptotically retarded coordinate systems in general relativity, one must choose coordinate conditions which define them without restricting the geometry of spacetime.

The problem of choosing coordinate conditions which are suitable for numerical evolution schemes is only now being resolved for the familiar Minkowski-like coordinate systems.^{2,4} For the unfamiliar ART coordinates it is much too soon to advocate a particular set of coordinate conditions. However, it may be useful to consider the properties of the simple ART coordinate system of Minkowski space.

Because the constant-time surfaces are hyperbolas in Minkowski space, they have constant extrinsic scalar curvature. In the usual notation,

$$\text{Tr}K = k = -3/A, \quad (6.1)$$

where K is the second fundamental form of a constant-time surface. This type of coordinate condition has been studied extensively by York and others and has been suggested as a way of realizing ART coordinates.²

The inverse spacetime metric tensor in ART coordinates is

$$g^{-1} = (2\sigma/A)^2 [(1 + \sigma^2)^{-2}n \times l - (1 - \sigma^2)^{-2}g^{-1}_{S^2}], \quad (6.2)$$

where

$$n = \partial/\partial t + \sigma^2\partial/\partial\sigma, \quad l = \partial/\partial t - \partial/\partial\sigma, \quad (6.3)$$

and

$$g^{-1}_{S^2},$$

is the inverse metric on the unit 2-sphere.

Equations (6.2) and (6.3) show that this example lends itself to a null-tetrad or spinor formalism. Such a formalism is appropriate near future null infinity but may not be very useful near a complicated source which admits no preferred null directions.

Most numerical evolutions of Einstein's equations have adopted a 3 + 1 splitting of spacetime properties. The spatial metric and second fundamental form as well as the lapse and shift function are usually the values that the computer evolves. In this example, the lapse function is found to be

$$N = A(\sigma^{-1} + \sigma)/2, \quad (6.4)$$

while the shift vector has the single nonzero component

$$N^1 = 1 - \sigma^2. \quad (6.5)$$

The second fundamental form or extrinsic curvature tensor of the constant-time surfaces is just the extrinsic curvature of a hyperbolic spacelike surface in Minkowski spacetime:

$$K^{ij} = -A^{-1}g^{ij}, \quad \text{Tr}K = -3/A. \quad (6.6)$$

The simplest ART coordinate condition that one can impose just fixes the lapse and shift functions to be those given by Equations (6.3) and (6.4). This simple approach will encounter focusing problems. The constant-position lines in such a coordinate system will intersect one another just as they do in hypersurface-orthogonal geodesic coordinates. However, because ART coordinates show outgoing waves reaching infinity in a finite coordinate time (which can be comparable to the wave crossing-time of the source), this focusing problem may not be as serious as it is in the usual, asymptotically Minkowski coordinates.

¹The earliest reference to this idea that I know of is W. C. Hernandez Jr. and C. W. Misner, *Astrophys. J.*, **143**, 452-464 (1966).

²Larry Smarr and James W. York, *Phys. Rev. D* **17**, 2529-2551 (1978).

³Abhay Ashtekar, "Asymptotic Structure of the Gravitational Field at Spatial Infinity" in *General Relativity and Gravitation*, Vol. 2 edited by A. Held (Plenum, New York, 1980), pp. 37-68; Abhay Ashtekar and R. O. Hansen, *J. Math. Phys.* **19**, 1542-1566 (1978). These papers cite a large number of additional references to earlier work.

⁴Larry Smarr, "Basic Concepts in Finite Differencing of Partial Differential Equations" and "Gauge Conditions, Radiation Formulae and the Two Black Hole Collision," in *Sources of Gravitational Radiation*, edited by Larry Smarr (Cambridge University, Cambridge, 1979), pp. 139-160 and 245-274.

A further note on the Hénon–Heiles problem

P. G. L. Leach

Department of Applied Mathematics, La Trobe University, Bundoora, 3083, Australia

(Received 12 November 1979; accepted for publication 14 March 1980)

The method of the Lie theory of extended groups is applied to the Hénon–Heiles problem. Only one generator for a one-parameter group is found. The corresponding first integral is the energy. It is inferred that no other exact integral exists.

PACS numbers: 03.20 + i

I. INTRODUCTION

The Hénon–Heiles problem¹ concerns the Hamiltonian

$$H = \frac{1}{2}(p_1^2 + p_2^2 + q_1^2 + q_2^2) + q_1^2 q_2 - \frac{1}{3}q_2^3 \quad (1.1)$$

($p_1 = \dot{q}_1, p_2 = \dot{q}_2$). It has been posed as a model for the motion of a galactic cluster. Computer analysis of the problem has suggested that for sufficiently small values of the energy, there exists a first integral independent of the energy. This has been termed the third integral (the first being the energy, the second being the angular momentum of the total system of which the Hamiltonian above is part). The tantalizing suggestiveness of the numerical results has led to much effort to find the third integral. It would be fair to say that progress has not been great and that the problem remains intractable. Indeed the term “notorious” has been applied to the problem,² a comment which doubtless stems from the sense of frustration produced.

In this note another method of attack is employed, that of the Lie theory of extended groups. Recently it has been used with considerable success on linear systems³ and with some success on a nonlinear time-dependent system.⁴ In the latter instance one of the chief results is that the possible existence of an invariant is determined by the nature of the time dependence.⁵

Before commencing the analysis, we give a brief summary of the method. Given a system of Newtonian equations of motion

$$N(\ddot{\mathbf{q}}, \dot{\mathbf{q}}, \mathbf{q}, t) = \mathbf{0}, \quad (1.2)$$

the system admits a one-parameter Lie group with generator

$$G(\mathbf{q}, t) = \xi(\mathbf{q}, t) \partial / \partial t + \boldsymbol{\eta}(\mathbf{q}, t) \cdot \nabla_{\mathbf{q}} \quad (1.3)$$

provided

$$G^{(2)}N = \mathbf{0} \quad (1.4)$$

whenever Eq. (1.2) is satisfied. $G^{(2)}$ is the second extension of G and is given by

$$G^{(2)} = G + \boldsymbol{\eta}^{(1)} \cdot \nabla_{\dot{\mathbf{q}}} + \boldsymbol{\eta}^{(2)} \cdot \nabla_{\ddot{\mathbf{q}}}, \quad (1.5)$$

where

$$\boldsymbol{\eta}^{(1)} = \dot{\boldsymbol{\eta}} - \dot{\xi} \dot{\mathbf{q}}, \quad \boldsymbol{\eta}^{(2)} = \ddot{\boldsymbol{\eta}} - \ddot{\xi} \dot{\mathbf{q}} - 2\dot{\xi} \ddot{\mathbf{q}}. \quad (1.6)$$

If such a generator exists, there exists a corresponding first integral, $I(\mathbf{q}, \dot{\mathbf{q}}, t)$, which is constructed by applying the double requirement that

$$G^{(1)}I = 0, \quad DI = 0, \quad (1.7)$$

where D represents the total time derivative. If more than

one generator of a one-parameter Lie group exists, there may be more than one first integral, although it does not follow automatically. As a trivial counterexample, the one-dimensional free particle has eight linearly independent generators; yet only three linearly independent first integrals are obtained.

We mention that the Lie theory is more general than Noether's theorem⁶ in its conception. The generators for the latter constitute a subset of those of the former. This distinction applies to nonlinear as well as to linear systems. An excellent example of this distinction is seen in the treatment of the classical Kepler problem by Prince and Eliezer.⁷

2. THE FORM OF THE GENERATORS

Applying the Lie method leads to sufficient complexity in the case of one-dimensional systems, let alone in systems of higher order, to warrant the determination of the permissible form of the generators for a given type of Newtonian equation before considering a particular problem. Suppose the system has Newtonian equations of the form

$$\ddot{q}_l + f_l(q, t) = 0, \quad l = 1, n. \quad (2.1)$$

Adopting the usual convention of summation on repeated indices, the twice-extended generator is

$$G^{(2)} = \xi \partial / \partial t + \eta_i \partial / \partial q_i + (\dot{\eta}_i - \dot{\xi} \dot{q}_i) \partial / \partial \dot{q}_i + (\ddot{\eta}_i - \ddot{\xi} \dot{q}_i - 2\dot{\xi} \ddot{q}_i) \partial / \partial \ddot{q}_i. \quad (2.2)$$

Applying this to Eq. (2.1) and separating out the terms which are of second and third order in the velocities, we have

$$\dot{q}_i \dot{q}_j \partial^2 \xi / \partial q_i \partial q_j = 0 \quad (2.3)$$

$$\dot{q}_i \dot{q}_j \partial^2 \eta_i / \partial q_i \partial q_j - 2 \dot{q}_i \dot{q}_j \partial^2 \xi / \partial q_i \partial t = 0. \quad (2.4)$$

From Eq. (2.3) it is apparent that

$$\xi(q, t) = a(t) + b_i(t) q_i. \quad (2.5)$$

Substituting this into Eq. (2.4), we have

$$\dot{q}_i \dot{q}_j \partial^2 \eta_i / \partial q_i \partial q_j = 2 \dot{q}_i \dot{q}_j \dot{b}_i. \quad (2.6)$$

Differentiating with respect to \dot{q}_m and \dot{q}_n in turn and assuming that η_i is sufficiently regular for the order of differentiation to be immaterial, we have

$$\partial^2 \eta_i / \partial q_m \partial q_n = \delta_{im} \dot{b}_n + \delta_{in} \dot{b}_m. \quad (2.7)$$

It then follows that

$$\eta_i(\mathbf{q}, t) = \dot{b}_k q_k q_i + c_{ik}(t) q_k + d_i(t). \quad (2.8)$$

Thus for a system of Newtonian equations of the type given

by Eq. (2.1), a generator of one-parameter Lie group has the form

$$G(\mathbf{q}, t) = (a + b_k q_k) \partial / \partial t + (b_k q_k q_l + c_{lk} q_k + d_l) \partial / \partial q_l, \quad (2.9)$$

where a , b , c , and d are functions of time to be determined according to the particular functions $f_i(\mathbf{q}, t)$.

3. EQUATIONS DETERMINING THE TIME-DEPENDENT FUNCTIONS

As the algebra involved in the determination of the functions a , b , c , and d tends to be messy no matter what the system may be, it is just as easy to consider a general two-dimensional system of which the Hénon–Heiles problem is a particular case. We take the Hamiltonian to be

$$H = \frac{1}{2}(p_1^2 + p_2^2 + q_1^2 + q_2^2) + Aq_1^3 + Bq_1^2 q_2 + Cq_1 q_2^2 + Dq_2^3, \quad (3.1)$$

the Hénon–Heiles case being given by $A = 0 = C$, $B = 1$, $D = -\frac{1}{2}$. The two Newtonian equations corresponding to Eq. (3.1) may be written as

$$\ddot{q}_l + q_l + f_{mn}{}^l q_m q_n = 0, \quad l = 1, 2, \quad (3.2)$$

where $f_{mn}{}^l$ is symmetric in the indices m and n and the repeated indices are summed now over 1 and 2 only. Applying the second extension of the operator given by Eq. (2.9) to Eq. (3.2), the coefficients of the terms of second and third order in the velocities vanish identically. The terms linear in the velocities are

$$3\ddot{b}_k(\dot{q}_k q_l + q_k \dot{q}_l) + 2\dot{c}_{lk} \dot{q}_k - \ddot{a} \dot{q}_l - b_k \dot{q}_l (q_k + f_{mn}{}^k q_m q_n) - 2b_k \dot{q}_k (q_l + f_{mn}{}^l q_m q_n) = 0. \quad (3.3)$$

From the coefficients of the second order terms in the displacements, it is obvious that

$$b_1 \equiv 0, \quad b_2 \equiv 0. \quad (3.4)$$

From the terms now remaining, it follows that

$$2c_{ij} = \dot{a} \delta_{ij} + \alpha_{ij} \quad (3.5)$$

where the four α_{ij} are as yet arbitrary constants.

Turning now to be velocity independent terms, those independent of the coordinates yield

$$\ddot{d}_l + d_l = 0, \quad (3.6)$$

those linear in the coordinates give

$$2d_m q_n f_{mn}{}^l + \dot{c}_{lm} q_m + 2\dot{a} q_l = 0, \quad (3.7)$$

and the second order terms are

$$2\dot{a} q_m q_n f_{mn}{}^l - c_{lk} q_m q_n f_{mn}{}^k + c_{mk} q_k q_n f_{mn}{}^l + c_{nk} q_k q_m f_{mn}{}^l = 0. \quad (3.8)$$

Differentiating Eq. (3.8) with respect to q_i and q_j in succession and making use of Eq. (3.5),

$$S\dot{a} f_{ij}{}^l + \alpha_{ki} f_{kj}{}^l + \alpha_{kj} f_{ki}{}^l - \alpha_{lk} f_{ij}{}^k = 0. \quad (3.9)$$

4. SOME POSSIBLE GENERATORS

From Eq. (3.6) it is evident that

$$d_l = E_l \sin t + F_l \cos t, \quad l = 1, 2. \quad (4.1)$$

Substituting for $c(t)$ in Eq. (3.7), four equations result, viz.

$$\ddot{a} + 4\dot{a} + 4d_m f^1{}_{m1} = 0, \quad (4.2)$$

$$\ddot{a} + 4\dot{a} + 4d_m f^2{}_{m2} = 0, \quad (4.3)$$

$$d_m f^1{}_{m2} = 0, \quad d_m f_{m1}{}^2 = 0. \quad (4.4)$$

The pair of equations in (4.4) are identical and, in terms of the coefficients in Eq. (3.1), are

$$Bd_1 + Cd_2 = 0. \quad (4.5)$$

A consistency condition between Eqs. (4.2) and (4.3) (in the case $\dot{a}(t) \neq 0$) requires

$$(3A - C)d_1 + (B - 3D)d_2 = 0. \quad (4.6)$$

If $\dot{a}(t) \equiv 0$, Eqs. (4.2) and (4.3) require that

$$3Ad_1 + Bd_2 = 0, \quad Cd_1 + 3Dd_2 = 0. \quad (4.7)$$

For the moment let us confine our attention to the case for which $a(t)$ is a constant. There exist three relations, Eqs. (4.5) and (4.7), between d_1 and d_2 . For these to be consistent A , B , C , and D are related by

$$B^2 = 3AC, \quad C^2 = 3BD, \quad BC = 9AD. \quad (4.8)$$

To within a scaling constant, possible values which the coefficients may take are

$$A \quad B \quad C \quad D \quad (4.9a)$$

$$1 \quad 0 \quad 0 \quad 0 \quad (4.9b)$$

$$0 \quad 0 \quad 0 \quad 1 \quad (4.9c)$$

$$1 \quad \pm 3x \quad 3x^2 \quad \pm x^3 \quad (4.9d)$$

$$-1 \quad \pm 3x \quad -3x^2 \quad \pm x^3 \quad (4.9e)$$

where x is a positive constant. The generators for such systems are

$$G_1 = \partial / \partial t, \quad (4.10)$$

corresponding to $a(t)$ constant and

$$G_2 = \sin t \partial / \partial q_2, \quad G_3 = \cos t \partial / \partial q_2$$

$$G_2 = \sin t \partial / \partial q_1, \quad G_3 = \cos t \partial / \partial q_1$$

$$G_2 = \sin t (\pm x \partial / \partial q_1 - \partial / \partial q_2),$$

$$G_3 = \cos t (\pm x \partial / \partial q_1 - \partial / \partial q_2)$$

$$G_2 = \sin t (\pm x \partial / \partial q_1 + \partial / \partial q_2),$$

$$G_3 = \cos t (\pm x \partial / \partial q_1 + \partial / \partial q_2), \quad (4.11)$$

corresponding to (4.9a) to (4.9d), respectively.

First integrals for the systems above are easily constructed by using the result that if I is a first integral, then so also is $G^{(1)}I$. Taking the first of (4.9), the energy is

$$E = \frac{1}{2}(\dot{q}_1^2 + \dot{q}_2^2 + q_1^2 + q_2^2) + Aq_1^3, \quad (4.12)$$

$$I_1 = G_2^{(1)}E = q_2 \sin t + \dot{q}_2 \cos t, \quad (4.13)$$

$$I_2 = G_3^{(1)}E = q_2 \cos t - \dot{q}_2 \sin t. \quad (4.14)$$

In the case of the third of (4.9),

$$E = \frac{1}{2}(\dot{q}_1^2 + \dot{q}_2^2 + q_1^2 + q_2^2) + A(q_1 \pm xq_2)^3, \quad (4.15)$$

$$I_1 = \sin t (\pm xq_1 - q_2) + \cos t (\pm x\dot{q}_1 - \dot{q}_2), \quad (4.16)$$

$$I_2 = \cos t (\pm xq_1 - q_2) - \sin t (\pm x\dot{q}_1 - \dot{q}_2), \quad (4.17)$$

The resemblance of Eqs. (4.16) and (4.17) to Eqs. (4.13) and

(4.14) is not accidental. Consider the transformation

$$\bar{t} = t \quad Q_1 = q_1 \pm xq_2, \quad Q_2 = \pm xq_1 + q_2. \quad (4.18)$$

Under this transformation the generators become

$$G_1 = \partial/\partial t, \quad (4.19)$$

$$G_2 = -(1+x^2)\sin t \partial/\partial Q_2,$$

$$G_3 = -(1+x^2)\cos t \partial/\partial Q_2, \quad (4.20)$$

and the energy is

$$E = \frac{1}{2}(1+x^2)(\dot{Q}_1^2 + \dot{Q}_2^2 + Q_1^2 + Q_2^2) + AQ_1^3. \quad (4.21)$$

The transformation has separated the system into two uncoupled parts so that four linearly independent first integrals exist. The integrals linear in the coordinate and velocity correspond to the harmonic oscillator part of Eq. (4.21). They represent $Q_2(0)$ and $\dot{Q}_2(0)$. That similar integrals do not exist for Q_1 is not surprising in view of the quadrature required to express Q_1 as a function of time.

The generators discussed above for the case when $a(t)$ is a constant do not apply to the Hénon–Heiles problem since the coefficients for that problem do not fit in with the scheme in (4.9). Allowing $a(t)$ to be not constant is of no use for the Hénon–Heiles problem as the consistency condition of Eq. (4.6) is not satisfied. Moreover, returning to Eq. (3.9), the admissible time-dependent forms of $a(t)$, which involve sines and cosines [cf. Eq. (4.2)] can only occur when all the f_{mn}^l are zero. The last remaining source of a generator is to be found in the α part of $c(t)$. From Eq. (3.9) we see that the α 's must satisfy

$$\alpha_{ki}f_{kj}^l + \alpha_{kj}f_{ki}^l - \alpha_{lk}f_{ij}^k = 0. \quad (4.22)$$

In terms of the coefficients in Eq. (3.1) the conditions in equation (4.22) constitute the system of equations

$$\begin{bmatrix} 3A & -B & 2B & 0 \\ 0 & 3A-C & C & B \\ 2B & 0 & 2C-3A & -B \\ -C & 2B-3D & 0 & 2C \\ C & B & 3D-B & 0 \\ 0 & 2C & -C & 3D \end{bmatrix} \begin{bmatrix} \alpha_{11} \\ \alpha_{12} \\ \alpha_{21} \\ \alpha_{22} \end{bmatrix} = 0. \quad (4.23)$$

Unfortunately, when the Hénon–Heiles values $A = 0 = C$, $B = 1$, $D = -\frac{1}{3}$ are substituted, all of the α 's must be zero and so no integral can arise from this source.

For the possible cases listed in (4.9), α 's do exist. In corresponding order they are

$$\begin{array}{cccc} \alpha_{11} & \alpha_{12} & \alpha_{21} & \alpha_{22} \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & \pm x & \pm x & 1 \\ 1 & \pm x & \pm x & 1 \end{array} \quad (4.24)$$

For example, the integral corresponding to the first case is

$$I_3 = t - \arctan q_2/\dot{q}_2. \quad (4.25)$$

Finally we note for other values of the coefficients A , B , C , and D , α -based generators do not exist.

5. DISCUSSION

We have seen that the Hénon–Heiles problem gives rise to only one one-parameter symmetry group. Consequently, any first integral of the motion must satisfy the requirement that

$$G^{(1)}(\mathbf{q}, \dot{\mathbf{q}}, t)I(\mathbf{q}, \dot{\mathbf{q}}, t) = 0, \quad (5.1)$$

where $G^{(1)}$ is the first extension of the generator G . In this case G is simply the generator of time translations and so

$$G^{(1)} = G = \partial/\partial t. \quad (5.2)$$

Thus the invariant satisfies the equation

$$\partial I(\mathbf{q}, \dot{\mathbf{q}}, t)/\partial t = 0. \quad (5.3)$$

In the formulation of this problem the velocity and momentum are identical, and we may rewrite (5.3) as

$$\partial I(\mathbf{q}, \mathbf{p}, t)/\partial t = 0, \quad (5.4)$$

from which it follows that I is a function of the canonical variables only. Since the invariant, now $I(\mathbf{q}, \mathbf{p})$, has zero total time-derivative,

$$[I, H]_{PB} = 0, \quad (5.5)$$

i.e., it has zero Poisson bracket with the Hamiltonian. We have already seen in an earlier paper⁸ that the only time-independent invariant with this property is the Hamiltonian itself.

Those who are familiar with the application of the Lie theory to linear systems will know that there exist generators for which the corresponding invariants contain time explicitly. In particular there are invariants corresponding to the initial conditions of the motion. Such invariants do not arise in the case of the Hénon–Heiles problem, nor is this lack of occurrence peculiar to it. A similar situation applied to the one-dimensional anharmonic oscillator⁴ and to the Kepler problem.⁹ We surmise that the apparent absence of initial condition type invariants for nonlinear systems may be due to nonlinear operations being required to invert expressions for variables to expressions for constants (of integration).

6. CONCLUSION

The Lie method has given a negative answer to the question of the existence of an integral other than the energy for the Hénon–Heiles problem. It is known that a formal integral exists for which various expansion techniques are available.¹⁰ The result of our investigations suggests that such series are indeed formal. The Lie method, being based on the Newtonian equations of motion, provides the largest possible set of generators of dynamical symmetries. For the Hénon–Heiles problem only one generator exists, and hence there is only one exact first integral, the energy (\equiv the Hamiltonian).

What then is to become of the Hénon–Heiles problem? In spite of the contrary evidence presented here, there is still the fact that the system does possess remarkable regularity at low energies. This suggests that the formal integral is a reasonable approximation for small enough values of the variables. It may be possible to use the idea of an approximate symmetry to construct a corresponding approximate integral. This is a matter for future investigation.

ACKNOWLEDGMENTS

I am indebted to G. E. Prince of La Trobe for stimulating discussions on the subject of the Lie method and for providing preprints of his work and to Professor A. Giorgilli of Milan, whose query regarding my earlier work on this problem⁸ prompted me to try again.

¹The literature is extensive. The original paper was M. Hénon and C. Heiles, *Astron. J.* **69**, 73–9 (1964). For a recent survey see R. C. Churchill, G. Pecelli, and D. L. Rod, "A survey of the Hénon–Heiles Hamiltonian with applications to related examples," in *Como Conference Proceedings on Stochastic Behaviour in Classical and Quantum Mechanical Systems* (Springer, New York, to appear). This contains a comprehensive set of references.

²T. Bountis, "Non-linear models in dynamics and statistical mechanics," Ph.D. thesis (University of Rochester, 1978).

³P. G. L. Leach, *J. Math. Phys.* (to appear); G. E. Prince, *J. Phys. A* (to appear); P. G. L. Leach, *J. Austr. Math. Soc. Ser. B* (to appear).

⁴P. G. L. Leach, "An exact invariant for a class of time-dependent anharmonic oscillators with cubic anharmonicity," Research Report AM79:10 (Department of Applied Mathematics, La Trobe University, 1979).

⁵The same point is made in a different context in W. Sarlet and L. Y. Bahar, "A direct construction of first integrals for certain non-linear dynamical systems," preprint (Department of Mechanical Engineering and Mechanics, Drexel University).

⁶The particular version referred to here is the one used for example by M. Lutzky, *J. Phys. A* **11**, 249–58 (1978) and not the one involving velocity-dependent transformations found for example in D. S. Djukic, *Arch. Mech. Stosow.* **26**, 243–9 (1974).

⁷G. E. Prince and C. J. Eliezer, "On the Lie symmetries of the classical Kepler problem," Research Report AM79:06 (Department of Applied Mathematics, La Trobe University, 1979)

⁸P. G. L. Leach, *J. Math. Phys.* (to appear).

⁹P. G. L. Leach, "Applications of the Lie theory of extended groups in Hamiltonian mechanics: the oscillator and the Kepler problem," Research Report AM80:01 (Department of Applied Mathematics, La Trobe University, 1980).

¹⁰A. Giorgilli and L. Galgani, *Celest. Mech.* **17**, 267–80 (1978) and the references cited therein. Also A. Giorgilli (private communication 18 June 1979).

Further comments on the behavior of acceleration waves of arbitrary shape

V. D. Sharma^{a)} and V. V. Menon^{b)}

Department of Aerospace Engineering, University of Maryland, College Park, Maryland 20742

(Received 25 February 1980; accepted for publication 12 December 1980)

Converging waves with nonzero initial critical amplitude are completely characterized. It is shown that for a converging wave a necessary and sufficient condition for the initial critical amplitude to be zero is that the converging wave is spherical.

PACS numbers: 03.40.Kf

I. INTRODUCTION

Bowen and Chen¹ have discussed the growth and decay behavior of converging and diverging waves, and they have completely characterized diverging waves with nonzero initial critical amplitude; but they do not refer to the corresponding characterization for converging waves. The purpose of the present paper is to characterize these converging waves completely. It is shown that all spherically converging waves will grow into shock waves before the formation of the focus, no matter how small their initial amplitude.

II. BEHAVIOR OF CONVERGING WAVES

The differential equation governing the amplitude $a(t)$ of an acceleration wave propagating into a homogeneous material medium, assumed to be at rest initially, is of the form¹

$$\frac{da}{dt} = -(\mu_0 - \frac{1}{2}u_n\bar{K})a + \beta_0 a^2, \quad (2.1)$$

where μ_0 is a constant depending on the type of material under study and the uniform conditions prevailing ahead of the wave, β_0 is a nonzero constant depending solely on the elastic response of the material, u_n is the constant normal speed (taken to be positive), \bar{K} is the mean curvature of the wavefront at any time t expressed as

$$\bar{K} = (\bar{K}_0 - 2K_0u_n t)/(1 - \bar{K}_0u_n t + K_0u_n^2 t^2), \quad (2.2)$$

where $\bar{K}_0 = k_1 + k_2$ is the initial mean curvature and $K_0 = k_1 k_2$ is the initial total curvature with k_1 and k_2 being the initial principal curvatures. When k_1 and k_2 are both nonpositive, the wave is divergent; and when one or both the initial principal curvatures are positive, the wave is convergent. In the following discussion, we shall consider only converging waves.

The solution of (2.1) in view of (2.2) can be written as

$$a(t) = [I_1(t)\exp(-\mu_0 t)]/[1/a(0) - \beta_0 I_2(t)], \quad (2.3)$$

where $a(0) \neq 0$ is the initial amplitude, and the functions $I_1(t)$ and $I_2(t)$ are given by

$$I_1(t) = \{(1 - k_1 u_n t)(1 - k_2 u_n t)\}^{-1/2}, \quad (2.4)$$

$$I_2(t) = \int_0^t \{(1 - k_1 u_n \tau)(1 - k_2 u_n \tau)\}^{-1/2} \exp(-\mu_0 \tau) d\tau. \quad (2.5)$$

In the analysis of converging waves, where *at least one of the initial principle curvatures is positive*, the integral $I_2(t^*)$ plays a crucial role, where t^* is the smallest positive root of the equation $(1 - k_1 u_n t^*)(1 - k_2 u_n t^*) = 0$. One can easily show that the integral $I_2(t^*)$ in (2.5) is infinite if and only if both k_1 and k_2 are positive and equal, i.e., $k_1 = k_2 > 0$. For, if $k_1 = k_2 > 0$, by substituting $z = t^* - t$, we find that the singularity $t \rightarrow t^*$ in the integrand of $I_2(t^*)$ is of the form $z^{-1}\phi(z)$ as $z \rightarrow 0$, where $\phi(z)$ is both bounded and bounded away from zero, and the function z^{-1} is not integrable on any interval $[0, T]$, $T > 0$. If $k_1 \neq k_2$ and at least one of k_1, k_2 is positive, then the integral $I_2(t^*)$ is finite; this follows from the argument that the singularity as $t \rightarrow t^*$ in the integrand of $I_2(t^*)$ is of the type $z^{-1}\psi(z)$ as $z \rightarrow 0$, where the function $\psi(z)$ is again bounded and bounded away from zero, and the function z^{-1} is integrable over every interval $[0, T]$, $T > 0$.

Thus for converging waves, irrespective of the sign of μ_0 , we have the following two situations:

(i) $k_1 \neq k_2$ and *at least one of them is positive*: in this case, when $\text{sgn}a(0) = \text{sgn}\beta_0$, it follows from (2.3) that *not all* converging waves will grow into shock waves, i.e., there exists a critical value of the initial wave amplitude, given by

$$\gamma = [|\beta_0|I_2(t^*)]^{-1}, \quad (2.6)$$

such that waves with initial amplitude less than γ form a focus (i.e., $|a(t)| \rightarrow \infty$ as $t \rightarrow t^*$), waves with initial amplitude greater than γ form a shock before the focus (i.e., there exists a positive $\hat{t} (< t^*)$ given by $I_2(\hat{t}) = [|\beta_0 a(0)]^{-1}$, such that $|a(t)| \rightarrow \infty$ as $t \rightarrow \hat{t}$), and waves with initial amplitude equal to γ form a shock and focus simultaneously (i.e., $\hat{t} = t^*$ and $|a(t)| \rightarrow \infty$ as $t \rightarrow t^*$).

(ii) $k_1 = k_2 > 0$: in this case, which corresponds to a spherically converging wave, the integral $I_2(t^*)$ is infinite and thus, the initial critical amplitude given by (2.6) vanishes; further, it follows from (2.3) that when $\text{sgn}a(0) = \text{sgn}\beta_0$, there exists a positive $\tilde{t} < t^*$, given by

$$\int_0^{\tilde{t}} (1 - k_1 u_n \tau)^{-1} \exp(-\mu_0 \tau) d\tau = 1/|\beta_0 a(0)|,$$

such that as t approaches \tilde{t} , the denominator of (2.3) vanishes, whereas the numerator remains finite, i.e., $|a(t)| \rightarrow \infty$ as $t \rightarrow \tilde{t}$. This means that all spherically converging waves will grow into shock waves before the formation of the focus, no matter how small be their initial amplitude. This result is, in

^{a)}On leave from Applied Mathematics Section, I. T., B. H. U., India.

^{b)}Applied Mathematics Section, I. T., B. H. U., India.

effect, that for a converging wave a necessary and sufficient condition for the initial critical amplitude to be zero is for the converging wave to be spherical.

It is interesting to note that when $\text{sgn}a(0) = -\text{sgn}\beta_0$, the denominator in (2.3), in both the situations mentioned above, is always bounded away from zero, and $|a(t)| \rightarrow \infty$ as $t \rightarrow t^*$, i.e., in this case all converging waves form a focus only.

ACKNOWLEDGMENTS

The authors are thankful to the referee for making certain points more explicit.

Research support to V.D.S. from the Ministry of Education, Govt. of India, is gratefully acknowledged.

¹R. M. Bowen and P. J. Chen, *J. Math. Phys.* **13**, 948 (1972).

Functional methods in random classical field theory

Jorge F. Willemsen

Schlumberger-Doll Research Center, Ridgefield, Connecticut 06877

(Received 15 July 1980; accepted 31 October 1980)

A functional which generates N distinct point solutions to a classical wave equation with random coefficients in the presence of external sources is constructed. Statistical averaging over the random coefficients is then implemented using replica and/or anticommuting field techniques.

PACS numbers: 03.50.De, 03.40.Kf

I. INTRODUCTION

Consider the wave equation

$$\left[\frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial \mathbf{x}^2} \right] \varphi(\mathbf{x}, t) = 0. \quad (1.1)$$

In many applications, the velocity of propagation, c , is adequately considered a fixed parameter determined from average physical characteristics of the medium supporting the wave motion. There are many interesting situations, however, in which the medium may appropriately be characterized in terms of physical characteristics which fluctuate on a small macroscopic scale.¹

The existence of such fluctuations implies that the speed " c " cannot be treated as a fixed parameter of the system. In this situation it is fruitful to write an index of refraction which contains a part characteristic of the large-scale average medium, and a second part which may be considered a random variable.

The problem of wave propagation in a fluctuating medium thus becomes one of solving an ensemble averaged version of Eq. (1.1),

$$\left[\frac{\partial^2}{\partial t^2} - c_0^2 \frac{\partial^2}{\partial \mathbf{x}^2} \right] \langle \varphi(\mathbf{x}, t) \rangle = \langle n(\mathbf{x}, t) \frac{\partial^2 \varphi}{\partial \mathbf{x}^2} \rangle. \quad (1.2)$$

Of course each element of the ensemble over which one is averaging is to be considered a unique, deterministic system. A wave will propagate in a particular sample of the fluctuating medium in a way which depends upon the precise manner in which the fluctuations occur in the particular sample. The real problem, then, is to determine not only how an average over propagations in samples of the medium behaves, but to infer how closely propagation in a particular sample resembles propagation in the averaged medium. In short, one must examine not only $\langle \varphi \rangle$, but higher moments as well.²

In this paper, we formulate the problem of computing the distinct N -point ensemble averaged correlations

$$\langle \Phi(1 \dots M; M+1 \dots N) \rangle \equiv \langle \varphi^*(\mathbf{x}_1, t_1) \dots \varphi(\mathbf{x}_{M+1}, t_{M+1}) \dots \rangle \quad (1.3)$$

in terms of generating functionals.^{3,4} In this manner, the averaging problem takes a form which is very familiar in the statistical mechanics of, e.g., quenched impurities and spin-glasses.⁵ The averaging procedure can then be implemented by borrowing techniques such as the replica method which have been successful in the study of condensed matter.^{6,7}

II. GENERATING FUNCTIONAL

For the sake of generality, consider the linear second order differential equation which follows from the stationarity requirement

$$\delta S[\varphi] = 0, \quad (2.1)$$

$$S = \int d^d x dt [\varphi^*(\mathbf{x}, t) L(\mathbf{x}, t) \varphi(\mathbf{x}, t) + J \varphi^* + J^* \varphi],$$

where the operator $L(\mathbf{x}, t)$ in d spatial dimensions contains coefficients which may be functions of spatial coordinates. In this equation, $J(\mathbf{x}, t)$ is a prescribed source, and we are interested in causal solutions to the equation.

In a standard manner, real sources give rise to real solutions of Eq. (2.1), but we shall, for convenience, more generally consider complex representations for the sources. Then we must solve for complex φ , and extract the physically meaningful real part at the end of the calculation.

By virtue of Eq. (2.1), it is clear that if the points (\mathbf{x}, t) are all distinct, the product of solutions, denoted Φ , satisfies the following equation:

$$\hat{L}(i) \Phi(1 \dots M; M+1 \dots N) = J(i) \Phi(1 \dots \bar{i} \dots M; \dots N) + J^*(i) \Phi(1 \dots M; \dots \bar{i} \dots N). \quad (2.2)$$

We employ a notation for functions of indexed variables $\psi(\mathbf{x}_i, t_i) \equiv \psi(i)$. The symbol \bar{i} indicates the i th argument is absent. We shall restrict our considerations to the case of distinct points.

Let us now attempt to construct a functional which generates the quantities Φ subject to the equation (2.2). We seek our generating functional in the form

$$\Phi = \frac{1}{i} \frac{\delta}{\delta j(1)} \dots \frac{1}{i} \frac{\delta}{\delta j(M)} \frac{1}{i} \frac{\delta}{\delta j^*(M+1)} \dots \frac{1}{i} \frac{\delta}{\delta j^*(N)} \mathcal{F} \Big|_{j=j^*=0} \quad (2.3)$$

Clearly the auxiliary currents $j(\mathbf{x}, t)$ are fictitious functions which cannot appear in any physical quantities of interest. Nevertheless, it is useful to reserve imposing the conditions $j=0$ to the end of the calculation.

For $j \neq 0$, then, we introduce the *ansatz*

$$\mathcal{F}(j, j^*) = z^j(j, j^*) / z^j(0, 0);$$

$$z^j(j, j^*) = \iint \mathcal{D}\varphi^* \mathcal{D}\varphi \exp(iS[\varphi]) \times \exp\left(i \int d^d x dt [j \varphi^* + j^* \varphi]\right). \quad (2.4)$$

It is readily verified that Eq. (2.2) is satisfied by the ansatz, with a correct normalization, once the conditions $j = 0$ are imposed. For $j \neq 0$, however, one has the set of Schwinger-Dyson equations

$$\left\{ \hat{L}(i) \frac{1}{i} \frac{\delta}{\delta j^*(i)} - [j(i) + J(i)] \right\} \mathcal{F} = 0. \quad (2.5)$$

Thus, the generating functional \mathcal{F} satisfies the desired conditions, but unfortunately suffers a drawback which is readily apparent in the expression

$$\varphi(1)\varphi(2) = \left[\int \hat{L}^{-1} J^* \right]_1 \left[\int \hat{L}^{-1} J^* \right]_2 - \hat{L}^{-1}(1,2). \quad (2.6)$$

Because all points are assumed to be distinct, the correct equation is satisfied by $\varphi(1)\varphi(2)$, but one would like to project pieces such as that displayed in the equation above which do not connect to the physically interesting currents J .

The required projection will be implemented as part of the averaging process, to which we now turn.

III. AVERAGING

The functional expression Eq. (2.4) above is convenient for purposes of averaging because the quantities to be averaged appear in the exponential. Although it was not explicitly stated earlier, the coefficients in $L(x,t)$ over which averaging is to be performed appear only linearly. Thus, one is basically asked to express the characteristic functional of the probability distributions over which the averaging is being performed.⁹

It is well known, however, that there exists a nontrivial complication in performing the averaging which arises from the normalizing factor $z^J(0,0)$ which is present in the functional. The normalizing factor contains L , which depends upon the random variables to be averaged. Thus, one has

$$\langle \mathcal{F} \rangle = \iint \mathcal{D}\varphi^* \mathcal{D}\varphi \exp[iJ + j\varphi^*] \exp[iJ^* + j^*\varphi] \times \left\langle \exp\left(i \int \varphi^* L \varphi\right) / z^J(0,0) \right\rangle. \quad (3.1)$$

A. Replicas

The so-called replica trick has been introduced in statistical mechanics to attempt a separation of the characteristic functional from the normalization factor in Eq. (3.1). Within the present context, this trick may be viewed as consisting of the following set of observations.

(1) Before averaging, the functional integrals defining \mathcal{F} are Gaussian. Not only can they be performed explicitly, their meaning, term by term, is known.

Thus, explicitly,

$$z^J(j, j^*) = \mathcal{N}_\infty [\text{Det} L]^{-1} \times \exp\left(-i \int \int [J^* + j^*] L^{-1} [J + j]\right). \quad (3.2)$$

The denominator, or normalizing factor, in (2.4) is of the same form, with $j = 0$.

Now, the Det which appears in both numerator and

denominator of \mathcal{F} consists of all closed loops, with propagator L^{-1} , which connect neither to J nor to j .

Similarly, the JJ factor which appears in the denominator multiplying the Det consists of all single lines representing L^{-1} which join distinct J points.

Clearly the role of the denominator is to leave an expression in which there are no closed loops, and in which each line L^{-1} couples to at least one j . But this implies that any method at all which accomplishes the desired cancellation of terms from the numerator will give a correct expression for \mathcal{F} .

One is thus motivated to achieve cancellation of undesired terms in the numerator without making reference to the denominator. To this end, consider the functional

$$z^J(j, j^*) \equiv \prod_{p=1}^K \iint \mathcal{D}\varphi_p^* \mathcal{D}\varphi_p \exp(i \sum_p \sum_q \varphi_p^* L \delta_{pq} \varphi_q) \times \exp(i \sum_p [\varphi_p^* (\alpha J_p + \beta j_p) + \varphi_p (\alpha J_p^* + \beta j_p^*)]). \quad (3.3)$$

Figure 1 displays the graphical structures generated by this expression.

It is routine to demonstrate that the following assertion is true:

$$\Phi(1 \dots M; M + 1 \dots N) = \lim_{\alpha, \beta, K \rightarrow 0} \frac{1}{(N!)^2} \left(\frac{\partial}{\partial \alpha} \right)^N \left(\frac{\partial}{\partial \beta} \right)^N K^{-N} \times \sum_{(\alpha_i)} \left[\frac{1}{i} \frac{\delta}{\delta j(1)} \right]_{\alpha_1} \dots \left[\frac{1}{i} \frac{\delta}{\delta j^*(N)} \right]_{\alpha_N} \times z^J(j, j^*)|_{j=j^*=0}. \quad (3.4)$$

(2) Equation (3.4) asserts that the replica method succeeds in killing all disconnected graphs. Furthermore, the α and β restrictions kill all JJ graphs, and in addition all jj graphs. This latter observation means that the minor problem with $\varphi(1)\varphi(2)$ cited in Eq. (2.6) is resolved.

All of the above holds true for the free theory, before averaging. It implies that the object to be averaged consists of precisely the desired pieces, if the limits of Eq. (3.4) are performed prior to averaging. But now one is to interchange the averaging operation with the limiting operation. As long as a graphical expansion method of the same character as

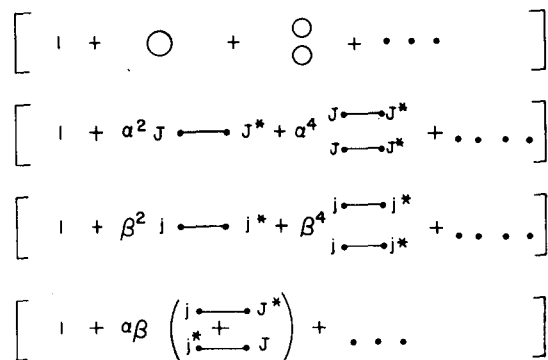


FIG. 1. Formal expansion of z^J in terms of Feynman graphs. Only the α and β coefficients are explicitly displayed. The graphs carry weights obtained by counting the number of configurations, and factors K for each line, open or closed.

that used to motivate the replica method is utilized after averaging, it is apparently valid to interchange the limits.⁶ More will be said on this point below.

Having made the necessary caveats on the known range of validity of the method, we are now in a position to construct

$$\langle z' \rangle = \int \mathcal{D}\varphi^* \mathcal{D}\varphi \langle \exp(i\varphi^* \cdot L \cdot \varphi) \rangle \times \exp\{i[(\mathbf{J} + \mathbf{j}) \cdot \varphi^* + (\mathbf{J} + \mathbf{j}^*) \cdot \varphi]\} \quad (3.5)$$

(3) The expression for the characteristic functional of most value in considering Eq. (3.5) is the cumulant expansion. For a single random variable, we remind the reader of the expressions⁸

$$\langle e^{i\xi\varphi} \rangle_{\xi} \equiv \exp\left[\sum_{p=1}^{\infty} K_p \left(\frac{i\varphi}{p!} \right)^p \right], \quad K_2 = \langle \xi^2 \rangle - \langle \xi \rangle^2, \text{ etc.} \quad (3.6)$$

Functional generalizations of these expressions are readily available.⁹ Consider as an example

$$S = \int \int \left[\rho(x) \left(\frac{\partial y}{\partial t} \right)^2 - T(x) \left(\frac{\partial y}{\partial x} \right)^2 + J(x)y(x) \right] dx dt. \quad (3.7)$$

If the tensions, or elastic constants, are distributed about a mean value T_0 , one has

$$\langle e^{iS} \rangle = e^{iS(T_0)} \exp \left\{ \sum_{p=2}^{\infty} \int \dots \int dx_1 dt_1 \dots dx_p dt_p \times \left[i \left(\frac{\partial y}{\partial x_1} \right)^2 \right] \dots \left[i \left(\frac{\partial y}{\partial x_p} \right)^2 \right] K_p(x_1, x_2, \dots, x_p) \right\}. \quad (3.8)$$

It should be evident from this example, which will be considered in more detail momentarily, that higher order cumulants can potentially become increasingly singular at short distances, creating severe ultraviolet convergence problems in the theory. A likely concomitant is that, much as in the theory of phase transitions, the high order cumulants will be relatively unimportant in the long-wavelength regime. However, care must be exercised in defining the dimensional coupling functions such as K_p in terms of observable quantities and cutoffs. These remarks will be expanded when we examine the linear chain in detail.

(4) The upshot of the averaging procedure is thus seen to be a replacement of the original problem by a nonlinear problem in which the strengths of the nonlinearities are directly related to the cumulants of the statistical distributions in question.

To the extent that the "induced" nonlinearities can be viewed as weak, the new problem can be treated by conventional perturbative methods. The additional complications introduced by the replicas and the α and β constraints are not severe, and collapse to the following rules in graphs:

A. Draw only graphs which connect J with j , initially with propagators L^{-1} .

B. Set down points with emerging lines to represent nonlinear vertices appropriate to the order of perturbation theory being computed.

C. "Break open" the lines joining j with J , and reconnect these with the lines emerging from the vertices in all possible ways.

D. Examine the replica index loop structure which emerges. Count a factor K for each initial jJ line, considering it a trace over the replica indices. Delete all index loop structures in the final diagrams which lead to higher powers of K upon tracing than the power K deduced from the free jJ lines.

B. Anticommuting field variables

It is evident from the discussion of "replicas" given above that the essential point of the replica method is to cancel the graphs generated by $\text{Det}iL$. But clearly, $\text{Det}iL$ is cancelled without recourse to graphs in the expression

$$\bar{z}^j(j, j^*) = (\text{Det}iL) z^j(j, j^*). \quad (3.9)$$

This simple observation is turned into a calculational tool with the introduction of a functional integral representation for $\text{Det}iL$ over anticommuting scalar fields:

$$\text{Det}iL \equiv \int \int \mathcal{D}\bar{\psi} \mathcal{D}\psi e^{i\bar{\psi}L\psi}. \quad (3.10)$$

The "trick" Eq. (3.10) has been introduced into gauge field theory by Fadeev and Popov.⁷ Its relevance to the class of problems heretofore studied using replica methods has been noted recently by McKane.⁷

Using the Fadeev-Popov technique, we have an expression formally free of $\text{Det}L$.

$$\bar{z}^j(j, j^*) \equiv \int \dots \int \mathcal{D}\varphi^* \mathcal{D}\varphi \mathcal{D}\bar{\psi} \mathcal{D}\psi e^{i\bar{\psi}L\psi} \times e^{i\varphi^* L \varphi e^{\varphi^*(J+J^*)} e^{i\varphi(J^*+j^*)}} \quad (3.11)$$

The averaging procedure modifies the integral in Eq. (3.11); using Eq. (3.7) as an example once again, we have

$$\langle \bar{z}^j(j, j^*) \rangle \supset \exp \sum_{p=1}^{\infty} \int \dots \int dx_1 dt_1 \dots dx_p dt_p \times \left\{ i \left[\left(\frac{\partial y}{\partial x_1} \right)^2 + \frac{\partial \bar{\psi}}{\partial x_1} \frac{\partial \psi}{\partial x_1} \right] \dots \times i \left[\left(\frac{\partial y}{\partial x_p} \right)^2 + \frac{\partial \bar{\psi}}{\partial x_p} \frac{\partial \psi}{\partial x_p} \right] K_p(x_1, \dots, x_p) \right\}.$$

Thus, $\langle \bar{z}^j \rangle$ is to be computed as a theory of interacting bosons and to fermions. The choice of dealing with $\langle \bar{z}^j \rangle$ or $\langle z^j \rangle$ in perturbation theory is left to the reader's taste.

In principle, however, the virtue of \bar{z}^j is that no interchange of averaging and graph selection occurs. Thus, one may seek genuinely nonperturbative effects in $\langle \bar{z}^j \rangle$ with confidence that any effects observed are at least as valid as the original functional representation \mathcal{F} .

One nonperturbative method which has proved fruitful in recent years has been the calculation of functional integrals such as $\langle \bar{z}^j \rangle$ by the method of steepest descents. One might also profitably bring renormalization group methods to bear on the problem of computing $\langle \bar{z}^j \rangle$ when a straightforward perturbative approach is not valid. The principal difficulty in applying modern field theoretic methods to the problem at hand appears to be that one gains little intuition from the anticommuting fields in the integral. But this should not be a crucial drawback to the method.

IV. DISORDERED LINEAR CHAIN IN PERTURBATION THEORY

To illustrate the methods described above, we return to the special one-dimensional case

$$\left[\rho(x) \frac{\partial^2}{\partial t^2} - \frac{\partial T}{\partial x} \frac{\partial}{\partial x} - T \frac{\partial^2}{\partial x^2} \right] y'(x,t) = J(x,t). \quad (4.1)$$

It is supposed that the density ρ is homogeneously statistically distributed about a mean ρ_0 , and the tension T is similarly distributed about a mean T_0 .

The average of the solution is clearly

$$\langle y'(x,t) \rangle = \iint dx' dt' \langle G(x,x';t,t') \rangle J(x',t'). \quad (4.2)$$

Inasmuch as we consider only time-independent fluctuations in ρ and T , it is convenient to introduce the Fourier transform

$$G(\omega, x) = \int dt e^{-i\omega t} G(x, 0; t, 0). \quad (4.3)$$

For this simple example, it suffices to compute

$$\left\langle \exp \left[-\frac{1}{2} \int dx \left[\omega^2 \rho(x) y^2 - T(x) \left(\frac{\partial y}{\partial x} \right)^2 \right] \right] \right\rangle_{\rho, T}. \quad (4.4)$$

We shall assume Gaussian distributions in ρ and T , and so write

$$\begin{aligned} \langle F \rangle_{\rho, T} &\equiv \iint \mathcal{D}\rho \mathcal{D}T F[\rho, T] \exp \left\{ -\frac{1}{\Delta_\rho^2} \int dx [\rho(x) - \rho_0]^2 \right\} \\ &\times \exp \left\{ -\frac{1}{\Delta_T^2} \int dx [T(x) - T_0]^2 \right\} / \\ &\iint \mathcal{D}\rho[\rho, T] \exp \left\{ -\frac{1}{\Delta_\rho^2} \int dx [\rho(x) - \rho_0]^2 \right\} \end{aligned} \quad (4.5)$$

Now, if $\rho(x)$ and $T(x)$ were smooth, differentiable functions, we could derive the sourceless version of (4.1) as the continuum limit of the discrete set of equations

$$M_j \frac{d^2 y_j}{dt^2} = k_{j+1} [y_{j+1} - y_j] - k_j [y_j - y_{j-1}]. \quad (4.6)$$

The "mean", or ideal, theory in which ρ and T simply assume their mean values is obtained from (4.6) under the conditions

$$\begin{aligned} M_j &= M = \rho_0 a, \quad \forall j; \\ k_j &= k = T_0/a, \quad \forall j. \end{aligned} \quad (4.7)$$

It is important to notice the manner in which the lattice constant enters into these expressions. For if we now ask a microscopic interpretation to assign the fluctuation parameter Δ_T which enters into (4.5), we have

$$\mathcal{P}[T(x) - T_0] = \exp \left(-\frac{1}{\Delta_T^2} a^3 \sum_i \delta k_i^2 \right).$$

That is,

$$a^3 \delta k_i^2 / \Delta_T^2 \approx 1$$

essentially measures the spread in the spring constants. But this implies a scaled spread in the actual value of the tension given by

$$\Delta T/a = \Delta_T/a^{3/2}.$$

A similar argument applies for the density spread parameter Δ_ρ .

Consequently, displaying the explicit powers of "a" inferred from the above consideration, we find

$$\begin{aligned} \langle z' \rangle &= \int \mathcal{D}y \exp \left\{ -\frac{i}{2} \int dx \left[\omega^2 \rho_0 y^2 + T_0 \left(\frac{\partial y}{\partial x} \right)^2 \right] \right\} \\ &\times \exp \left[-a \frac{\omega^4 (\Delta_\rho)^2}{16} \int dx (y^2(x))^2 \right] \\ &\times \exp \left[-\frac{a (\Delta T)^2}{16} \int dx \left(\frac{\partial y}{\partial x} \cdot \frac{\partial y}{\partial x} \right)^2 \right]. \end{aligned} \quad (4.8)$$

Next, in perturbation theory, graphs generated by the induced nonlinear interaction may diverge in momentum space, as can readily be deduced using standard power-counting arguments.¹⁰ Figure 2 shows an example of the replica index routings which are possible at nontrivial order in the perturbation, and of a graph which diverges. The divergence is regulated by introducing a cutoff at momentum (or, more properly speaking, wavenumber) $2\pi/a$.

There is a competition, therefore, between explicit powers of "a" which enter into defining the fluctuation parameters, and internally generated powers of $1/a$. Retaining only terms up to linear in a , the graphical expansion to second

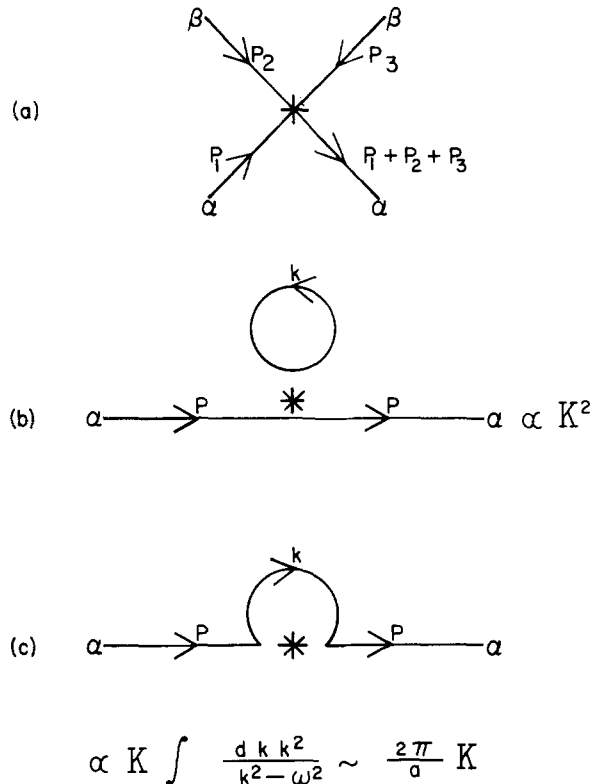


FIG. 2. (a) A vertex induced by averaging. The (*) denotes the vertex carries a factor $p_1 p_2 p_3 (p_1 + p_2 + p_3)$. (b) A bare L^{-1} line dressed by interaction with the vertex of Fig. 2(a). The point vertex has been split open to display the replica index routing. The two independent routes supply a factor K^2 . (c) A second possible index routing, which supplies a factor K . In both Figs. 2(b) and (c), the closed momentum (wavenumber) loop provides a divergent integral which is regulated by cutting the integral off at wavenumber "a".

order in the proper self-energy can be summed to give

$$\langle G(\omega, k) \rangle^{-1} = \rho_0 \left\{ \omega^2 - c^2 \left[1 - \pi \left(\frac{\Delta T}{T_0} \right)^2 \right] k^2 - ia \frac{\pi \omega}{\rho_0 c} \left[\left(\frac{\omega}{c} \right)^2 \left(\frac{\Delta \rho}{\rho_0} \right)^2 + k^2 \left(\frac{\Delta T}{T_0} \right)^2 \right] \right\}. \quad (4.9)$$

Thus, one observes:

(1) A finite renormalization of the speed of sound;
 (2) Pseudodissipation at order “ a ”. The term “pseudo” refers to the fact that total energy is conserved in the system, but backscattering can deplete the amplitude in a given direction of propagation.

(3) Dispersion at order “ a^2 ” (not displayed). From the discussion relating to Eq. (4.6), it should be clear that one is investigating a physical system using a set of equations appropriate to the long wavelength limit. In particular, one has not abandoned the information that the macroscopic system is more properly described by microphysics starting at some length scale “ a ”. The dispersive terms are necessary to provide a sound mathematical meaning to the “long wavelength” limit.

Consider now a source of the specific form

$$J(x, t) = \mathcal{F}_0 \delta(x) \delta(t).$$

In one dimension, it is easily seen that this source corresponds to providing an acceleration impulse to the medium. Using the derived average Green function, the average causal signal at $x > 0$ is

$$\langle y'(x, t) \rangle \approx \frac{\mathcal{F}_0}{4\rho c^*} \left[1 - \text{Erf} \left(\frac{\tau c^*}{2(\alpha a x)^{1/2}} \right) + \left(\frac{\alpha a}{\pi x} \right)^{1/2} \exp \left(-\frac{\tau^2 c^{*2}}{4\alpha a x} \right) \right],$$

where

$$\tau \equiv t - x/c^*, c^* \equiv c \left[1 - \pi/2 (\Delta T/T)^2 \right],$$

$$\alpha \equiv \frac{1}{2} \pi \left[(\Delta \rho/\rho)^2 + (\Delta T/T)^2 \right].$$

Observe that all of the dimensional parameters in the problem have organized themselves into the combinations $\alpha a/x$ and $c^* \tau/(\alpha a x)^{1/2}$. Our approximation has consisted of treating α and “ a ” as small parameters. But there is no real meaning to saying that “ a is small.” Some further criterion based upon wavenumber—lattice spacing dimensionless combinations is provided by higher terms in the perturbation expansion, as was argued earlier. However, even without displaying such higher order terms, we see in Eq. (4.10) still a *different* possible criterion for the validity of perturbation theory emerging—it is that the range may play an essential role in determining the validity of the perturbative expansion. This is not unexpected, and corresponds physically to the likely possibility that small fluctuation-induced disturbances can add up over long propagation paths.⁴

V. SUMMARY

The essential point of this paper is the introduction of a generating functional for “ N -point amplitudes” of a wave equation with random variables entering as coefficients. The

generating functional which we have introduced is not restricted to any special approximation to the full wave equation.

When a perturbative treatment of the fluctuating random variables is appropriate, the replica approach to calculating averages over fluctuations is convenient because the graphical rules the method generates are simple.

In addition, it is straightforward to write the Schwinger–Dyson equations for the averaged functional $\langle z' \rangle$, and so infer sets of SD equations for any amplitudes of interest. These may be viewed as resummations of the perturbative expansion, i.e., equations inferrable in principle from a careful enumeration of graphs, and so can be considered valid in their “replica” form. It should then be correct to attempt nonperturbative methods for the solution of the set of equations. An analogue here is the Coulomb problem: no finite set of graphs can give rise to bound states, but bound positronium solutions to the Bethe–Salpeter equation can be viewed as corresponding to the sum of an infinite set of graphs.

An important virtue of the functional approach, then, is the convenient summary of perturbation theory which it entails. In addition, it may be possible to reach interesting and valid nonperturbative results via functional methods, which would be difficult to generate by other means. The safest avenue for this approach is the representation \bar{z}' , using anti-commuting fields.

ACKNOWLEDGMENTS

I have benefitted from stimulating conversations with colleagues at Schlumberger–Doll Research Center, and from perceptive remarks by Roger Dashen.

- ¹U. Frisch, in *Probabilistic Methods in Applied Mathematics*, (Academic, New York, 1968); B. J. Uscinski, *The Elements of Wave Propagation in Random Media* (McGraw-Hill, N. Y., 1977); S. Flatte, *Sound Transmission Through a Fluctuating Ocean* (Cambridge U.P., New York, 1979).
- ²Frisch (Ref. 1) discusses “ergodicity” questions which arise in interpreting ensemble averaged random medium calculations. We shall assume, as is standard, that the ensemble average provides statistical information about individual experiments.
- ³Functional methods are mentioned briefly in Frisch (Ref. 1), with references to earlier literature. A highly successful calculation using Feynman path integrals is related in Flatte (Ref. 1) based on the original work by R. Dashen (Ref. 4).
- ⁴R. Dashen, *J. Math. Phys.* **20**, 894 (1979).
- ⁵S. K. Ma, *Modern Theory of Critical Phenomena* (Benjamin, Reading, 1976); P. W. Anderson and S. Edwards, *J. Phys. F* **5**, 965 (1975).
- ⁶V. J. Emery, *Phys. Rev. B* **11**, 239 (1975); G. Grinstein and A. Luther, *ibid.* **B 13**, 1329 (1976); P. G. de Gennes, *Phys. Lett. A* **38**, 339 (1972).
- ⁷L. D. Fadeev and V. N. Popov, *Phys. Lett. B* **25**, 29 (1976); L. D. Fadeev, *Theor. and Math. Phys.* **1**, 1 (1969) [Translation of *Theoret. Mat. Fiz.*]. A. J. McKane, “Reformulation of $n \rightarrow 0$ Models Using Anticommuting Scalar Fields,” Cambridge Preprint, DAMPT 80/3.
- ⁸N. Arley and K. R. Bach, *Introduction to the Theory of Probability and Statistics* (Wiley, New York, 1950).
- ⁹S. K. Ma, Ref. 5. An important physical consideration to be treated elsewhere involves defining the cumulants experimentally.
- ¹⁰J. D. Bjorken and S. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965).

The maximality of Lorentz spaces and their importance with respect to classical and quantum electrodynamics

W. Gessner

Unterer Katzenbergweg 7, D 8700 Würzburg, Germany

(Received 7 March 1980; accepted for publication 5 September 1980)

The indefinite metric state space \mathcal{S}_M of the covariant form of the quantized Maxwell field M contains, as is known, a family of continuously many, isomorphic, isometric pre-Hilbert spaces \mathcal{L}^q , called Lorentz spaces, each of which corresponds to one square-integrable, prescribed, classical, spatial distribution $q(\mathbf{x})$ of the total charge $Q = 0$. The quotient spaces $\mathcal{L}^q/\mathcal{N}^q$ modulo the subspace $\mathcal{N}^q \subset \mathcal{L}^q$ of all elements with norm 0 are indeed Hilbert spaces in \mathcal{S}_M and it appears that any QED which has to do with \mathcal{S}_M has to be formulated not on \mathcal{S}_M as a whole but on the family of these Lorentz spaces. To support this assumption we embed any $\mathcal{L}^q/\mathcal{N}^q$ in \mathcal{L}^q in an isomorphic-isometric way and thus get Hilbert spaces ℓ^q in \mathcal{S}_M , but not in a unique way; we will show that the different possibilities of this embedding correspond exactly with the different gauges. The main results about any embedding space ℓ^q are, however, that it is a *maximal* Hilbert space in \mathcal{S}_M (under a premise referring to the expectation values of charge distribution), and that any Hilbert space of \mathcal{S}_M which is "physically important" in some sense is necessarily one of these ℓ^q . In this way the family $\{\ell^q | q \in \Omega\}$ of Lorentz spaces (defined by the index set Ω) has some outstanding properties so that the ℓ^q are now characterized by these qualities (and no longer in the heuristical way via generalization of the classical Lorentz condition). Starting now with the prominent role of the ℓ^q we get not only a deeper understanding of the Lorentz condition of classical electrodynamics—the properties of the ℓ^q lead us automatically to the definition of a positive-definite state space of QED with the use only of these ℓ^q . (Our considerations refer not to full but to some restricted QED; the restriction is primarily given by the above-mentioned index set Ω so that extensions to full QED seem possible). We show that our definition of a state space is consistent with time evolution, given by the Hamiltonian H , and that the QED on the basis of this state space is a constraint-free theory because the otherwise-necessary selection rules of Lorentz condition and charge conservation are now superfluous (and not present in a hidden form either). Furthermore, the properties of Lorentz spaces lead us automatically to a new concept of observables, all of which commute from the beginning with the operator of charge distribution. As a special observable we discuss the number operator $N(\mathbf{k})$ of photons by showing that this, in general, cannot be of the form $a_\mu^+ a^\mu$. A modified form of $N(\mathbf{k})$ and with it a reformulation of the Hamiltonian H of QED is given. All these considerations go back to the properties of the Lorentz spaces and thus, basically, to the canonical quantization of the Maxwell field.

PACS numbers: 03.50.De, 12.20.Ds

1. INTRODUCTION

The state space \mathcal{S}_M of the quantized Maxwell field¹ M is one of the basic mathematical structures of physics. In any QED it is a "partner space" within the whole state space of the theory. Beyond QED the space \mathcal{S}_M will have a certain model character, for the state space of gravitons,² e.g., or with respect to the state space of the unified gauge theory of the weak and electromagnetic interaction.

In this paper \mathcal{S}_M is to be investigated for itself and within the framework of QED respectively. If a quantized four-current density $J_\mu(\mathbf{x})$ in the Schrödinger picture is introduced together with the state space \mathcal{S}_J , spanned by the creation operators of electrons and positrons in the usual way, then the first choice for the state space of this QED is the tensor product $\mathcal{S}_J \otimes \mathcal{S}_M$.

But in this way we do not get a useful state space, for in \mathcal{S}_M the Cauchy-Schwarz inequality cannot generally hold, as we know¹ (we use the familiar Gupta-Bleuler-metric). As a consequence, the space $\mathcal{S}_J \otimes \mathcal{S}_M$ does not admit the statis-

tical interpretation of a quantum theory in the sense of Born.³ The cause of this defect lies in the canonical commutation relations of the operators $A_\mu(\mathbf{x})$, $\Pi_\nu(\mathbf{x}')$ in the Schrödinger picture, which mainly determine the structure of \mathcal{S}_M .

One way out of this dilemma is the possibility of looking for subspaces of \mathcal{S}_M which are positive-semidefinite with respect to the Gupta-Bleuler metric or even Hilbert spaces. Then we have to examine in what way these spaces can be used as partner spaces of \mathcal{S}_J . The discovery and examination of such subspaces is first of all a purely mathematical problem. In order to test their relationship with physics one can introduce a pseudo-interaction on only \mathcal{S}_M as an important heuristic aid: we prescribe any classical, conserved, four-current and examine to what extent its interaction with M can be described on the detected subspace of \mathcal{S}_M .

The task of finding positive-semidefinite subspaces of \mathcal{S}_M has so far been treated as follows: Gupta⁴ and Bleuler⁵ introduced a generalization of the Lorentz condition of classical electrodynamics for the free M . They named such ele-

ments α of \mathcal{S}_M "physical" which satisfied the condition

$$L(\mathbf{k})\alpha = 0 \text{ for almost any } \mathbf{k} \in \mathbb{R}^3, \quad (1)$$

where $L(\mathbf{k}) = k^2 a_0(\mathbf{k})$. In this way they introduced the eigenspace \mathcal{L}^0 of the operator $L(\mathbf{k})$ to the eigenvalue 0 and showed that \mathcal{L}^0 is positive-semidefinite. In order to be able to describe at least the above mentioned pseudo-interaction on \mathcal{S}_M , in a further step¹ the following generalization of (1) was made:

$$L(\mathbf{k})\alpha = q(\mathbf{k})\alpha \text{ for almost any } \mathbf{k} \in \mathbb{R}^3, \quad (2)$$

where $q = q(\mathbf{k})$ is taken from the set \mathcal{Q} of all complex-valued square-integrable functions over the \mathbb{R}^3 so that $q^*(\mathbf{k}) = q(-\mathbf{k})$. Consequently, the Fourier transform

$$q(\mathbf{x}) = (2\pi)^{-3/2} \int d^3k (2|\mathbf{k}|)^{1/2} e^{i\mathbf{k}\cdot\mathbf{x}} q(\mathbf{k}) \quad (3)$$

is real-valued and square-integrable. To each eigenfunction q of this kind the eigenspace

$$\mathcal{L}^q = \{ \alpha \in \mathcal{S}_M \mid L(\mathbf{k})\alpha = q(\mathbf{k})\alpha \text{ for almost any } \mathbf{k} \in \mathbb{R}^3 \}$$

was introduced and called "Lorentz space". For these Lorentz spaces the following statements were proved:

Theorem 1:

(a) Any Lorentz space \mathcal{L}^q is positive-semidefinite; the space of equivalence classes $\mathcal{L}^q/\mathcal{N}^q$ modulo \mathcal{N}^q : = $\{ \alpha \in \mathcal{L}^q \mid \langle \alpha \mid \alpha \rangle = 0 \}$ is Hilbert space with respect to the scalar product of \mathcal{S}_M .

(b) Any pair of different Lorentz spaces $\mathcal{L}^q, \mathcal{L}^p (p \neq q)$ satisfies $\mathcal{L}^q \cap \mathcal{L}^p = \{0\}$; but between \mathcal{L}^q and \mathcal{L}^p there exists an isometric isomorphism $\mathcal{L}^q \cong \mathcal{L}^p$.

(c) In the span⁶ $\text{sp}\{\mathcal{L}^q, \mathcal{L}^p\}$ of two different Lorentz spaces the Cauchy-Schwarz inequality cannot hold, so that any subspace of \mathcal{S}_M which contains the span of any pair of different Lorentz spaces is indefinite by necessity.

(d) For test purposes a pseudointeraction between the quantized Maxwell field M and any prescribed, classical, stationary four-current $j_\mu(\mathbf{x})$ may be introduced. But it can only be described on the Lorentz space \mathcal{L}^q , for which $q(\mathbf{x}) = j_0(\mathbf{x})$, where $q(\mathbf{k})$ is given by Eq. (3).

According to (d) any classical, prescribed, charge density determines its own Lorentz space. One might rather expect that this is true for any total charge. In the cases to be discussed the charge vanishes because of the square-integrability of q and Eq. (3). The one-to-one correspondence between charge densities and Lorentz spaces, however, permits an interesting physical interpretation: according to Theorem 1 the \mathcal{L}^q form positive-semidefinite "islands" in the indefinite \mathcal{S}_M . The admitted charge densities correspond with these islands; on the other hand these correspond with the photon robes which surround the charge densities, and which build up the Coulomb fields as quantum theoretical expectation values. Thus the one-to-one correspondence indicates at the same time the inseparability of charge densities and photon robes. The indefinite character of \mathcal{S}_M which renders possible this inseparability, no longer appears as a formal disadvantage, but has a direct physical meaning.¹

The properties of the Lorentz spaces \mathcal{L}^q , which we have described so far, and their physical interpretation, suggest the conclusion that each interaction theory which has to

do with \mathcal{S}_M must be brought in connection with these \mathcal{L}^q , not with \mathcal{S}_M as a whole. Because of the far-reaching consequences of such a conclusion the Lorentz spaces, especially their quotient spaces $\mathcal{L}^q/\mathcal{N}^q$, will have to be further investigated. As we need Hilbert subspaces of \mathcal{S}_M , we embed any quotient space $\mathcal{L}^q/\mathcal{N}^q$ isomorphically in \mathcal{L}^q . The resulting subspace ℓ^q of \mathcal{L}^q is not defined in a unique way. One result is, that the different possibilities for embedding $\mathcal{L}^q/\mathcal{N}^q$ in \mathcal{L}^q correspond exactly with the gauges (Sec. 2). However now the Hilbert space ℓ^q in \mathcal{L}^q is selected: in Sec. 3 it can be shown that ℓ^q is a *maximal* Hilbert space in \mathcal{S}_M (under an additional premise referring to the expectation values of charge densities). This maximality of any ℓ^q means a considerable refinement of Theorem 1(c). Vice versa, any physically important Hilbert space in \mathcal{S}_M is necessarily one of our ℓ^q . In this way the important role of the Lorentz spaces has been proved, and this permits us to forget about the heuristic way for the detection of the Lorentz spaces via generalization of the classical Lorentz condition, and to start directly from the outstanding properties of the Lorentz spaces. Then we can attempt to answer some important questions:

(i) In classical electrodynamics (CED) the Lorentz condition (CLC) plays a dubious role: it is added "from outside" to the Hamilton equations of motion of the classical potential A_μ and its canonical momentum Π , and is motivated only afterwards by the results (we arrive at Maxwell's equations). A deeper understanding of CLC is impossible within the framework of CED. Now the first step from CED towards QED is the theory of pseudo-interaction between photons and any prescribed classical four-current as it is used here as a heuristic aid. This theory can, however, be completely described on the family $\{\mathcal{L}^q\}$ of Lorentz spaces. If we return from this theory to CED, we will automatically get CLC there. Thus, CLC is a consequence of the outstanding properties of the Lorentz spaces in \mathcal{S}_M (Sec. 4).

(ii) The Lorentz spaces should, of course, above all play a decisive role in a "true" QED with quantized four-current $J_\mu(\mathbf{x})$. We construct (Sec. 5) the positive-definite state space of such a QED, which results almost automatically from the properties of the ℓ^q , and define "observables" on this state space according to the Lorentz subspace structure of \mathcal{S}_M . The properties of the Lorentz spaces lead immediately to the statement that these observables commute with the operator of charge density (theorem of Strocchi and Wightman⁷). However, our results refer not to full but to some restricted QED, defined essentially by the above mentioned index set \mathcal{Q} . Extensions to full QED seem possible but exceed the intention of this paper (Sec. 5).

(iii) As an example of an observable in the sense of Sec. 5 we discuss the operator $N(\mathbf{k})$ of photon number (Sec. 6). We show that $N(\mathbf{k})$, as a consequence of our definitions, is a non-negative operator, but in general it cannot have the form $a_\mu^+(\mathbf{k})a^\mu(\mathbf{k})$ which is usually given.

2. THE ISOMETRIC EMBEDDING OF THE QUOTIENT SPACES $\mathcal{L}^q/\mathcal{N}^q$ IN \mathcal{L}^q AND THE GAUGE PROBLEM

First we repeat the definition of the Lorentz spaces $\mathcal{L}^q \subset \mathcal{S}_M$ and give it an alternative form which will become

important for later considerations.

Ω may be the set of all complex-valued functions $q = q(\mathbf{k})$ over R^3 with the property that $q(\mathbf{k})$ and $(|\mathbf{k}|)^{1/2}q(\mathbf{k})$ are square-integrable in the Lebesgue sense and that $q^*(\mathbf{k}) = q(-\mathbf{k})$ for any $\mathbf{k} \in R^3$. Then it follows that

$$q(\mathbf{x}) := (2\pi)^{-3} \int d^3k (2|\mathbf{k}|)^{1/2} e^{i\mathbf{k}\cdot\mathbf{x}} q(\mathbf{k})$$

is real-valued and that $Q := \int d^3x q(\mathbf{x}) = 0$. The function $q(\mathbf{x})$ has—according to Theorem 1(d)—the meaning of any stationary charge density with vanishing total charge.

We choose any $q \in \Omega$ and define the subspace \mathcal{L}^q of \mathcal{S}_M as the eigenspace of the “Lorentz operator” $L(\mathbf{k}) := k^\nu a_\nu(\mathbf{k})$ to the eigenvalue function $q = q(\mathbf{k})$:

$$\mathcal{L}^q := \{ \alpha \in \mathcal{S}_M \mid L(\mathbf{k})\alpha = q(\mathbf{k})\alpha \text{ for almost any } \mathbf{k} \in R^3 \}. \quad (4)$$

The definition of \mathcal{S}_M is given in the Appendix. The characteristics of the spaces \mathcal{L}^q which have been known up to now are briefly summarized in Theorem 1 and explained in Ref. 1. Especially, \mathcal{L}^q is positive-semidefinite, i.e., it contains elements $\alpha \neq 0$ with $\langle \alpha | \alpha \rangle = 0$. We introduce the subspace

$$\mathcal{N}^q := \{ \alpha \in \mathcal{L}^q \mid \langle \alpha | \alpha \rangle = 0 \}. \quad (5)$$

The quotient space $\mathcal{L}^q / \mathcal{N}^q$ is not only positive-definite, but even a Hilbert space with respect to the (Gupta—Bleuler) metric $\langle | \rangle$ of \mathcal{S}_M ; we call it a Lorentz space, too.

Instead of defining \mathcal{L}^q as the eigenspace of $L(\mathbf{k})$ to the eigenvalue function $q = q(\mathbf{k})$ as above, one can also begin with the operator

$$L(\mathbf{x}) := (2\pi)^{-3} \int d^3k (2|\mathbf{k}|)^{1/2} e^{i\mathbf{k}\cdot\mathbf{x}} L(\mathbf{k}). \quad (6)$$

Then \mathcal{L}^q is the eigenspace of $L(\mathbf{x})$ to the eigenvalue function $q = q(\mathbf{x})$ of Eq. (3). According to these two possibilities for defining \mathcal{L}^q , the notation \mathcal{L}^q leaves open whether one thinks especially of the eigenspace of $L(\mathbf{k})$ to $q(\mathbf{k})$ or of $L(\mathbf{x})$ to $q(\mathbf{x})$. However, the latter has the advantage that $L(\mathbf{x})$ is self-adjoint on each Lorentz space (because \mathcal{L}^q is not a Hilbert space, the term “self-adjoint on \mathcal{L}^q ” remains to be defined. We call any operator Ω self-adjoint on \mathcal{L}^q if $\Omega \mathcal{L}^q \subseteq \mathcal{L}^q$ and $\langle \alpha | \Omega \alpha \rangle \in R$ for any $\alpha \in \mathcal{L}^q$).

Only the Lorentz spaces \mathcal{L}^0 and $\mathcal{L}^0 / \mathcal{N}^0$ contain the state $\omega_0 := (1, 0, 0, \dots)$ of the bare vacuum (Appendix), which can also be defined by $\langle \omega_0 | \omega_0 \rangle = 1$ and $a_\mu(\mathbf{k})\omega_0 = 0$ for all μ and \mathbf{k} . The quotient space $\mathcal{L}^q / \mathcal{N}^q$ can be embedded isomorphically in \mathcal{S}_M in many ways. Because of the isometry of any two Lorentz spaces, $\mathcal{L}^q / \mathcal{N}^q$ could also be embedded in any \mathcal{L}^p with $p \neq q$. Because of the meaning of q as a charge density, however, we think $\mathcal{L}^q / \mathcal{N}^q$ embedded in “its” Lorentz space \mathcal{L}^q . For any $q \in \Omega$ such an isomorphic-isometric embedding of $\mathcal{L}^q / \mathcal{N}^q$ can be chosen. We denote this subspace of \mathcal{L}^q by ℓ^q . The Lorentz spaces ℓ^q are Hilbert spaces in \mathcal{S}_M and—up to the zero element 0 of \mathcal{S}_M —pairwise disjoint (Theorem 1, (b)). Thus the Hilbert spaces ℓ^q represent “islands” in the indefinite \mathcal{S}_M which, as it appears, are of special importance for physics. However, no one island ℓ^q is embedded in \mathcal{L}^q in a unique, physically motivated way. For example, with ℓ^q the subspace $\ell^q := W\{kg\}$ ℓ^q is also a possible embedding of $\mathcal{L}^q / \mathcal{N}^q$ in \mathcal{L}^q , where $W\{kg\}$ is the

Weyl operator (Appendix) produced by the four-vector function $g_\mu(\mathbf{k}) := k_\mu g(\mathbf{k})$ to any scalar function $g(\mathbf{k})$. In this way we get continuously many isomorphic-isometric embeddings of $\mathcal{L}^q / \mathcal{N}^q$ in \mathcal{L}^q .

If one claims that all possible embeddings of $\mathcal{L}^q / \mathcal{N}^q$ in \mathcal{L}^q are equally valid physically, then the change from one embedding to another one can only mean a gauge transformation in a sense to be defined. Operators are not influenced at all by this change of the embedding, for we maintain, as in Ref. 1, that the dynamical variables $A_\mu(\mathbf{x})$, $\Pi_\nu(\mathbf{x})$ of the theory are gauge-invariant *ab ovo*, which is then also true for their functionals, especially the Hamiltonian. Therefore, we define a gauge transformation in the following way:

Definition 1: Let ℓ_1^q and ℓ_2^q be two embeddings of $\mathcal{L}^q / \mathcal{N}^q$ in \mathcal{L}^q . Then a *gauge transformation* consists in the replacement of the transition element $\langle \alpha_1 | \Omega \beta_1 \rangle$ to any $\alpha_1, \beta_1 \in \ell_1^q$ by the element $\langle \alpha_2 | \Omega \beta_2 \rangle$, where Ω is any operator with $\Omega \mathcal{L}^q \subseteq \mathcal{L}^q$ and α_2, β_2 are elements from ℓ_2^q so that $\alpha_2 - \alpha_1 \in \mathcal{N}^q$ and $\beta_2 - \beta_1 \in \mathcal{N}^q$.

We now need a special algebra $\underline{\Omega}^q$ of operators to any fixed Lorentz space \mathcal{L}^q :

Let $\underline{\Omega}^q$ be the set of all operators Ω so that $\Omega \mathcal{L}^q \subseteq \mathcal{L}^q$ and Ω is self-adjoint on \mathcal{L}^q .

Lemma 2a:

(a) Any operator $\Omega \in \underline{\Omega}^q$ can be defined on $\mathcal{L}^q / \mathcal{N}^q$ also, and so on any embedding space ℓ^q .

(b) The matrix elements of any $\Omega \in \underline{\Omega}^q$ are *gauge invariant*, i.e., $\langle \alpha_1 | \Omega \beta_1 \rangle = \langle \alpha_2 | \Omega \beta_2 \rangle$ for all $\alpha_1, \alpha_2, \beta_1, \beta_2$ according to definition 1.

(c) The algebra $\underline{\Omega}^q$ is irreducible on $\mathcal{L}^q / \mathcal{N}^q$ and so on any embedding space ℓ^q if we agree that irreducibility means: any element $\beta \neq 0$ from $\mathcal{L}^q / \mathcal{N}^q$ can be mapped onto any element $\alpha \in \mathcal{L}^q / \mathcal{N}^q$ by a suitable $\Omega \in \underline{\Omega}^q$.

Proof:

(a) We have to show: $\Omega \mathcal{N}^q \subseteq \mathcal{N}^q$. Let $\alpha \in \mathcal{N}^q$ and $\beta \in \mathcal{L}^q$. Then the Cauchy—Schwarz-inequality, which holds on \mathcal{L}^q , says:

$$|\langle \beta | \Omega \alpha \rangle|^2 = |\langle \Omega \beta | \alpha \rangle|^2 \leq \langle \Omega \beta | \Omega \beta \rangle \langle \alpha | \alpha \rangle = 0, \text{ so that } \langle \beta | \Omega \alpha \rangle = 0.$$

With $\beta := \Omega \alpha \in \mathcal{L}^q$ we get the assertion.

(b) This follows immediately from $\Omega \mathcal{N}^q \subseteq \mathcal{N}^q$.

(c) The irreducibility means that there is always an operator $\Pi_{\alpha\beta}$ from $\underline{\Omega}^q$ so that $\alpha = \Pi_{\alpha\beta} \beta$, where α and $\beta \neq 0$ are prescribed elements from $\mathcal{L}^q / \mathcal{N}^q$.

In the case of $\langle \alpha | \beta \rangle \neq 0$ one takes the projector $\Pi_{\alpha\beta} := |\alpha\rangle \langle \alpha| / \langle \alpha | \beta \rangle$ which is obviously an operator from $\underline{\Omega}^q$, for $\Pi_{\alpha\beta} \beta = e^{i\varphi} \alpha$ with some $\varphi \in R$; because of the usual ray-equivalence in the Hilbert space $\mathcal{L}^q / \mathcal{N}^q$, the elements α and $e^{i\varphi} \alpha$ are identified. In the case of $\langle \alpha | \beta \rangle = 0$ one takes $\Pi_{\alpha\beta} := 1 / \langle \beta | \beta \rangle \cdot \{ |\alpha\rangle \langle \beta| + |\beta\rangle \langle \alpha| \}$ and obtains directly $\Pi_{\alpha\beta} \beta = \alpha$. ■

Lemma 2b:

(a) Let ℓ_1^q and ℓ_2^q be two different embeddings of $\mathcal{L}^q / \mathcal{N}^q$ in \mathcal{L}^q . Then $\ell_1^q \cap \ell_2^q = \{0\}$.

(b) To any element $\alpha \in \mathcal{L}^q$ with $\alpha \notin \mathcal{N}^q$ there exists precisely one embedding ℓ^q of $\mathcal{L}^q / \mathcal{N}^q$ in \mathcal{L}^q so that $\alpha \in \ell^q$.

Proof:

(a) If we had $u := \ell_1^q \cap \ell_2^q \neq \{0\}$, then every operator $\Omega \in \underline{\Omega}^q$ would satisfy $\Omega u \subseteq u$. This contradicts the irreducibility of the algebra $\underline{\Omega}^q$ (Lemma 2a).

(b) Choose $\ell^q := \{\Omega \alpha \mid \Omega \in \underline{\Omega}^q\}$. ■

It is impossible to pick out one embedding as especially significant among all possible embeddings of $\mathcal{L}^q/\mathcal{N}^q$ in \mathcal{L}^q . In the case $q = 0$ the subspace ℓ of "transversal" photons¹ could be given a special role. By the destruction operators $a^{(0)}(\mathbf{k})$ of "scalar" and $a^{(3)}(\mathbf{k})$ of "longitudinal" photons (Appendix) the subspace $\ell \subset \mathcal{L}^0$ is defined as follows:

$$\ell := \{\alpha \in \mathcal{S}_M \mid a^{(0)}(\mathbf{k})\alpha = 0 = a^{(3)}(\mathbf{k})\alpha \text{ for almost any } \mathbf{k}\}. \quad (7)$$

In Ref. 1 we have shown that ℓ is Hilbert space in \mathcal{S}_M and isomorphic to the quotient space $\mathcal{L}^0/\mathcal{N}^0$. So, ℓ is some embedding of $\mathcal{L}^0/\mathcal{N}^0$ in \mathcal{L}^0 . Because ℓ contains the bare vacuum $\omega_0 := (1, 0, 0, \dots)$ (Appendix), we have as a consequence of Lemma 2b:

Lemma 2c: ℓ is the embedding ℓ_0^0 of $\mathcal{L}^0/\mathcal{N}^0$ in \mathcal{L}^0 which contains the bare vacuum ω_0 . Each embedding different from ℓ_0^0 then contains no longer the bare vacuum ω_0 , but instead a "modified" (Friedrichs⁸) vacuum ω , which can not be characterized by the vanishing of $a_\mu(\mathbf{k})\omega$.

It has already been mentioned that from any embedding ℓ^q one gets another possible embedding $\ell_2^q := W\{kg\}\ell^q$ by the application of the Weyl operator $W\{kg\}$ to any scalar function $g = g(\mathbf{k})$. The following theorem says that in this way one will get all embeddings of $\mathcal{L}^q/\mathcal{N}^q$ in \mathcal{L}^q .

Theorem 2: Let ℓ_1^q and ℓ_2^q be two embeddings of $\mathcal{L}^q/\mathcal{N}^q$ in \mathcal{L}^q . Then there exists a scalar function $g = g(\mathbf{k})$ so that $\ell_2^q = W\{kg\}\ell_1^q$.

Proof: It is sufficient to limit the proof to the case $q = 0$, for—according to the Appendix—there is always a Weyl operator $W\{g\}$ so that $\mathcal{L}^0 = W\{g\}\mathcal{L}^q$. Then $\ell_1^0 := W\{g\}\ell_1^q$ and $\ell_2^0 := W\{g\}\ell_2^q$ are embeddings of $\mathcal{L}^0/\mathcal{N}^0$ in \mathcal{L}^0 . If it can be shown that there is a scalar function g in the sense of the theorem then it follows that

$$\begin{aligned} \ell_2^0 &= W\{-g\}\ell_2^0 = W\{-g\}W\{kg\}\ell_1^0 \\ &= W\{kg\}W\{-g\}\ell_1^0 = W\{kg\}\ell_1^0. \end{aligned} \quad (8)$$

ℓ^0 is now taken as any embedding of $\mathcal{L}^0/\mathcal{N}^0$ in \mathcal{L}^0 , and ℓ_0^0 as the embedding which contains the bare vacuum ω_0 . Now it is sufficient to show that there is always a scalar function g so that $\ell^0 = W\{kg\}\ell_0^0$. Then there are scalar functions g_1 and g_2 so that $\ell_1^0 = W\{kg_1\}\ell_0^0$ and $\ell_2^0 = W\{kg_2\}\ell_0^0$. From this follows immediately $\ell_2^0 = W\{k(g_2 - g_1)\}\ell_1^0$.

In order to prove $\ell^0 = W\{kg\}\ell_0^0$, it is sufficient to show with reference to the modified vacuum ω of ℓ^0 (which belongs to the same equivalence class modulo \mathcal{N}^0 as the element ω_0) that there exists a scalar function g with the property $\omega = W\{kg\}\omega_0$. Then we have $\ell^0 = W\{kg\}\ell_0^0$, for $W\{kg\}\ell_0^0$ is a possible embedding of $\mathcal{L}^0/\mathcal{N}^0$ in \mathcal{L}^0 and contains the element $\omega \neq 0$ from ℓ^0 , and consequently must be identical with ℓ^0 according to Lemma 2b. In order to show that $\omega = W\{kg\}\omega_0$ with a suitable scalar function g we define first the components of ω in the usual way (Appendix) as follows:

$$\varphi := \langle \omega_0 \mid \omega \rangle$$

and

$$\varphi_{\mu_1, \dots, \mu_n}(\mathbf{k}_1, \dots, \mathbf{k}_n) := [1/(n!)^{1/2}] \langle \omega_0 \mid a_{\mu_1}(\mathbf{k}_1) \dots a_{\mu_n}(\mathbf{k}_n) \omega \rangle \quad (9)$$

for all $n \in \mathbb{N}$. Now we assert that there is a scalar function $g = g(\mathbf{k})$ so that for all $n \in \mathbb{N}$:

$$\varphi_{\mu_1, \dots, \mu_n}(\mathbf{k}_1, \dots, \mathbf{k}_n) = [1/\sqrt{n!}] k_{\mu_1} \dots k_{\mu_n} g(\mathbf{k}_1) \dots g(\mathbf{k}_n). \quad (10)$$

We prove this by *induction*: in the case $n = 1$ we have to show that

$$\varphi_\mu(\mathbf{k}) = \langle \omega_0 \mid a_\mu(\mathbf{k}) \omega \rangle = k_\mu g(\mathbf{k}).$$

To prove this we use the Heisenberg operators for the free case

$$A_\mu(x) := (2\pi)^{-\frac{3}{2}} \int d^3k / (2|\mathbf{k}|)^{1/2} e^{ikx} a_\mu(\mathbf{k}) + cc, \quad (11)$$

$$\begin{aligned} F_{\mu\nu}(x) &:= i(2\pi)^{-\frac{3}{2}} \\ &\quad \times \int d^3k / (2|\mathbf{k}|)^{1/2} e^{ikx} \{a_\mu(\mathbf{k})k_\nu - a_\nu(\mathbf{k})k_\mu\} + cc. \end{aligned} \quad (12)$$

A straightforward calculation yields $[F_{\mu\nu}(x), L(\mathbf{k})] = 0$ so that $F_{\mu\nu}(x)\mathcal{L}^q \subseteq \mathcal{L}^q$ for any $q \in \mathbb{Q}$. Further on, $F_{\mu\nu}(x)$ is self-adjoint on any \mathcal{L}^q . As a consequence, the matrix elements of $F_{\mu\nu}(x)$ are gauge-invariant (Lemma 2a) so that

$$\langle \omega_0 \mid F_{\mu\nu}(x) \omega \rangle = \langle \omega_0 \mid F_{\mu\nu}(x) \omega_0 \rangle = 0. \quad (13)$$

We generalize this equation for later purposes to the following statement:

$$\langle \omega_0 \mid F_{\mu\nu}(x) a_{\mu_1}(\mathbf{k}_1) \dots a_{\mu_n}(\mathbf{k}_n) \omega \rangle = 0 \text{ for all } n \in \mathbb{N}. \quad (14)$$

Here $\beta := a_{\mu_1}(\mathbf{k}_1) \dots a_{\mu_n}(\mathbf{k}_n) \omega$ is an element from \mathcal{L}^0 and lies in the same equivalence class modulo \mathcal{N}^0 as the element $\beta' := a_{\mu_1}(\mathbf{k}_1) \dots a_{\mu_n}(\mathbf{k}_n) \omega_0$. According to Lemma 2a, the matrix element $\langle \omega_0 \mid F_{\mu\nu}(x) \beta \rangle$ is gauge-invariant, so that $\langle \omega_0 \mid F_{\mu\nu}(x) \beta \rangle = \langle \omega_0 \mid F_{\mu\nu}(x) \beta' \rangle = 0$. This is Eq. (14).

The term $\langle \omega_0 \mid F_{\mu\nu}(x) \omega \rangle = 0$ represents now the rotation-free integrand of the curve integral

$$\int_0^x d\xi^\mu \langle \omega_0 \mid A_\mu(\xi) \omega \rangle, \quad (15)$$

where we have integrated along any smooth curve $\xi^\mu = \xi^\mu(s)$, $s =$ curve parameter, from the point 0 to the point $x = (x_\mu)$ of the Minkowski space. Therefore, the integral (15) represents a function $A(x)$ of the point x alone, for which we have

$$A_\mu(x) = \langle \omega_0 \mid A_\mu(x) \omega \rangle. \quad (16)$$

If one applies this equation to the Fourier integral

$$A(x) = (2\pi)^{-\frac{3}{2}} \int d^3k e^{ikx} A(\mathbf{k})$$

and to the equation

$$\langle \omega_0 \mid A_\mu(x) \omega \rangle = (2\pi)^{-\frac{3}{2}} \int d^3k / (2|\mathbf{k}|)^{1/2} e^{ikx} \varphi_\mu(\mathbf{k}) \quad (17)$$

which follows from the definition of φ_μ , we get immediately the form $\varphi_\mu(\mathbf{k}) = k_\mu g(\mathbf{k})$.

Induction: We start with the equation

$$\langle \omega_0 \mid a_{\mu_1}(\mathbf{k}_1) \dots a_{\mu_n}(\mathbf{k}_n) \omega \rangle = k_{\mu_1} \dots k_{\mu_n} g(\mathbf{k}_1) \dots g(\mathbf{k}_n) \quad (18)$$

and discuss the term $\langle \omega_0 \mid a_{\mu_{n+1}}(\mathbf{k}_{n+1}) a_{\mu_1}(\mathbf{k}_1) \dots a_{\mu_n}(\mathbf{k}_n) \omega \rangle$.

Analogously to the method used in the case $n = 1$ we replace $a_{\mu_{n+1}}(\mathbf{k}_{n+1})$ in Eq. (18) by the potential $A_{\mu_{n+1}}(x)$. The curve integral analogous to Eq. (15) represents again a function of x alone, because the rotation of the integrand vanishes (Eq. 14). So all conclusions of the beginning of the induction can be repeated and one gets the desired form

$$\langle \omega_0 | a_{\mu_1}(\mathbf{k}_1) \cdots a_{\mu_{n+1}}(\mathbf{k}_{n+1}) \omega \rangle = k_{\mu_1} \cdots k_{\mu_{n+1}} g(\mathbf{k}_1) \cdots g(\mathbf{k}_{n+1}). \quad (19)$$

Since we can now write all components of ω with the help of a single scalar function g , we define with this g the Weyl operator $W\{kg\}$ and assert: $\omega = W\{kg\}\omega_0$.

In order to prove this it is sufficient to show that the components $\varphi, \varphi_{\mu_1, \dots, \mu_n}$ of ω result, if one calculates

$$[1/(n!)^{1/2}] \langle \omega_0 | a_{\mu_1}(\mathbf{k}_1) \cdots a_{\mu_n}(\mathbf{k}_n) W\{kg\} \omega_0 \rangle.$$

But this follows immediately from the familiar relationship (Appendix):

$$a_{\mu}(\mathbf{k}) W\{kg\} \omega_0 = k_{\mu} g(\mathbf{k}) W\{kg\} \omega_0 \quad (20)$$

and the equation

$$\langle \omega_0 | W\{kg\} \omega_0 \rangle = 1. \blacksquare$$

Theorem 2 effects the connection between the concept of gauges which is used here and the one used in Ref. 1. There we started from the Hilbert space \mathcal{H}^{tr} of "transversal" photons (our space $\mathcal{L} = \mathcal{L}_0^0$) and introduced the subspaces \mathcal{H}_g^0 in \mathcal{L}^0 by the definition $\mathcal{H}_g^0 = W\{kg\} \mathcal{H}^{\text{tr}}$. These spaces \mathcal{H}_g^0 are identical with our $\mathcal{L}_g^0 = W\{kg\} \mathcal{L}_0^0$. However, it was not discussed in Ref. 1, whether any embedding of $\mathcal{L}^0/\mathcal{N}^0$ in \mathcal{L}^0 can be reached from \mathcal{H}^{tr} by applying the Weyl operator $W\{kg\}$ to a suitable scalar function g . This gap is now filled.

3. THE MAXIMALITY AND UNIQUENESS OF THE LORENTZ SPACES \mathcal{L}^q

In the preceding section we considered the different embeddings $\mathcal{L}_1^q, \mathcal{L}_2^q, \dots$ of the quotient space $\mathcal{L}^q/\mathcal{N}^q$ in \mathcal{L}^q . Now we select for any $q \in \mathcal{Q}$ precisely one Hilbert space \mathcal{L}^q in any way. Because then any charge distribution determines its own \mathcal{L}^q the question arises whether it is possible to extend at least one of these spaces \mathcal{L}^q by the "adjunction" of suitable elements of \mathcal{S}_M in such a way that a subspace $\mathcal{u} \supset \mathcal{L}^q$ arises which is a Hilbert space, too. The following considerations will essentially show that this is impossible, so that the \mathcal{L}^q are maximal Hilbert spaces in \mathcal{S}_M .

Lemma 3a: Let \mathcal{L}^p and \mathcal{L}^q be two (different) Lorentz spaces and $\alpha \neq 0$ any element from \mathcal{L}^p . Then there exists an element $\beta \in \mathcal{L}^q$ such that $\langle \alpha | \beta \rangle \neq 0$.

In this way, two Lorentz spaces are never orthogonal to each other.

Proof: We exploit the isometric isomorphism between the Lorentz spaces (Appendix). First, the function $p = p(\mathbf{k})$ defines the four-vector $p^\mu = \delta_0^\mu p(\mathbf{k})/k_0$ and by it the Weyl operator $W\{p\}$ which has the property to depict some embedding \mathcal{L}_p^0 of $\mathcal{L}^0/\mathcal{N}^0$ in \mathcal{L}^0 onto the given \mathcal{L}^p :

$$\mathcal{L}^p = W\{p\} \mathcal{L}_p^0.$$

According to Theorem 2 there exists a gauge operator $W\{kg\}$ to some scalar function $g = g(\mathbf{k})$ so that $\mathcal{L}_p^0 = W\{kg\} \mathcal{L}_0^0$, where \mathcal{L}_0^0 is the space \mathcal{L} of "transversal" pho-

tons (Lemma 2c). Altogether we have $\mathcal{L}^p = W\{p + kg\} \mathcal{L}_0^0$. Consequently, there exists an element $\alpha_0 \in \mathcal{L}_0^0$ such that

$$\alpha = W\{p + kg\} \alpha_0. \quad (21)$$

By analogous conclusions there exists to $q^\mu = \delta_0^\mu q(\mathbf{k})/k_0$ and a suitable scalar function $h = h(\mathbf{k})$ the Weyl operator $W\{q + kh\}$ such that $\mathcal{L}^q = W\{q + kh\} \mathcal{L}_0^0$. Now let us define the element $\beta \in \mathcal{L}^q$ by the equation

$$\beta = W\{q + kh\} \alpha_0. \quad (22)$$

We will prove $\langle \alpha | \beta \rangle \neq 0$.

Up to a nonvanishing exponential we get first for $\langle \alpha | \beta \rangle = \langle \alpha_0 | W\{-p - hg\} W\{q + kh\} \alpha_0 \rangle$ the expression $\langle \alpha_0 | W\{q - p + k(h - g)\} \alpha_0 \rangle$. The four-vector f^μ which determines this Weyl operator obviously has the form $f^\mu = \delta_0^\mu f(\mathbf{k}) + k^\mu e(\mathbf{k})$ with scalar functions $f(\mathbf{k}), e(\mathbf{k})$ such that the exponent of this Weyl operator can be written as

$$\int d^3k \{ f^\mu a_\mu^+ - f_\mu^* a^\mu \} = \int d^3k \{ f(\mathbf{k}) a_0^+(\mathbf{k}) - f^*(\mathbf{k}) a_0(\mathbf{k}) + e(\mathbf{k}) L^+(\mathbf{k}) - e^*(\mathbf{k}) L(\mathbf{k}) \}. \quad (23)$$

In the scalar product $\langle \alpha_0 | W\{f\} \alpha_0 \rangle$, we separate now $W\{f\}$ into the factors

$$\exp \int d^3k f(\mathbf{k}) a_0^+(\mathbf{k}) \cdot \exp \int d^3k e(\mathbf{k}) L^+(\mathbf{k}) \cdot \exp \left[- \int d^3k e^*(\mathbf{k}) L(\mathbf{k}) \right] \cdot \exp \left[- \int d^3k f^*(\mathbf{k}) a_0(\mathbf{k}) \right] \quad (24)$$

up to a nonvanishing exponential. The exponentials on the right we apply to the element α_0 on the right. Because $L(\mathbf{k}) \alpha_0 = 0$ and $a_0(\mathbf{k}) = a^{(0)}(\mathbf{k}) \alpha_0 = 0$ this yields the element α_0 itself. In an analogous way we apply both first exponentials to the element $\langle \alpha_0 |$ on the left. So we reach a nonvanishing exponential $\langle \alpha_0 | \alpha_0 \rangle = \langle \alpha | \alpha \rangle \neq 0$ because $\alpha \neq 0$ is element of a Hilbert space. \blacksquare

Lemma 3b: Let \mathcal{L}^p and \mathcal{L}^q be two different Lorentz spaces and $\alpha \neq 0$ any element from \mathcal{L}^p . Then the Cauchy-Schwarz inequality cannot hold in the span $\text{sp}\{\mathcal{L}^q, \alpha\}$ which is the smallest subspace of \mathcal{S}_M containing \mathcal{L}^q and α . As a consequence, $\text{sp}\{\mathcal{L}^q, \alpha\}$ is necessarily indefinite.

Proof: We go back to the spaces \mathcal{L}^p and \mathcal{L}^q and will prove: If $\alpha \in \mathcal{L}^p$ is any element with $\langle \alpha | \alpha \rangle \neq 0$, then the Cauchy-Schwarz inequality does not hold in the space $\text{sp}\{\mathcal{L}^q, \alpha\}$.

We prove this statement in the following way: We show that there exists an element $v \in \mathcal{N}^q$ with the property $\langle v | \alpha \rangle \neq 0$. If, in contradiction to Lemma 3b, the Cauchy-Schwarz inequality were to hold on the span of \mathcal{L}^q and α , the above elements would satisfy

$$|\langle v | \alpha \rangle|^2 \leq \langle v | v \rangle \langle \alpha | \alpha \rangle = 0, \quad \text{so that } \langle v | \alpha \rangle = 0.$$

According to Lemma 3a there exists an element $\beta \in \mathcal{L}^q$ such that $\langle \alpha | \beta \rangle \neq 0$. Then the elements $\beta_0 = W\{-q\} \beta$ and $\alpha_0 = W\{-p\} \alpha$ are both elements from \mathcal{L}^0 . Now we define $v = W\{q\} \int d^3k h(\mathbf{k}) L^+(\mathbf{k}) \beta_0$, where $h(\mathbf{k})$ is any complex-valued, square-integrable function over R^3 . We show that $v \in \mathcal{N}^q$.

Proof:

$$\langle v | v \rangle = \int d^3k \int d^3k' h^*(\mathbf{k}) h(\mathbf{k}') \langle \beta_0 | L(\mathbf{k}) L^+(\mathbf{k}') \beta_0 \rangle = 0 \text{ be-}$$

cause of $[L(\mathbf{k}), L^+(\mathbf{k}')] = 0$ and $L(\mathbf{k})\beta_0 = 0$.

In the next step we show that $\nu \in \mathcal{L}^q$.

$$\begin{aligned} L(\mathbf{k}')\nu &= L(\mathbf{k}')W\{q\} \int d^3k h(\mathbf{k})L^+(\mathbf{k})\beta_0 \\ &= W\{q\}W\{-q\}L(\mathbf{k}')W\{q\} \int d^3k h(\mathbf{k})L^+(\mathbf{k})\beta_0 \\ &= W\{q\}L(\mathbf{k}') \int d^3k h(\mathbf{k})L^+(\mathbf{k})\beta_0 \\ &\quad + q(\mathbf{k}')W\{q\} \int d^3k h(\mathbf{k})L^+(\mathbf{k})\beta_0 \\ &= W\{q\} \int d^3k h(\mathbf{k})L^+(\mathbf{k})L(\mathbf{k}')\beta_0 + q(\mathbf{k}')\nu \\ &= q(\mathbf{k}')\nu. \end{aligned}$$

Let us now compute

$$\begin{aligned} \langle \nu | \alpha \rangle &= \int d^3k h^*(\mathbf{k}) \langle \beta_0 | L(\mathbf{k})W\{-q\} \alpha \rangle \\ &= \int d^3k h^*(\mathbf{k}) \langle \beta_0 | W\{-q\}W\{q\}L(\mathbf{k})W\{-q\} \alpha \rangle \\ &= \int d^3k h^*(\mathbf{k}) \{ \langle \beta | L(\mathbf{k})\alpha \rangle - q(\mathbf{k})\langle \beta | \alpha \rangle \} \\ &= \int d^3k h^*(\mathbf{k}) \{ p(\mathbf{k}) - q(\mathbf{k}) \} \langle \beta | \alpha \rangle. \end{aligned}$$

Because $\langle \beta | \alpha \rangle \neq 0$ and $p \neq q$ in the Lebesgue sense, we can always find a function $h(\mathbf{k})$, which is different from zero. So we have $\langle \nu | \alpha \rangle \neq 0$. This result is in contradiction to the presumed Cauchy-Schwarz inequality. ■

In particular, no Hilbert space ℓ^q can be extended to another Hilbert space by the adjunction of any element $\alpha \neq 0$ of some ℓ^p . Now we generalize this statement by dropping the condition $\alpha \in \ell^p$. However, we must include a condition about the expectation values of $L(\mathbf{x})$ which is physically motivated: As ℓ^q consists of eigenvectors of $L(\mathbf{x})$ with a possible charge density as eigenvalue function, a physically useful extension of ℓ^q will at least have to produce expectation values of $L(\mathbf{x})$ which can be interpreted as admitted charge densities.

Lemma 3c: Let ℓ^q be any Lorentz space and $\alpha \in \mathcal{S}_M$ any element so that $\alpha \notin \ell^q$, $\langle \alpha | \alpha \rangle \neq 0$ and $q_\alpha = q_\alpha(\mathbf{k}) := \langle \alpha | L(\mathbf{k})\alpha \rangle \in \mathcal{Q}$. Then $\text{sp}\{\ell^q, \alpha\}$ cannot be Hilbert space in \mathcal{S}_M .

Proof: We prove this by contradiction again and assume that $\mathcal{u} := \text{sp}\{\ell^q, \alpha\}$ is a Hilbert space. Then we can apply the familiar spectral theory to the operator $L(\mathbf{x})$ which is self-adjoint on the Hilbert space \mathcal{u} because $q_\alpha \in \mathcal{Q}$.

To each point $\mathbf{x} \in R^3$ there is then a special spectral family of projections $\{E^q(\mathbf{x}) | q \in R\}$ with the usual properties

$$\begin{aligned} \text{(a)} & E^q(\mathbf{x}) \leq E^p(\mathbf{x}) \quad \text{if } q < p, \\ \text{(b)} & E^{q+\epsilon}(\mathbf{x}) \rightarrow E^q(\mathbf{x}) \quad \text{for } \epsilon \rightarrow +0 \quad \text{and all } q \in R, \\ \text{(c)} & E^q(\mathbf{x}) \rightarrow 0 \quad \text{for } q \rightarrow -\infty; \quad E^q(\mathbf{x}) \rightarrow 1 \quad \text{for } q \rightarrow +\infty, \end{aligned} \quad (25)$$

so that $L(\mathbf{x})$ can be written in the form

$$L(\mathbf{x}) = \int_{-\infty}^{+\infty} q dE^q(\mathbf{x}). \quad (26)$$

Further on we introduce the additional projections

$$E^{q-}(\mathbf{x}) := \lim_{\epsilon \rightarrow +0} E^{q-\epsilon}(\mathbf{x}) \quad \text{and}$$

$$dE^q(\mathbf{x}) := E^q(\mathbf{x}) - E^{q-}(\mathbf{x}). \quad (27)$$

Up to this point q has been any real parameter, not a function in each case.

Now we choose any function $q = q(\mathbf{k}) \in \mathcal{Q}$. The real-valued function $q(\mathbf{x})$ is taken according to Eq. (3). To this function $q = q(\mathbf{x})$ we define now the projection E^q of \mathcal{u} in \mathcal{u} by the equation

$$E^q \mathcal{u} := \bigcap_{\substack{\mathbf{x} \in R^3 \\ q = q(\mathbf{x})}} E^q(\mathbf{x}) \mathcal{u}. \quad (28)$$

In an analogous way we define with the help of $E^{q-}(\mathbf{x})$ the projector E^{q-} and then $dE^q := E^q - E^{q-}$. Obviously we have

$$dE^q \mathcal{u} = \ell^q. \quad (29)$$

Because any projector $\mathcal{u} \rightarrow \mathcal{u}$ maps α either onto 0 or onto α itself, we consider for any given $\mathbf{x} \in R^3$ the set of real numbers

$$\mathfrak{B}(\mathbf{x}) = \{p \in R \mid E^p(\mathbf{x})\alpha = \alpha\}. \quad (30)$$

Each of these sets is not empty and has a lower bound because of (c) so that $\mathfrak{B}(\mathbf{x})$ contains a smallest number $p' = p'(\mathbf{x})$. Then for any $\mathbf{x} \in R^3$ we have

$$E^{p'(\mathbf{x})}(\mathbf{x})\alpha = \alpha, \quad \text{but} \quad E^{p'(\mathbf{x})-}(\mathbf{x})\alpha = 0. \quad (31)$$

This means, however

$$dE^{p'(\mathbf{x})}(\mathbf{x})\alpha = \alpha$$

and so

$$dE^{p'}\alpha = \bigcap_{\mathbf{x} \in R^3} dE^{p'(\mathbf{x})}(\mathbf{x})\alpha = \alpha. \quad (32)$$

So we get $\alpha \in \ell^{p'}$, for $p' \in \mathcal{Q}$. Because $p' \neq q$, Lemma 3b yields the assertion. ■

Theorem 3 (Maximality of Lorentz spaces): There exists no Hilbert space $\mathcal{u} \subset \mathcal{S}_M$ with the properties

- (a) \mathcal{u} contains one ℓ^q ($\ell^q \subset \mathcal{u}$),
- (b) every element $\alpha \in \mathcal{u}$ satisfies: $q_\alpha = q_\alpha(\mathbf{k})$

$:= \langle \alpha | L(\mathbf{k})\alpha \rangle \in \mathcal{Q}$. Among all Hilbert spaces of \mathcal{S}_M with property (b) the Lorentz spaces ℓ^q are consequently maximal.

Proof: If in contradiction to Theorem 3 any Hilbert space with the properties (a) and (b) exists, then choose any element $\alpha \in \mathcal{u}$ with $\alpha \notin \ell^q$ and construct the subspace $\text{sp}\{\ell^q, \alpha\}$ of \mathcal{u} . Lemma 3c shows that \mathcal{u} cannot be a Hilbert space. ■

In Sec. 2 we used the algebra $\underline{\Omega}^q$ of operators Ω such that $\Omega \mathcal{L}^q \subseteq \mathcal{L}^q$ and Ω is self-adjoint on \mathcal{L}^q (see Lemma 2a). Now we need also the algebra

$$\underline{\Omega} := \bigcap_{q \in \mathcal{Q}} \underline{\Omega}^q. \quad (33)$$

Every operator from $\underline{\Omega}$ maps any Lorentz space into itself and is self-adjoint on any Lorentz space. Examples in this section will show that $\underline{\Omega}$ is not empty.

Lemma 3d:

(a) $\underline{\Omega}$ is irreducible (see Lemma 2a) on any quotient space $\mathcal{L}^q / \mathcal{N}^q$ and so on any ℓ^q .

(b) Let ℓ^p and ℓ^q be two different Lorentz spaces and

choose $0 \neq \alpha \in \ell^p, 0 \neq \beta \in \ell^q$. Then there exists an operator $\Omega \in \underline{\Omega}$ so that the transition element $\langle \alpha | \Omega \beta \rangle$ admits no statistical interpretation (in the sense of Born).

Proof:

(a) We choose two elements $\alpha \neq 0$ and β from $\mathcal{L}^q / \mathcal{N}^q$ and have to construct an operator $P_{\alpha\beta} \in \underline{\Omega}$ such that $\beta = P_{\alpha\beta} \alpha$. Because of Lemma 2a we have a projector $\Pi_{\alpha\beta}$ with $\beta = \Pi_{\alpha\beta} \alpha$; but $\Pi_{\alpha\beta}$ is defined only on \mathcal{L}^q and not on any \mathcal{L}^p with $p \neq q$. To construct our $P_{\alpha\beta}$ with the help of $\Pi_{\alpha\beta}$ we take q fixed and choose any $p \in \underline{\Omega}$. There always exists a Weyl operator $W\{g\}$ (Appendix) such that $\mathcal{L}^q = W\{g\} \mathcal{L}^p$. The definition

$$P_{\alpha\beta} := W\{-g\} \Pi_{\alpha\beta} W\{g\} \quad (34)$$

gives a mapping of \mathcal{L}^p to any $p \in \underline{\Omega}$ into itself which is self-adjoint on \mathcal{L}^p and goes over to $\Pi_{\alpha\beta}$ if p is replaced by q .

(b) For the element $\alpha \in \ell^p$ we can find (Lemma 3a) an element $\alpha' \in \ell^q$ such that the Cauchy-Schwarz inequality doesn't hold with respect to the transition element $\langle \alpha | \alpha' \rangle$. Because of (a) we have an operator $\Omega \in \underline{\Omega}$ such that $\alpha' = \Omega \beta$. In this way, $\langle \alpha | \alpha' \rangle = \langle \alpha | \Omega \beta \rangle$ admits no statistical interpretation. ■

So far the Lorentz spaces have two characteristics, which will become decisive for the following considerations: They are maximal Hilbert spaces in \mathcal{S}_M , in which $L(\mathbf{x})$ is self-adjoint, and their family is complete in so far as each square-integrable charge density of total charge 0 is represented by precisely one of them. These two properties are considered sufficient for the following considerations, which will attempt to found the theory, as far as it concerns \mathcal{S}_M , only on these ℓ^q (respectively, \mathcal{L}^q). The limitation $Q = 0$ should not become too important, for if in a physical process $Q \neq 0$, then this process should not be influenced when in astronomical distances the charge $-Q$ is fixed so that the whole charge vanishes.

If one accepts the exclusive use (an additional cause for this is given by Theorem 4 later on) of the ℓ^q , then one should preferably also use such operators on \mathcal{S}_M as are suitable for this subspace structure of \mathcal{S}_M . First of all the operators of our algebra $\underline{\Omega}$ (which map any Lorentz space into itself and are self-adjoint on it) should play a special role. But Lemma 3d shows that $\langle \alpha | \Omega \beta \rangle$ with a suitable $\Omega \in \underline{\Omega}$ admits no statistical interpretation, if α, β are taken from different Lorentz spaces. So we have to be careful that only elements α, β from one and the same space ℓ^q are used to compute scalar products $\langle \alpha | \Omega \beta \rangle$. On the other hand, if an operator Ω had the property that $\alpha \in \ell^q$ but $\Omega \alpha \notin \ell^q$, then the transition element $\langle \alpha | \Omega \alpha \rangle$ would be computed in the space $\text{sp}\{\ell^q, \alpha\}$, i.e., in a space in which the Cauchy-Schwarz inequality does not generally hold, which is necessary for a statistical interpretation. Therefore we define for the time observables on \mathcal{S}_M alone according to the requirements of the Lorentz spaces in \mathcal{S}_M . This definition should not refer to the ℓ^q , however, but to the spaces \mathcal{L}^q , because ℓ^q is not physically determined in a unique way. When a quantized four-current is introduced later on we will have to modify this temporary concept of observables.

Definition 2: An operator Ω is called *observable*, if there

exists any nonempty subset $\underline{\Omega}_\Omega \subseteq \underline{\Omega}$ such that to every $q \in \underline{\Omega}_\Omega$ we have

- (i) $\Omega \mathcal{L}^q \subseteq \mathcal{L}^q$.
 - (ii) Ω is self-adjoint on \mathcal{L}^q .
- (35)

A trivial example of such an observable is the operator $L(\mathbf{x})$ of charge density with $\underline{\Omega}_{L(\mathbf{x})} = \underline{\Omega}$.

Observables Ω with $\underline{\Omega}_\Omega = \underline{\Omega}$ (our algebra $\underline{\Omega}$!) could be given a special name, e.g., global observables, but we will not do so because of the temporary definition of the set $\underline{\Omega}$ (cf. Sec. 5). Lemma 2a can now be extended to

Lemma 3e:

- (a) Any observable Ω can be defined on all quotient spaces $\mathcal{L}^q / \mathcal{N}^q$ and so on all embedding spaces ℓ^q , if $q \in \underline{\Omega}_\Omega$.
- (b) Any observable has gauge-invariant matrix elements on every \mathcal{L}^q with $q \in \underline{\Omega}_\Omega$.
- (c) Let q be any fixed function from $\underline{\Omega}$. Then the set of all observables Ω with $q \in \underline{\Omega}_\Omega$ is an irreducible Lie algebra on $\mathcal{L}^q / \mathcal{N}^q$ and so on any embedding space ℓ^q .

It follows from the definition of Lorentz spaces that each observable Ω commutes on $\mathcal{L}^q (q \in \underline{\Omega}_\Omega)$ with $L(\mathbf{k})$, respectively, $L(\mathbf{x})$. Thus one gets a practical criterion to test in concrete cases whether a given operator is an observable according to definition 2 and how the function set $\underline{\Omega}_\Omega$ has to be selected.

Examples:

(a) We define the electromagnetic field tensor $F_{\mu\nu}(\mathbf{x})$ in the Schrödinger picture by

$$F_{\mu\nu}(\mathbf{x}) := i(2\pi)^{-\frac{3}{2}} \int \frac{d^3k}{(2|\mathbf{k}|)^{1/2}} e^{i\mathbf{k}\mathbf{x}} \{a_\mu(\mathbf{k})k_\nu - a_\nu(\mathbf{k})k_\mu\} + cc.$$

Because of $[F_{\mu\nu}(\mathbf{x}), L(\mathbf{k})] = 0$ the field tensor $F_{\mu\nu}(\mathbf{x})$ is (global) observable without limitation of its domain, so that $\underline{\Omega}_{F_{\mu\nu}} = \underline{\Omega}$.

(b) The canonical variables

$$A_\mu(\mathbf{x}) = (2\pi)^{-\frac{3}{2}} \int \frac{d^3k}{(2|\mathbf{k}|)^{1/2}} e^{i\mathbf{k}\mathbf{x}} a_\mu(\mathbf{k}) + cc,$$

$$\Pi_\mu(\mathbf{x}) = -i(2\pi)^{-\frac{3}{2}} \int d^3k (|\mathbf{k}|/2)^{1/2} e^{i\mathbf{k}\mathbf{x}} a_\mu(\mathbf{k}) + cc \quad (36)$$

are, in contrast, not observables as they don't commute with $L(\mathbf{k})$. The fact that they are not observables is connected with the gauge problem (Sec. 2).

(c) Usually (cf., however, Sec 6) the operator $n(\mathbf{k})$ of photon number is defined as

$$n(\mathbf{k}) := a_\mu^+(\mathbf{k}) a^\mu(\mathbf{k}). \quad (37)$$

Because of $[L(\mathbf{k}), n(\mathbf{k}')] = L(\mathbf{k}) \delta(\mathbf{k} - \mathbf{k}')$ this commutator is zero only on L^0 so that $\underline{\Omega}_{n(\mathbf{k})} = \{0\}$. Accordingly, the free Hamiltonian

$$H_0 := \int d^3k |\mathbf{k}| a_\mu^+(\mathbf{k}) a^\mu(\mathbf{k}) \quad (38)$$

is an observable only if its domain is restricted to L^0 , which is the Lorentz space of the interaction-free theory.

(d) In order to test our concept of observables in a more rigorous way we introduce the often mentioned pseudointeraction between M and any prescribed, classical, explicitly time-dependent four-current $j_\mu(\mathbf{x}, t)$. With the Fourier

components

$$j_\mu(\mathbf{x}, t) = (2\pi)^{-\frac{3}{2}} \int d^3k \sqrt{2|\mathbf{k}|} e^{i\mathbf{k}\cdot\mathbf{x}} j_\mu(\mathbf{k}, t), \quad (39)$$

the Hamiltonian

$$H(t) := H_0 + \int d^3x j_\mu(\mathbf{x}, t) A^\mu(\mathbf{x}) \quad (40)$$

takes the form

$$H(t) = H_0 + \int d^3k \{ j_\mu^*(\mathbf{k}, t) a^\mu(\mathbf{k}) + \tilde{j}^\mu(\mathbf{k}, t) a_\mu^+(\mathbf{k}) \}. \quad (41)$$

Because of

$$\begin{aligned} [L(\mathbf{k}), H(t)] &= |\mathbf{k}| L(\mathbf{k}) + k^\mu j_\mu(\mathbf{k}, t) \\ &= |\mathbf{k}| \{ L(\mathbf{k}) - j_0(\mathbf{k}, t) \}, \end{aligned} \quad (42)$$

where in the last step the charge conservation has been used, the Hamiltonian $H(t)$ is an observable only if its domain is restricted to the Lorentz space \mathcal{L}^q with $q = j_0$. So we get $\mathcal{Q}_{H(t)} = \{j_0(\mathbf{k}, t)\}$. If the time t is fixed, then the momentary value of the charge density selects "its" Lorentz space \mathcal{L}^q with $q = j_0$ as the correct state space of the pseudo-interaction (cf. Theorem 1d). With the time running, the state space $\mathcal{L}^{q(t)}$ (where $q(t) = q[t](\mathbf{k}) = j_0(\mathbf{k}, t)$) denotes the "curve" of functions in the set of \mathcal{Q} wanders, however, in \mathcal{S}_M . Contradictions with respect to Lemma 3d can never occur in the calculation of transition probabilities because, according to Definition 2, any observable Ω maps $\mathcal{L}^{q(t)}$ into $\mathcal{L}^{q(t)}$, so that the theory at each moment t remains limited to the momentary $\mathcal{L}^{q(t)}$. The wandering of the state space $\mathcal{L}^{q(t)}$ can also¹ be written in the form

$$\mathcal{L}^{q(t)} = U(t) \mathcal{L}^{q(0)}, \quad (43)$$

where $U(t)$ is the time evolution operator determined by the Hamiltonian of Eq. (41).

Now we prescribe in any Lorentz space \mathcal{L}^q precisely one embedding ℓ^q in any way. By this, the family of Hilbert spaces $\ell^{q(t)}$ is defined, but $U(t)\ell^{q(0)}$ will in general not be identical with the chosen embedding $\ell^{q(t)}$. According to Theorem 2 there must, however, exist a scalar function $g[t] = g[t](\mathbf{k})$ in such a way that

$$\ell^{q(t)} = \mathcal{W}\{kg[t]\} U(t)\ell^{q(0)}. \quad (44)$$

In this way the definite selection of embedding spaces ℓ^q in \mathcal{L}^q has necessarily the consequences that gauge operators $\mathcal{W}\{kg[t]\}$ will appear.

Up to now we confined our discussion to Lorentz spaces. Could it not be that there are other Hilbert spaces \mathcal{u} in \mathcal{S}_M which are important for physics also? But what does "important for physics" mean? Certainly, such a space \mathcal{u} should allow us to describe at least the pseudo-interaction between M and any suitably chosen, classical, stationary four-current $j_\mu(\mathbf{x})$. Therefore, \mathcal{u} should contain an element $\alpha \neq 0$ and with it the whole Schrödinger curve $U(t)\alpha$ drawn through α , where $U(t)$ is the time evolution operator determined by the Hamiltonian of Eq. (41) to the four-current $j_\mu(\mathbf{x})$. On the other hand, all (global) observables Ω with $\mathcal{Q}_\Omega = \mathcal{Q}$, especially the operator $L(\mathbf{x})$ of charge density, should also play a role on \mathcal{u} , i.e., they should map \mathcal{u} into itself and should be self-adjoint on \mathcal{u} . So the physical theory on \mathcal{u} is

not essentially poorer than the theory on any ℓ^q . Finally, these observables Ω together with the Hamiltonian H above, should admit a quantum theory (in the Schrödinger picture), at least on the states $U(t)\alpha$. This requires, that to any $\Omega \in \mathcal{Q}$ the operator $\dot{\Omega} := i[\Omega, H]$ also is defined on these states and has the meaning of a time derivative of Ω . These seem the minimal demands for a physically important state space in our case.

Theorem 4 (Uniqueness of Lorentz spaces): Any Hilbert space $\mathcal{u} \subset \mathcal{S}_M$ which is "important for physics" in the above sense is necessarily one of the Lorentz spaces ℓ^q .

Proof: We commute the time derivation [Eqs. (42) and (6)]

$$\begin{aligned} \dot{L}(\mathbf{x}) = i[L(\mathbf{x}), H] &= i(2\pi)^{-\frac{3}{2}} \int d^3k (2|\mathbf{k}|)^{1/2} |\mathbf{k}| e^{i\mathbf{k}\cdot\mathbf{x}} \\ &\quad \times \{ L(\mathbf{k}) - j_0(\mathbf{k}) \}. \end{aligned} \quad (45)$$

Because of the stationary charge distribution, $\dot{L}(\mathbf{x})$ is the zero operator on the element α . This means that $L(\mathbf{k})\alpha = j_0(\mathbf{k})\alpha$ for almost any \mathbf{k} so that $\alpha \in \mathcal{L}^q (q = j_0)$. Then there exists precisely one embedding ℓ^q which contains α . Because of the inseparability of the algebra $\underline{\mathcal{Q}}$ (Lemma 3d) \mathcal{u} then contains with α every element of ℓ^q , so that $\ell^q \subseteq \mathcal{u}$. At last, $\mathcal{u} = \ell^q$ follows from the maximality of ℓ^q . ■

4. CLASSICAL ELECTRODYNAMICS FROM THE POINT OF VIEW OF THE LORENTZ SUBSPACE STRUCTURE OF \mathcal{S}_M

As is well known, classical electrodynamics (CED) have a formal deficiency: the Hamilton equations of motion of the classical potential $A_\mu(\mathbf{x})$ and its canonical momentum $\Pi_\nu(\mathbf{x})$ will not generally lead to Maxwell's equations. The Hamilton functional reads

$$\begin{aligned} H(x_0) &= \frac{1}{2} \int d^3x \{ \Pi^\nu(x) \Pi_\nu(x) + [\nabla A^\nu(x)] [\nabla A_\nu(x)] \} \\ &\quad + \int d^3x j^\nu(x) A_\nu(x). \end{aligned} \quad (46)$$

$j^\nu(x) = j^\nu(x_0, \mathbf{x})$ denotes the components of any given, conserved, four-current. $A_\nu(x)$ is the four-potential, $\Pi^\nu(x)$ its canonical momentum. The Hamilton equations of motion are

$$\frac{\partial}{\partial x_0} A_\nu(x) = \delta H(x_0) / \delta \Pi^\nu = \Pi_\nu(x), \quad (47)$$

$$\frac{\partial}{\partial x_0} \Pi_\nu(x) = -\delta H(x_0) / \delta A^\nu = -j_\nu(x) + \nabla^2 A_\nu(x).$$

They yield by iteration

$$\square A_\nu(x) = j_\nu(x). \quad (48)$$

But, the electromagnetic field tensor

$$F_{\mu\nu}(x) := \partial_\mu A_\nu(x) - \partial_\nu A_\mu(x) \quad (49)$$

doesn't immediately satisfy Maxwell's equations

$$\partial^\mu F_{\mu\nu}(x) = j_\nu(x). \quad (50)$$

Instead, we arrive only at

$$\begin{aligned} \partial^\mu F_{\mu\nu}(x) &= \partial^\mu \partial_\mu A_\nu(x) - \partial^\mu \partial_\nu A_\mu(x) \\ &= \square A_\nu(x) - \partial_\nu \partial^\mu A_\mu(x) = j_\nu(x) - \partial_\nu [\partial^\mu A_\mu(x)]. \end{aligned} \quad (51)$$

Only those solutions, which at the same time satisfy the classical Lorentz condition (CLC)

$$\partial^\mu A_\mu(x) = 0, \quad (52)$$

which demands a special gauge (Lorentz gauge), also satisfy Maxwell's equations and are, therefore, called "physical" solutions.

Thus, CLC is added from outside to the Hamilton equations of motion and is motivated only by the result which it effects. A deeper understanding, and with this a derivation of CLC, has not yet been achieved.

Now the first step from CED to QED is the theory of pseudointeraction between the quantized Maxwell field M and any prescribed, classical, four-current, which is used in this paper as a heuristic tool. This theory can and must be described on the family of Lorentz spaces, as we have seen. In order to get these Lorentz spaces, the historical way via generalizations of CLC is no longer necessary. On the contrary, the Lorentz spaces can now be characterized on their own through their outstanding mathematical and physical properties (Sec. 3). In reversal of the historical way these characteristics of the Lorentz spaces should now permit a derivation and with this a deeper understanding of CLC. In order to achieve this we only have to take a "step back" from the theory of pseudointeraction to CED.

Let $A_\mu^-(x)$ be the part of the operator of Eq. (36) which consists only of photon annihilators and define (Schrödinger picture) $\mathcal{P}A_0^-(x) = i[H(x_0), A_0^-(x)]$, where $H(x_0) = H(t)$ is to be taken from Eq. (41).

Theorem 5: The operator identity $\partial^\nu A_\nu^-(x) = 0$ holds on the family $\{\mathcal{L}^q | q = j_0(x)\}$ of Lorentz spaces. If the theory of pseudointeraction is restricted to these spaces, the Lorentz condition in this theory is automatically satisfied. Returning from this theory to CED we get automatically CLC. Conversely, CLC is necessary in CED in order to permit a transition from CED to the theory of pseudo-interaction and further on to QED.

Proof: $H(x_0)$ is, according to Sec. 3, example d, an observable only if its domain is limited to the Lorentz space $\mathcal{L}^{q(x_0)}$ belonging to $q[x_0] = j_0(\mathbf{k}, x_0)$. The time evolution operator $U(x_0)$ which belongs to $H(x_0)$ has the property (Eq. (43)) $\mathcal{L}^{q(x_0)} = U(x_0)\mathcal{L}^{q(0)}$. The Schrödinger curve $\alpha(x_0)$ through any element $\alpha \in \mathcal{L}^{q(0)}$ is consequently given by $\alpha(x_0) = U(x_0)\alpha \in \mathcal{L}^{q(x_0)}$ so that $L(\mathbf{k})\alpha(x_0) = q[x_0](\mathbf{k})\alpha(x_0)$ for almost any \mathbf{k} .

A simple calculation yields

$$\partial^0 A_0^-(x) = i(2\pi)^{-\frac{3}{2}} \int \frac{d^3k}{(2|\mathbf{k}|)^{1/2}} e^{ikx} \{k^0 \alpha_0(\mathbf{k}) - q[x_0](\mathbf{k})\}.$$

Consequently, we arrive at

$$\partial^\nu A_\nu^-(x) = i(2\pi)^{-\frac{3}{2}} \int \frac{d^3k}{(2|\mathbf{k}|)^{1/2}} e^{ikx} \{L(\mathbf{k}) - q[x_0](\mathbf{k})\}. \quad (53)$$

Because of $L(\mathbf{k})\alpha(x_0) = q[x_0](\mathbf{k})\alpha(x_0)$ we get the assertion. ■

As a consequence, an additional Lorentz condition becomes superfluous in quasiclassical electrodynamics on the family of Lorentz spaces. Thus, CED finds its formal completion in these quasiclassical electrodynamics, which are based essentially on the concept of photons.

In this way, the described lack of CED is repaired by the canonical quantization of M in a surprising manner. Conversely CLC anticipates, in some sense, the concept of photons!

With these considerations we can, by the way, also understand why the generalization of a seemingly unimportant additional condition to CED could result in a criterion for the selection of Hilbert spaces in \mathcal{S}_M , i.e., for the solution of a purely mathematical problem.

5. THE IMPORTANCE OF LORENTZ SPACES FOR THE STATE SPACE OF QED

We are looking for a state space of the theory of interaction between the quantized Maxwell field M and the quanta of a four-current $J_\mu(x)$, given as an operator in the Schrödinger picture, e.g., the electrons and positrons $b_{r^+}(\mathbf{k})v, d_{r^+}(\mathbf{k})v$ ($v = \text{vacuum}, r = 1, 2$) of the Dirac four-current $J_\mu(x) = e\bar{\psi}(x)\gamma_\mu\psi(x)$. The creation operators b_{r^+}, d_{r^+} span the Fock space \mathcal{S}_J in the usual way, and this space must be a partner of \mathcal{S}_M in some way.

The first choice for the state space of the interacting system $J + M$ is the tensor product $\mathcal{S}_J \otimes \mathcal{S}_M$ of the partner spaces. However, two elements $a \otimes \alpha, b \otimes \beta$ ($a, b, \in \mathcal{S}_J, \alpha, \beta \in \mathcal{S}_M$) of this tensor product is ordered to the scalar product

$$\langle a \otimes \alpha | b \otimes \beta \rangle := \langle a | b \rangle \langle \alpha | \beta \rangle. \quad (54)$$

Because of the special properties of the Lorentz spaces, the scalar product $\langle \alpha, \beta \rangle$ which appears here is responsible for the fact that in general the Cauchy-Schwarz inequality cannot hold in $\mathcal{S}_J \otimes \mathcal{S}_M$, so that a statistical interpretation in the sense of Born is impossible. Therefore, the state space must be constructed in such a way that only the Hilbert spaces \mathcal{L}^q (definitely selected in some way) or, more generally, the spaces \mathcal{L}^q of \mathcal{S}_M are used, and that "mixtures" of elements from different Lorentz spaces can never appear when scalar products are computed. Consequently the state space to be constructed must consist of "sectors"

$$\mathcal{S}^q := \mathcal{L}^q \otimes \mathcal{L}^q, \quad (55)$$

in which \mathcal{L}^q is a definitely selected Hilbert space in \mathcal{S}_M . The closest possibility to define the partner space \mathcal{L}^q consists in introducing it as the eigenspace of the operator $J_0(x)$ of charge density to the eigenvalue function $q = q(x)$ according to Eq. (3). As an alternative to this definition we could also use the Fourier component $J_0(\mathbf{k})$, which is given through the more general definition

$$J_\mu(x) = (2\pi)^{-\frac{3}{2}} \int d^3k (2|\mathbf{k}|)^{1/2} e^{ikx} J_\mu(\mathbf{k}). \quad (56)$$

Let $J_\mu(x)$ be self-adjoint on \mathcal{S}_J so that

$$J_\mu^+(\mathbf{k}) = J_\mu(-\mathbf{k}). \quad (57)$$

The Hamiltonian H which describe the interaction between J and M will then take the form

$$H = H_0 + \int d^3k \{J_\mu^+(\mathbf{k})a^\mu(\mathbf{k}) + J^\mu(\mathbf{k})a_\mu^+(\mathbf{k})\}, \quad (58)$$

in which H_0 is the free part. We are going to introduce the term $Q(x) := J_0(x)$, respectively $Q(\mathbf{k}) := J_0(\mathbf{k})$, for the free-

quently occurring operator of the charge density. To any $q \in \mathcal{Q}$ the space \mathcal{K}^q is then defined by

$$\mathcal{K}^q := \{a \in \mathcal{S}_J \mid Q(\mathbf{k})a = q(\mathbf{k})a \text{ for almost any } \mathbf{k} \in \mathbb{R}^3\}. \quad (59)$$

We must further on take care that all parameter functions $q \in \mathcal{Q}$ in the state space to be constructed really play a role, but on the other hand, that products between elements from different sectors \mathcal{S}^q will never appear. Thus we are almost forced to define the positive-definite state space of the interaction theory of the system $J + M$ as the direct sum of the \mathcal{S}^q :

$$\mathcal{S} := \bigoplus_{q \in \mathcal{Q}} \mathcal{S}^q = \bigoplus_{q \in \mathcal{Q}} \mathcal{K}^q \otimes \mathcal{L}^q. \quad (60)$$

The introduction of the direct sum means the introduction of a new scalar product, which, however, is reduced to the already-defined scalar product when we limit ourselves to a single sector \mathcal{S}^q . Each tensor product $\mathcal{K}^q \otimes \mathcal{L}^q$ expresses from the beginning the inseparability of a possible charge density from its photon robe. Furtheron, \mathcal{S}^q seems to demonstrate a close connection between the quantization of M and charge quantization.

Of course, the direct sum \mathcal{S} is mathematically not defined as such, because of the continuous set \mathcal{Q} . In order to express this it would be advisable to use the sign $\int_{\mathcal{S}^q}$ instead of \bigoplus . An element $A \in \mathcal{S}$ is generally given by a continuous set of components $A^q \in \mathcal{S}^q$ and could be written as $A = \int_{\mathcal{S}^q} A^q$. But we will neither discuss here what such a continuous representation of any element $A \in \mathcal{S}$ means physically, nor whether a theory of the continuous direct sums with reference to the Lorentz spaces can be established in a satisfactory manner. We regard \mathcal{S} here only as a formal conglomerate of the sectors \mathcal{S}^q .

Our definition of the family of the \mathcal{S}^q and so the state space \mathcal{S} depends decisively on our choice of the function set \mathcal{Q} . \mathcal{Q} has been introduced according to the requirements of the theory of pseudo-interaction between M and any given, square-integrable four-current. Therefore, \mathcal{Q} will not even describe the soft-photon effects¹⁰⁻¹³ connected with a classical four-current. Kibble¹³ gives some remarks (p. 321 of his work) which would lead us possibly to a certain extension of the space \mathcal{S}_M and of the set \mathcal{Q} so that the infrared problems could be handled on the basis of the supplementarily-introduced sectors \mathcal{S}^q . But in this paper we will not do so; we intend here only to discuss the benefit which the Lorentz spaces \mathcal{L}^q and the sectors \mathcal{S}^q bring to the (temporarily restricted) QED. By the way, Kibble¹³ does use only the Lorentz space \mathcal{L}^0 , especially the coherent states from \mathcal{L}^0 . This is in accordance with his intentions to give a (soft-photon processes including) S -Matrix theory of four-currents. Because the in- and out-vectors of any scattering theory are free states, Kibbles's states belong to \mathcal{L}^0 because they are taken as coherent.

However, our set \mathcal{Q} is too small not only with respect to soft-photon effects. We introduce here also the eigenspace \mathcal{K}^q of the charge operator $Q(\mathbf{k})$ in its Fock space \mathcal{S}_J . In general, $Q(\mathbf{k})$ will admit eigenfunctions which do not belong to \mathcal{Q} . Also from this fact the necessity arises to extend \mathcal{Q} to some set $\mathcal{Q}^* \supset \mathcal{Q}$ so that the enlarged state space

$$\mathcal{S}^* := \bigoplus_{q \in \mathcal{Q}^*} \mathcal{S}^q \quad (61)$$

possibly becomes the state space of the full QED. Obviously, this enlargement causes a lot of trouble, which we are not willing to discuss here.

Summarizing, our space \mathcal{S} defined by the index set \mathcal{Q} will represent not the full QED (provided that QED can be regarded as a self-consistent, full theory at all) but, in some sense, a "hard core" of it. If in this restricted QED the concept of the sectors \mathcal{S}^q proves true, the same concept with additionally-introduced sectors $\mathcal{S}^q (q \in \mathcal{Q}^*)$ in possibly enlarged spaces $\mathcal{S}_J^*, \mathcal{S}_M^*$ instead of our $\mathcal{S}_J, \mathcal{S}_M$ will also be important with respect to full QED.

We will now discuss some advantages of our state space \mathcal{S} and have especially to show that our definition of \mathcal{S} does not contradict the time evolution operator $U(t)$ which is given through the Hamiltonian of Eq. (58).

Before we turn to this question we give the new concept of observables as it follows necessarily from our construction of \mathcal{S} . If there is given any operator $\Omega: \mathcal{S}^q \rightarrow \mathcal{S}^q$, i.e., $\Omega(a \otimes \alpha) = a' \otimes \alpha' (a, a' \in \mathcal{K}^q; \alpha, \alpha' \in \mathcal{L}^q)$, then we define the operators $\Omega_k: \mathcal{K}^q \rightarrow \mathcal{K}^q$, respectively $\Omega_l: \mathcal{L}^q \rightarrow \mathcal{L}^q$, by the equations $\Omega_k a := a'$, respectively $\Omega_l \alpha := \alpha'$, and write $\Omega = \Omega_k \otimes \Omega_l$. We write the identical mappings in any \mathcal{K}^q , respectively \mathcal{L}^q , in the form 1_k , respectively 1_l . In this way we can write, e.g.,

$$L(\mathbf{k}) = 1_k \otimes L(\mathbf{k}) \text{ and } Q(\mathbf{k}) = Q(\mathbf{k}) \otimes 1_l. \quad (62)$$

Definition 3: An operator $\Omega = \Omega_k \otimes \Omega_l$ is called *observable* if there exists any nonempty subset $\mathcal{Q}_\Omega \subseteq \mathcal{Q}$ such that to every $q \in \mathcal{Q}_\Omega$ we have

- (i) Ω_l is observable in the sense of Definition 2 ($\mathcal{Q}_\Omega, := \mathcal{Q}_\Omega$).
- (ii) $\Omega_k \mathcal{K}^q \subseteq \mathcal{K}^q$.
- (iii) Ω_k is self-adjoint on \mathcal{K}^q .

From this definition follows immediately: If $\Omega = \Omega_k \otimes \Omega_l$ is any observable and $q \in \mathcal{Q}_\Omega$, we have

$$[\Omega_k, Q(\mathbf{k})]_{\mathcal{K}^q} = 0 \text{ and } [\Omega_l, L(\mathbf{k})]_{\mathcal{L}^q} = 0. \quad (63)$$

In particular, any observable Ω commutes on any sector $\mathcal{S}^q (q \in \mathcal{Q}_\Omega)$ with the operator of charge density. Thus one theorem of Strocchi and Wightman⁷ follows here directly from the construction of the state space \mathcal{S} , the definition of which is basically a consequence of the canonical quantization of the Maxwell field.

It should be noted, by the way, that the Hamiltonian H of Eq. (58) will not be an observable in our sense, otherwise the time-evolution operator $U(t)$ determined by H would have to map every sector \mathcal{S}^q with $q \in \mathcal{Q}_H$ into itself, so that the expectation value of the charge density would be constant in time.

Before we come to the question of whether the time-evolution operator $U(t)$ doesn't contradict our concept of \mathcal{S} we will show that the Lorentz condition, as it is otherwise necessary in QED, and in connection with it the charge conservation, are automatically valid in any QED on the basis of our state space \mathcal{S} . Thus any QED on \mathcal{S} is a *constraint-free* theory, because selection rules, as they are otherwise represented by the Lorentz condition or the charge conservation, are superfluous. Both conditions will not appear here even in

a hidden form, for the construction of \mathcal{S} and the definition of observables which is closely connected with it are based only on the outstanding properties of the Lorentz spaces; i.e., it goes in principal back to the canonical quantization of the Maxwell field. In this connection it is not important that the way to these Lorentz spaces has been shown heuristically by the classical Lorentz condition.

Let $A_{\mu}^{-}(\mathbf{x})$ be the part of the operator of Eq. (36) which consists of only photon annihilators and define (Schrödinger picture) $\partial^0 A_0^{-}(\mathbf{x}) := i[H, A_0^{-}(\mathbf{x})]$, where H is to be taken from Eq. (58). Similarly define $\partial^0 J_0(\mathbf{x}) := i[H, J_0(\mathbf{x})]$.

Theorem 6: The Lorentz condition

$$\partial^{\nu} A_{\nu}^{-}(\mathbf{x}) = 0 \quad (64)$$

and the charge conservation

$$\partial^{\nu} J_{\nu}(\mathbf{x}) = 0 \quad (65)$$

hold automatically as operator identities on \mathcal{S} .

Proof: With the use of only the commutation relations of the photon creation and annihilation operators one calculates

$$\partial^0 A_0^{-}(\mathbf{x}) = i(2\pi)^{-3} \int \frac{d^3 k}{(2|\mathbf{k}|)^{1/2}} e^{i\mathbf{k}\mathbf{x}} \{k^0 a_0(\mathbf{k}) - Q(\mathbf{k})\}.$$

Together with the formal derivation $\partial^{\nu} A_{\nu}^{-}(\mathbf{x})$ for $r = 1, 2, 3$ we get

$$\partial^{\nu} A_{\nu}^{-}(\mathbf{x}) = i(2\pi)^{-3} \int \frac{d^3 k}{(2|\mathbf{k}|)^{1/2}} e^{i\mathbf{k}\mathbf{x}} \{L(\mathbf{k}) - Q(\mathbf{k})\}. \quad (66)$$

Now, $L(\mathbf{k}) - Q(\mathbf{k})$ is the zero operator on any \mathcal{S}^q and so on \mathcal{S} .

The commutation relations of the photon creation and annihilation also yield

$$[L(\mathbf{k}), H] = k_0 L(\mathbf{k}) + k^{\nu} J_{\nu}(\mathbf{k}), \quad (67)$$

from which follows (for $r = 1, 2, 3$)

$$\begin{aligned} \partial^0 J_0(\mathbf{x}) = & -i(2\pi)^{-3} \int d^3 k (2|\mathbf{k}|)^{1/2} e^{i\mathbf{k}\mathbf{x}} \{k_0 [L(\mathbf{k}) - Q(\mathbf{k})] \\ & + k^{\nu} J_{\nu}(\mathbf{k})\}. \end{aligned} \quad (68)$$

Because of $L(\mathbf{k}) - Q(\mathbf{k}) = 0$ on \mathcal{S} and the formal derivation

$$\partial^{\nu} J_{\nu}(\mathbf{x}) = i(2\pi)^{-3} \int d^3 k (2|\mathbf{k}|)^{1/2} e^{i\mathbf{k}\mathbf{x}} k^{\nu} J_{\nu}(\mathbf{k}) \quad (r = 1, 2, 3)$$

one gets immediately $\partial^{\nu} J_{\nu}(\mathbf{x}) = 0$ on \mathcal{S} . ■

Now we have to discuss whether the time evolution operator $U(t)$ doesn't contradict our definition of \mathcal{S} . We will show that this concept is consistent with time evolution if the function set Ω is enlarged to some set $\Omega^* \supset \Omega$ as discussed above.

Let Ω^* be a set of functions $q = q(\mathbf{k})$ such that Ω^* contains our set Ω and all eigenfunctions of the operator $Q(\mathbf{k})$ on \mathcal{S}_J . Generalize also the definition on \mathcal{S}^q in such a way that any function $q \in \Omega^*$ may also appear as index. Then we prove:

Theorem 7: Start at $t = 0$ from any element

$a \otimes \alpha \in \mathcal{K}^q \otimes \mathcal{L}^q = \mathcal{S}^q$, where $q = :q[0]$ is chosen from Ω^* . Then to any $t > 0$ there exists a function $q[t]$ from Ω^* such that

$$U(t)(a \otimes \alpha) \in \mathcal{S}^{q[t]}. \quad (69)$$

Proof: First we prove for any $t > 0$ the equation

$$L(\mathbf{k})U(t)_{\mathcal{S}^{q[0]}} = Q(\mathbf{k})U(t)_{\mathcal{S}^{q[0]}}. \quad (70)$$

We use charge conservation (Theorem 6) $\partial^{\nu} J_{\nu}(\mathbf{x}) = 0$ on \mathcal{S} . Because we work in the Schrödinger picture, this means ($r = 1, 2, 3$):

$$[Q(\mathbf{k}), H] = k^{\nu} J_{\nu}(\mathbf{k}) = k^{\nu} J_{\nu}(\mathbf{k}) \otimes 1_l. \quad (71)$$

This reads for ($\nu = 0, 1, 2, 3; r = 1, 2, 3$)

$$k^{\nu} J_{\nu}(\mathbf{k}) \otimes 1_l = -k_0 Q(\mathbf{k}) \otimes 1_l + [Q(\mathbf{k}), H]. \quad (72)$$

From the Hamiltonian H of Eq. (58) we get by the commutation relations of the photon creation, respectively annihilation, operators the equation

$$\begin{aligned} [L(\mathbf{k}), H] &= k_0 L(\mathbf{k}) + k^{\nu} J_{\nu}(\mathbf{k}) \\ &= 1_k \otimes k_0 L(\mathbf{k}) + k^{\nu} J_{\nu}(\mathbf{k}) \otimes 1_l \\ &= k_0 \{L(\mathbf{k}) - Q(\mathbf{k})\} + [Q(\mathbf{k}), H]. \end{aligned} \quad (73)$$

On any sector \mathcal{S}^q , $L(\mathbf{k}) - Q(\mathbf{k})$ is the zero operator according to the definition of \mathcal{S}^q , so that we have $[L(\mathbf{k}), H] = [Q(\mathbf{k}), H]$ on any \mathcal{S}^q . Iteration of this equation yields

$[[L(\mathbf{k}), H], H] = [[Q(\mathbf{k}), H], H]$ on \mathcal{S} and so on. Because of $U(t) = \exp(-iHt)$ and the formal series $\exp(iHt) \times L(\mathbf{k}) \exp(-iHt) = L(\mathbf{k}) + it[L(\mathbf{k}), H] + \dots$ [similarly with respect to $Q(\mathbf{k})$] we get on \mathcal{S} ,

$$L(\mathbf{k}, t) := U^+(t)L(\mathbf{k})U(t) = U^+(t)Q(\mathbf{k})U(t) := Q(\mathbf{k}, t) \quad (74)$$

Multiplying this by $U(t)$ from the left we get Eq. (70).

Thus we can be sure that from the equation $L(\mathbf{k})(a \otimes \alpha) = q(\mathbf{k})(a \otimes \alpha) = Q(\mathbf{k})(a \otimes \alpha)$ the equation $L(\mathbf{k})(a(t) \otimes \alpha(t)) = Q(\mathbf{k})(a(t) \otimes \alpha(t))$ follows at least once for each element $a(t) \otimes \alpha(t) := U(t)(a \otimes \alpha)$ of the Schrödinger curve drawn through the element $a \otimes \alpha$. The question remains whether $a(t) \otimes \alpha(t)$ lies in a sector $\mathcal{S}^{q[t]}$ every time. We take $a(t) \otimes \alpha(t) = U(t)(a \otimes \alpha)$ and use the operators $L(\mathbf{k}) = 1_k \otimes L(\mathbf{k})$ and $Q(\mathbf{k}) = Q(\mathbf{k}) \otimes 1_l$. It follows that

$$\begin{aligned} [1_k \otimes L(\mathbf{k})][a(t) \otimes \alpha(t)] &= [Q(\mathbf{k}) \otimes 1_l][a(t) \otimes \alpha(t)], \text{ i.e.,} \\ a(t) \otimes [L(\mathbf{k})\alpha(t)] &= [Q(\mathbf{k})a(t)] \otimes \alpha(t). \end{aligned} \quad (75)$$

Such an equation can only be true if we have, with a function $q[t](\mathbf{k})$,

$$L(\mathbf{k})\alpha(t) = q[t](\mathbf{k})\alpha(t) \text{ and } Q(\mathbf{k})a(t) = q[t](\mathbf{k})a(t). \quad (76)$$

Because Ω^* has so been chosen that it contains $q[t]$ and because the definition of any sector \mathcal{S}^q accordingly has been generalized, these equations indeed define the sector $\mathcal{S}^{q[t]}$. ■

We finish this section with some remarks about approximation theories for QED on the basis of our state space $\mathcal{S}^* = \int \mathcal{S}^q (q \in \Omega^*)$. Any approximation might consist in a suitably chosen subset $\Omega^{\text{red}} \subset \Omega^*$ (in this sense, our index set Ω also defines such an approximation theory). Instead of the full state space \mathcal{S}^* we start with a reduced state space at $t = 0$

$$\mathcal{S}^{\text{red}}(t = 0) := \bigoplus_{q \in \Omega^{\text{red}}} \mathcal{S}^q.$$

The consequence is that with time running the state space $\mathcal{S}^{\text{red}}(t) := U(t)\mathcal{S}^{\text{red}}(t = 0)$ of the approximation theory will move through \mathcal{S}^* . In this way one gets a wandering state space, similar to the theory of pseudointeraction between M and any prescribed, classical, four-current (Sec. 3, example d). There cannot be any contradictions with respect to Sec. 3, however, because according to Definition 3 any observable

will always map this state space $\mathcal{S}^{\text{red}}(t)$ into itself at any moment $t \geq 0$.

6. THE NUMBER OPERATOR OF PHOTONS AS AN OBSERVABLE

To apply our discussions we are looking for the number operator of photons. In the literature it is usually given as

$$n(\mathbf{k}): = a_{\nu}^{\dagger}(\mathbf{k})a^{\nu}(\mathbf{k}). \quad (77)$$

In Sec. 3, example c, however, we have already seen that $n(\mathbf{k})$ is an observable only if the set $\mathcal{D}_{n(\mathbf{k})}$ consists of only the zero function; this means $n(\mathbf{k})$ is restricted to the space \mathcal{L}^0 . One may look upon this fact as a formalism, but the following lemma will show that $n(\mathbf{k})$ on Lorentz spaces \mathcal{L}^q with $q \neq 0$ cannot have the physical meaning of a photon number.

Lemma 6a:

(a) In any Lorentz space \mathcal{L}^q with $q \neq 0$ there are elements α such that $\langle \alpha | n(\mathbf{k}) \alpha \rangle < 0$. This contradicts the assumption that the term $d^3k \langle \alpha | n(\mathbf{k}) \alpha \rangle$ could be a photon number.

(b) Furthermore, the expectation values of $n(\mathbf{k})$ on any space \mathcal{L}^q with $q \neq 0$ are not generally gauge invariant, i.e., the value $\langle \alpha | n(\mathbf{k}) \alpha \rangle$ of any element $0 \neq \alpha \in \mathcal{L}^q$ changes when α is replaced by another element $\alpha' \in \mathcal{L}^q$ which is equivalent to α modulo \mathcal{N}^q .

Proof:

(a) The function $q = q(\mathbf{k}) \neq 0$ from \mathcal{D} may be given. We define the four-vector function $\varphi_{\mu}(\mathbf{k})$ by $\varphi^0(\mathbf{k}) := q(\mathbf{k})/k_0$ and $\varphi_r(\mathbf{k}) := 0$ for $r = 1, 2, 3$. Then we have $\alpha := \mathcal{W}\{\varphi\}\omega_0$ is an element from \mathcal{L}^q (Appendix). With respect to this element α we compute the expectation value

$$\begin{aligned} \langle \alpha | n(\mathbf{k}) \alpha \rangle &= \langle \omega_0 | \mathcal{W}\{-\varphi\} a_{\nu}^{\dagger}(\mathbf{k}) \mathcal{W}\{\varphi\} \mathcal{W}\{-\varphi\} a^{\nu}(\mathbf{k}) \mathcal{W}\{\varphi\} \omega_0 \rangle \\ &= \langle \omega_0 | [a_{\nu}^{\dagger}(\mathbf{k}) + \varphi^{\dagger}(\mathbf{k})] [a^{\nu}(\mathbf{k}) + \varphi^{\nu}(\mathbf{k})] \omega_0 \rangle \\ &= \varphi^{\dagger}(\mathbf{k}) \varphi^{\nu}(\mathbf{k}) = -q^*(\mathbf{k})q(\mathbf{k})/k_0^2 < 0 \quad \text{for some } \mathbf{k} \in R^3. \end{aligned}$$

(b) We now choose any element $\alpha \in \mathcal{L}^q$ with $\langle \alpha | \alpha \rangle \neq 0$. According to Sec. 2 we apply the gauge operator $\mathcal{W}\{kg\}$ to any scalar function $g = g(\mathbf{k})$. The element $\alpha' := \mathcal{W}\{kg\}\alpha \in \mathcal{L}^q$ is then equivalent to α modulo \mathcal{N}^q . A straightforward calculation gives

$$\begin{aligned} \langle \alpha' | n(\mathbf{k}) \alpha' \rangle &= \langle \alpha | [a_{\nu}^{\dagger}(\mathbf{k}) + k_{\nu}g^*(\mathbf{k})] [a^{\nu}(\mathbf{k}) + k^{\nu}g(\mathbf{k})] \alpha \rangle \\ &= \langle \alpha | n(\mathbf{k}) \alpha \rangle + [g^*(\mathbf{k})q(\mathbf{k}) + g(\mathbf{k})q^*(\mathbf{k})] \langle \alpha | \alpha \rangle. \end{aligned}$$

Therefore, we can arrange by a proper selection of \mathbf{k} and g that in all cases $q \neq 0$ we have $\langle \alpha' | n(\mathbf{k}) \alpha' \rangle \neq \langle \alpha | n(\mathbf{k}) \alpha \rangle$. ■

Thus we have shown that at least in the case $q \neq 0$ the correct operator $N(\mathbf{k})$ of the number of photons remains to be looked for. As $N(\mathbf{k})$ is to be defined on \mathcal{S}_M , i.e., $N(\mathbf{k}) = 1_k \otimes N(\mathbf{k})$, we introduce as an heuristic means the pseudointeraction between M and any prescribed, classical, four-current $j_{\nu}(x)$ with the following intention: The prescribed classical four-current $j_{\nu}(x)$ defines the time evolution operator $U(x_0)$ by the Hamiltonian $H(x_0) = H(t)$ of Eq. (41). Then we have $\mathcal{L}^{q(x_0)} = U(x_0)\mathcal{L}^{q(0)}$, where $q[x_0](\mathbf{k}) = j_0(\mathbf{k}, x_0)$, so that the element $\alpha(x_0) := U(x_0)\alpha$ of the Schrödinger curve drawn through α is always in the correct Lorentz space $\mathcal{L}^{q(x_0)}$ if the initial element α was selected in $\mathcal{L}^{q(0)}$. The expectation values of the field tensor $F_{\mu\nu}(x)$ of

Eq. (12) with respect to $\alpha(x_0)$ will then give the expectation values $\mathbf{E}(x)$ and $\mathbf{B}(x)$ of the electric and magnetic field strengths. The classical field energy must be connected with the expectation value of the operator $N(\mathbf{k})$ to be looked for in the following way:

$$\int d^3k |\mathbf{k}| \langle \alpha(x_0) | N(\mathbf{k}) \alpha(x_0) \rangle = \frac{1}{2} \int d^3x [\mathbf{E}^2(x) + \mathbf{B}^2(x)]. \quad (78)$$

A tightening of this postulate (78) will automatically lead us to the operator $N(\mathbf{k})$ we have been looking for.

Lemma 6b: We take the Schrödinger curve $\alpha(x_0) = U(x_0)\alpha \in \mathcal{L}^{q(x_0)}$ as it has just been introduced. With the help of the expectation value

$$g_{\mu}(\mathbf{k}, x_0) := \langle \alpha(x_0) | a_{\mu}(\mathbf{k}) \alpha(x_0) \rangle / \langle \alpha | \alpha \rangle \quad (79)$$

we can then write the classical field energy, which is connected with the prescribed $j_{\mu}(x)$, where $q[x_0](\mathbf{k}) = j_0(\mathbf{k}, x_0)$, as follows:

$$\begin{aligned} \frac{1}{2} \int d^3x [\mathbf{E}^2(x) + \mathbf{B}^2(x)] &= \int d^3k k_0 \left\{ g_{\mu}^* g^{\mu} - \frac{1}{k_0} [q g_0^* + q^* g_0] \right\}. \quad (80) \end{aligned}$$

Proof: From $F_{\mu\nu}(x)$ and the definition of g_{μ} one gets first the Fourier components in the form

$$\mathbf{E}(\mathbf{k}, x_0) = i(2k_0)^{-1/2} \{ \mathbf{k} [g_0(\mathbf{k}, x_0) + g_0^*(-\mathbf{k}, x_0)] - k_0 [g(\mathbf{k}, x_0) - g^*(-\mathbf{k}, x_0)] \}, \quad (81)$$

$$\mathbf{B}(\mathbf{k}, x_0) = i(2k_0)^{-1/2} \{ \mathbf{k} \times g(\mathbf{k}, x_0) + \mathbf{k} \times g^*(-\mathbf{k}, x_0) \}, \quad (82)$$

where we have combined the components g_1, g_2, g_3 with the "vector" \mathbf{g} . From this we get the assertion by the equation

$$\begin{aligned} \frac{1}{2} \int d^3x [\mathbf{E}^2(x) + \mathbf{B}^2(x)] &= \frac{1}{2} \int d^3k [\mathbf{E}(\mathbf{k}, x_0) \mathbf{E}^*(\mathbf{k}, x_0) \\ &\quad + \mathbf{B}(\mathbf{k}, x_0) \mathbf{B}^*(\mathbf{k}, x_0)], \quad (83) \end{aligned}$$

if one also considers the premise with reference to the Lorentz spaces in the form

$$k^{\nu} g_{\nu}(\mathbf{k}, x_0) = q[x_0](\mathbf{k}). \quad \blacksquare \quad (84)$$

Because of Lemma 6b we demand for the expectation value of $N(\mathbf{k})$ the equation

$$\begin{aligned} \langle \alpha(x_0) | N(\mathbf{k}) \alpha(x_0) \rangle &:= g_{\nu}^*(\mathbf{k}, x_0) g^{\nu}(\mathbf{k}, x_0) - (1/k_0) \\ &\quad \times \{ g_0^*(\mathbf{k}, x_0) q(\mathbf{k}, x_0) + g_0(\mathbf{k}, x_0) q^*(\mathbf{k}, x_0) \}. \quad (85) \end{aligned}$$

As $N(\mathbf{k})$ is here applied on $\alpha(x_0) \in \mathcal{L}^{q(x_0)}$ and as $N(\mathbf{k})$ should go over to the operator $n(\mathbf{k})$ in the case $q = 0$, this leads us necessarily to the definition

Definition 4: We introduce as the number operator of photons:

$$N(\mathbf{k}) := n(\mathbf{k}) - (1/k_0) \{ a_0^{\dagger}(\mathbf{k}) L(\mathbf{k}) + L^{\dagger}(\mathbf{k}) a_0(\mathbf{k}) \}. \quad (86)$$

We get some other forms of the operator $N(\mathbf{k})$ when we introduce the usual transversal photons (Appendix) with the number operator

$$n^{tr}(\mathbf{k}) := a^{+(1)}(\mathbf{k}) a^{(1)}(\mathbf{k}) + a^{+(2)}(\mathbf{k}) a^{(2)}(\mathbf{k}) \quad (87)$$

and "bad ghosts" in the sense of Ref. 14 with the destruction operators

$$a_b(\mathbf{k}) := \frac{1}{\sqrt{2}} \{ a^{(3)}(\mathbf{k}) - a^{(0)}(\mathbf{k}) \} = \frac{1}{\sqrt{2} \cdot k_0} \cdot L(\mathbf{k}) \quad (88)$$

and the number operator

$$n_b(\mathbf{k}) := a_b^+(\mathbf{k})a_b(\mathbf{k}) = \frac{1}{2k_0^2} L^+(\mathbf{k})L(\mathbf{k}). \quad (89)$$

Then we get $N(\mathbf{k}) = n^{\text{tr}}(\mathbf{k}) + 2n_b(\mathbf{k})$ and

$$N(\mathbf{k}) = n^{\text{tr}}(\mathbf{k}) + \frac{1}{k_0^2} L^+(\mathbf{k})L(\mathbf{k}). \quad (90)$$

All these equations show that $N(\mathbf{k})$ will coincide with $n(\mathbf{k})$ only on the space \mathcal{L}^0 , i.e., in the absence of any electrical-charge distribution. Equation (90) shows especially that $n^{\text{tr}}(\mathbf{k})$ gives the density of the "radiating" photons, whereas the term $L^+(\mathbf{k})L(\mathbf{k})/k_0^2$ gives the density of the photons bounded by the given charge distribution. Only the sum of both parts is an observable in our sense!

The new operator $N(\mathbf{k})$ now possesses all properties which were missing in Lemma 6a with $n(\mathbf{k})$.

Theorem 8: The operator $N(\mathbf{k})$ is an observable with $\mathcal{Q}_{N(\mathbf{k})} = \mathcal{Q}$ (especially gauge-invariant according to Lemma 3e) and non-negative.

Proof: By straightforward computation one finds $[L(\mathbf{k}), N(\mathbf{k}')] = 0$. Thus $N(\mathbf{k})$ maps every Lorentz space into itself. The property of $N(\mathbf{k})$ of being non-negative follows immediately from Eq. (90), for the number operator of the transversal photons is non-negative. ■

We return to the interacting system $J + M$ (Sec. 5) with the state space $\mathcal{S} = \oplus \mathcal{S}^r$. According to Eq. (58) the Hamiltonian is given by

$$H = H_0^J + \int d^3k k_0 n(\mathbf{k}) + \int d^3k \{J^\nu(\mathbf{k})a_{\nu}^+(\mathbf{k}) + J_{\nu}^+(\mathbf{k})a^{\nu}(\mathbf{k})\}, \quad (91)$$

where H_0^J is the Hamiltonian of the free electrons and positrons (more generally, the free quanta of J). Now we rewrite H in the form

$$H = H_0^J + \int d^3k k_0 \left[n(\mathbf{k}) - \frac{1}{k_0} [a_0^+(\mathbf{k})Q(\mathbf{k}) + Q^+(\mathbf{k})a_0(\mathbf{k})] \right] + \int d^3x J_r(\mathbf{x})A^r(\mathbf{x}), \quad (r = 1, 2, 3). \quad (92)$$

The expectation values of $Q(\mathbf{k})$ and $L(\mathbf{k})$, and $Q^+(\mathbf{k})$ and $L^+(\mathbf{k})$, respectively, coincide on \mathcal{S} , so that we can write for H in a short form:

$$H = \int d^3k k_0 N(\mathbf{k}) + \int d^3x J_r(\mathbf{x})A^r(\mathbf{x}) + H_0^J. \quad (93)$$

The Gupta-Bleuler-metrics have been eliminated formally from this remarkable form of H , for $r = 1, 2, 3$. The operator $J_0 = Q$ will no longer appear explicitly. Instead, the operator $N(\mathbf{k})$ takes the role of J_0 , as $N(\mathbf{k})$ introduces the eigenvalue function q of $J_0 = Q$ on any sector \mathcal{S}^q into the calculation.

APPENDIX

The following section summarizes physical and mathematical prerequisites of this paper. It will also serve to give our terminology.

A. The space \mathcal{S}_M

The physical principle for the definition of the state space \mathcal{S}_M of the quantized Maxwell field M is provided by the canonical commutation relations

$$[A_\nu(\mathbf{x}), H^\nu(\mathbf{x}')] = i\delta_\mu^\nu \delta(\mathbf{x} - \mathbf{x}') \quad (A1)$$

of the dynamical variables of M which are to be represented on \mathcal{S}_M . If one introduces first the operators $a_\mu(\mathbf{k}), a_\nu^+(\mathbf{k})$ in a formal way by the definitions

$$A_\mu(\mathbf{x}) = (2\pi)^{-\frac{3}{2}} \int \frac{d^3k}{\sqrt{2|\mathbf{k}|}} \{e^{i\mathbf{k}\cdot\mathbf{x}} a_\mu(\mathbf{k}) + e^{-i\mathbf{k}\cdot\mathbf{x}} a_\mu^+(\mathbf{k})\}$$

and

$$H^\nu(\mathbf{x}') = -i(2\pi)^{-\frac{3}{2}} \int d^3l \sqrt{\frac{|\mathbf{l}|}{2}} \{e^{i\mathbf{l}\cdot\mathbf{x}'} a^{\nu}(\mathbf{l}) - e^{-i\mathbf{l}\cdot\mathbf{x}'} a^{+\nu}(\mathbf{l})\}, \quad (A2)$$

then Eq. (A1) is satisfied if one postulates

$$[a_\mu(\mathbf{k}), a_\nu^+(\mathbf{l})] = g_{\mu\nu} \delta(\mathbf{k} - \mathbf{l})$$

and

$$[a_\mu(\mathbf{k}), a_\nu(\mathbf{l})] = 0 = [a_\mu^+(\mathbf{k}), a_\nu^+(\mathbf{l})], \quad (A3)$$

where the Minkowski metric $g_{\mu\mu} = (-1, +1, +1, +1)$ and natural units have been used. To represent now these formal operators on a space \mathcal{S}_M we denote by K^0 the empty set and by K^n the set $(\mathbf{k}_1, \mu_1; \dots; \mathbf{k}_n, \mu_n)$ of $n = 1, 2, \dots$ pairs of real variables \mathbf{k}_ν, μ_ν where any \mathbf{k}_ν varies continuously over the \mathbb{R}^3 and any μ_ν assumes the values $0, 1, 2, 3$. Define further the symbol $\int dK^n$ by

$$\int dK^n \dots := \int d^3k_1 \dots d^3k_n \cdot \sum_{\mu_1=0}^3 \dots \sum_{\mu_n=0}^3, \quad (A4)$$

where $\int d^3k$ denotes the elementary Lebesgue integral over the \mathbb{R}^3 . Consider also the Fock space \mathcal{F} of all sequences

$$\alpha := \{\alpha_0(K^0), \alpha_1(K^1), \alpha_2(K^2), \dots\}, \quad (A5)$$

with the following properties: The n th component $\alpha_n = \alpha_n(K^n)$ of α is a complex number for $n = 0$, and for $n = 1, 2, \dots$ it is a complex-valued function, symmetric in the pairs $(\mathbf{k}_\nu, \mu_\nu)$ and defined in such a way that $\int dK^n |\alpha_n(K^n)|^2$ exists. The Hilbert scalar product of \mathcal{F} is given by

$$[\alpha, \beta] := \alpha_0^* \beta_0 + \sum_{n=1}^{\infty} \int dK^n \alpha_n^*(K^n) \beta_n(K^n) \quad (A6)$$

and exists for any $\alpha, \beta \in \mathcal{F}$ if $\alpha \in \mathcal{F}$ means $\|\alpha\| := \sqrt{[\alpha, \alpha]} < \infty$, as usual. We assume that \mathcal{F} has been completed already. The norm $\|\dots\|$ on \mathcal{F} defines in particular a complete Banach space \mathcal{B} . On this Banach space we introduce the scalar product (Gupta-Bleuler form)

$$\langle \alpha | \beta \rangle := \alpha_0^* \beta_0 + \sum_{n=1}^{\infty} \int dK^n \alpha_n^*(K^n) g_{\mu_1 \mu_1} \dots g_{\mu_n \mu_n} \beta_n(K^n), \quad (A7)$$

which is used exclusively in this paper. \mathcal{S}_M is then defined as the pair

$$\mathcal{S}_M := (\mathcal{B}, \langle \alpha | \beta \rangle). \quad (A8)$$

In all subspaces of \mathcal{S}_M which are discussed here this scalar product $\langle | \rangle$ is always used automatically. In close analogy to Fock spaces¹⁵ we define destruction and creation opera-

tors $a_\mu(\mathbf{k}), a_\mu^+(\mathbf{k})$ by

$$(a_\mu(\mathbf{k})\alpha)_n(K^n) := (n+1)^{1/2}\alpha_{n+1}(\mathbf{k}, \mu; K^n), \quad n=0,1,\dots \text{ and}$$

$$(a_\mu^+(\mathbf{k})\alpha)_n(K^n) := \begin{cases} 0, & n=0 \\ \frac{1}{\sqrt{n}} \sum_{\nu=1}^{\infty} g_{\mu\nu} \delta(\mathbf{k}-\mathbf{k}_\nu) \alpha_{n-1}(K^n \setminus (\mathbf{k}_\nu, \mu_\nu)), & n=1,2,\dots \end{cases} \quad (\text{A9})$$

with

$$(\mathbf{k}, \mu; K^n) := (\mathbf{k}, \mu; \mathbf{k}_1, \mu_1; \dots; \mathbf{k}_n, \mu_n)$$

and

$$K^n \setminus (\mathbf{k}_\nu, \mu_\nu) := (\mathbf{k}_1, \mu_1; \dots; \mathbf{k}_{\nu-1}, \mu_{\nu-1}; \mathbf{k}_{\nu+1}, \mu_{\nu+1}; \dots; \mathbf{k}_n, \mu_n).$$

These operators satisfy Eq. (71) and are formal adjoints of each other relative to $\langle | \rangle$.

As the "bare vacuum" ω_0 we introduce the element $\omega_0 := (1, 0, 0, \dots) \in \mathcal{S}_M$. With its help we can also write the components $\alpha_n(K^n)$ of any element $\alpha \in \mathcal{S}_M$ in the important form

$$\alpha_n(K^n) = (n!)^{-1/2} \langle \omega_0 | a_{\mu_1}(\mathbf{k}_1) \dots a_{\mu_n}(\mathbf{k}_n) \alpha \rangle, \quad n=1,2,\dots,$$

$$\alpha_0 = \langle \omega_0 | \alpha \rangle. \quad (\text{A10})$$

For many purposes it is useful to introduce modified operators $a^{(\lambda)}(\mathbf{k}), a^{+(\lambda)}(\mathbf{k})$ instead of the operators $a_\mu(\mathbf{k}), a_\mu^+(\mathbf{k})$ by the definitions

$$a_\mu(\mathbf{k}) = \sum_{\lambda=0}^3 e_\mu^{(\lambda)}(\mathbf{k}) a^{(\lambda)}(\mathbf{k}), \quad a_\mu^+(\mathbf{k}) = \sum_{\lambda=0}^3 e_\mu^{(\lambda)}(\mathbf{k}) a^{+(\lambda)}(\mathbf{k}), \quad (\text{A11})$$

where the $e_\mu^{(\lambda)}(\mathbf{k}), \lambda=0,1,2,3$, are real "polarization" four-vectors with the properties (for $r=1,2,3$)

$$\begin{aligned} e_0^{(0)}(\mathbf{k}) &= 1; & e_r^{(0)}(\mathbf{k}) &= 0; & e_r^{(1)}(\mathbf{k}) e^{(1)r}(\mathbf{k}) &= 1 = e_r^{(2)}(\mathbf{k}) e^{(2)r}(\mathbf{k}); \\ e_0^{(1)}(\mathbf{k}) &= 0 = e_0^{(2)}(\mathbf{k}); & k^r e_r^{(1)}(\mathbf{k}) &= 0 = k^r e_r^{(2)}(\mathbf{k}); \\ e_0^{(3)}(\mathbf{k}) &= 0; & e_r^{(3)}(\mathbf{k}) &= k_r / |\mathbf{k}|; & e_r^{(1)}(\mathbf{k}) e^{(2)r}(\mathbf{k}) &= 0; \end{aligned} \quad (\text{A12})$$

repeated indices r indicate the sum over $r=1,2,3$. These vectors satisfy the relations

$$e_\mu^{(\lambda)}(\mathbf{k}) e^{(\lambda)\mu}(\mathbf{k}) = g^{\lambda\lambda}. \quad (\text{A13})$$

and induce the concept of destruction operators for scalar ($\lambda=0$), transversal ($\lambda=1,2$) and longitudinal ($\lambda=3$) photons with reference to the new operators $a^{(\lambda)}(\mathbf{k})$. The commutation relations are now

$$[a^{(\lambda)}(\mathbf{k}), a^{+(\mu)}(\mathbf{l})] = g^{\lambda\mu} \delta(\mathbf{k}-\mathbf{l}). \quad (\text{A14})$$

The Lorentz operator $L(\mathbf{k}) = k^\nu a_\nu(\mathbf{k})$ takes the special form

$$L(\mathbf{k}) = k_0 \{ a^{(3)}(\mathbf{k}) - a^{(0)}(\mathbf{k}) \}, \quad k_0 := |\mathbf{k}|. \quad (\text{A15})$$

B. Weyl operators on \mathcal{S}_M

Let $f_\mu(\mathbf{k})$ be any square-integrable four-vector function. Then we expect that the Weyl operator

$$W\{f\} := \exp \int d^3k \{ f^\mu(\mathbf{k}) a_\mu^+(\mathbf{k}) - f_\mu^*(\mathbf{k}) a^\mu(\mathbf{k}) \} \quad (\text{A16})$$

can be defined on \mathcal{S}_M by its power series. We bypass the question of the exact domain of $W\{f\}$ in \mathcal{S}_M (cf. however the appendix of Ref. 1). The following relations are the main

tools of our conclusions:

$$W^+\{f\} = W\{-f\},$$

$$W\{f+g\} = W\{f\} W\{g\} \exp \left\{ -i \text{Im} \int d^3k f^\mu(\mathbf{k}) g_\mu^*(\mathbf{k}) \right\}, \quad (\text{A17})$$

$$\begin{aligned} W\{f\} &= \exp \left\{ -\frac{1}{2} \int d^3k f_\mu^*(\mathbf{k}) f^\mu(\mathbf{k}) \right\} \\ &\cdot \exp \left\{ \int d^3k f^\mu(\mathbf{k}) a_\mu^+(\mathbf{k}) \right\} \\ &\cdot \exp \left\{ -\int d^3k f_\mu^*(\mathbf{k}) a^\mu(\mathbf{k}) \right\} \end{aligned}$$

Their consequence is, in particular

$$\langle W\{f\} \alpha | W\{f\} \alpha \rangle = \langle \alpha | \alpha \rangle. \quad (\text{A18})$$

Coherent states^{16,17} of \mathcal{S}_M are elements of the form

$$\begin{aligned} W\{f\} \omega_0 &= \exp \left\{ -\frac{1}{2} \int d^3k f_\mu^*(\mathbf{k}) f^\mu(\mathbf{k}) \right\} \\ &\cdot \exp \left\{ \int d^3k f^\mu(\mathbf{k}) a_\mu^+(\mathbf{k}) \right\} \omega_0. \end{aligned} \quad (\text{A19})$$

They satisfy

$$a_\mu(\mathbf{k}) W\{f\} \omega_0 = f_\mu(\mathbf{k}) W\{f\} \omega_0, \quad (\text{A20})$$

so that any coherent state is an element of the Lorentz space \mathcal{L}^q where q is given by the equation

$$q = q(\mathbf{k}) = k^\mu f_\mu(\mathbf{k}). \quad (\text{A21})$$

The Weyl operators define isometric mappings¹ between the Lorentz spaces. Such a mapping $\mathcal{L}^p \rightarrow \mathcal{L}^q$ is given in the following way: Let $p_\mu(\mathbf{k}), q_\mu(\mathbf{k})$ be functions such that $k^\mu p_\mu(\mathbf{k}) = p(\mathbf{k})$ and $k^\mu q_\mu(\mathbf{k}) = q(\mathbf{k})$. Then the Weyl operator $W\{q-p\}$ gives such a mapping.

If one takes $g_\mu(\mathbf{k}) = k_\mu g(\mathbf{k})$ for any scalar function $g(\mathbf{k})$ there arises the Weyl operator $W\{kg\}$ which maps any Lorentz space \mathcal{L}^q into itself because $k^2 = k^\nu k_\nu = 0$. The Weyl operators occur, by the way, in a quite natural way as time-evolution operators $U(t)$, if one introduces¹ on \mathcal{S}_M the pseudointeraction between M and any prescribed classical four-current.

¹W. Gessner and V. Ernst, J. Math. Phys. **21**, 93 (1980).

²A. Ashtekar and R. Geroch, Rep. Prog. Phys. **37**, 1211 (1974).

³M. Born, Z. Phys. **58**, 803 (1926).

⁴S. Gupta, Proc. Phys. Soc. (London) A **63**, 681 (1950); A **64**, 859 (1951).

⁵K. Bleuler, Helv. Phys. Acta **23**, 567 (1950).

⁶J. Bognár, *Indefinite Inner Product Spaces* (Springer, Berlin, 1974).

⁷F. Strocchi and A. S. Wightman, J. Math. Phys. **15**, 2198 (1974).

⁸K. O. Friedrichs, *Mathematical Aspects of the Quantum Theory of Fields* (Interscience, New York, 1953).

⁹J. M. Jauch, *Foundations of Quantum Mechanics* (Addison-Wesley, Reading, Mass., 1968).

¹⁰F. Bloch and A. Nordsieck, Phys. Rev. **52**, 54 (1937).

¹¹J. R. Klauder and J. McKenna, J. Math. Phys. **6**, 68 (1965).

¹²J. R. Klauder, J. McKenna, and E. J. Woods, J. Math. Phys. **7**, 822 (1966).

¹³T. W. B. Kibble, J. Math. Phys. **9**, 315 (1968).

¹⁴H. P. Dürr and E. Rudolph, Nuovo Cimento **62 A**, 411 (1969).

¹⁵S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Harper & Row, New York, 1959).

¹⁶R. J. Glauber, Phys. Rev. **84**, 395 (1951); Phys. Rev. **130**, 2529 (1963);

- Phys. Rev. **131**, 2766 (1963).
- ¹⁷J. Gomatam, Phys. Rev. D **6**, 1292 (1971).
- ¹⁸J. M. Cook, Trans. Am. Math. Soc. **74**, 222 (1953); J. Math. Phys. **2**, 33 (1961).
- ¹⁹J. M. Jauch and F. Rohrlich, *The Theory of Photons and Electrons* (Addison-Wesley, Cambridge, Mass., 1955).
- ²⁰K. L. Nagy, *State Vector Spaces with Indefinite Metric in Quantum Field Theory* (Akadémiai Kiadó, Budapest, 1966).
- ²¹W. Heitler, *The Quantum Theory of Radiation* 3rd ed. (Clarendon, Oxford, 1954).
- ²²J. von Neumann, *Mathematische Grundlagen der Quantenmechanik* (Springer, Berlin, 1932).
- ²³N. J. Achieser and I. M. Glasman, *Theorie der linearen Operatoren im Hilbertraum*, 5th ed. (Akademie, Berlin, 1968).

Phase-space approach to relativistic quantum mechanics. III. Quantization, relativity, localization and gauge freedom

Gerald Kaiser

Mathematics Department, University of Lowell, Lowell, Massachusetts 01854

(Received 23 September 1980; accepted for publication 26 November 1980)

We examine the relationship between the mathematical structures of classical mechanics, quantum mechanics, and special relativity, with a view toward building a consistent framework for all three. The usual idea of “canonical quantization,” with its emphasis on the transition from functions over classical phase space to operators, appears to be inconsistent with relativistic covariance. On the other hand, the spectral condition of relativistic quantum mechanics, which is merely a covariant statement that the energy is nonnegative, naturally leads to the construction of a covariant extension of classical phase space. By giving up the idea of sharp “localizability” in space—known to be at odds with covariance—and adopting instead a notion of “soft” localizability in phase space, a consistent theory of relativistic quantum mechanics is seen to emerge whose structure naturally incorporates the classical symplectic geometry as well. Furthermore, the new theory deals directly and covariantly with extended particles rather than point particles and is free of the various inconsistencies known to plague the usual theory. The new notion of “microlocality” in phase space leads to a new form of gauge freedom which is similar to the usual one but simpler and more powerful, using methods of complex analysis. The phase-space version of Yang–Mills theory is worked out.

PACS numbers: 03.65.Ca, 11.10.Np

1. INTRODUCTION

In recent years, the relationship between classical and quantum mechanics has received much renewed attention. Most efforts have centered around the problem of “quantization” (for a sample of the literature see Refs. 1–3 and the references therein) and its dual, the “classical limit” problem (see Refs. 3,4 and the references therein). One major point of view which has emerged in connection with the quantization problem is that there is no entirely satisfactory, exact solution (see, for example, Ref. 4, p. 249), although its study has led to some deep and interesting results such as the orbit method in Lie group representation theory and a better understanding of the semiclassical approximation. In the study of the classical limit problem, considerable progress has been made—generally by using phase-space formulations of quantum mechanics such as the various “coherent-state” representations. Thus it appears that while the classical limit problem lends itself to rigorous mathematical analysis, “quantization” remains an art, relying on one set or another of mysterious but sometimes useful “prescriptions.” This should not be so surprising, since quantum mechanics is the more fundamental description of Nature; hence the quantization problem appears analogous to guessing a relativistic theory from its nonrelativistic limit.

Yet, the dream of a perfect “quantization scheme” dies hard. The reasons for this are not difficult to discover:

In recent years, classical mechanics has been generalized and reduced, in principle, to a branch of global symplectic geometry.^{5–7} By contrast, the accepted, standard formalism of quantum mechanics is far from being definitive; it depends on the existence of *Cartesian* “canonical coordinates” which are to be suddenly and mysteriously trans-

formed into operators on a Hilbert Space, with Poisson brackets going to commutators. Aside from its conceptual deficiency, this process is both mathematically and physically unsatisfactory. It is not “intrinsic” (and hence fails, for example, when the classical phase space is curved or multiply connected), and even in the standard theory it is riddled with ambiguities (the operator-ordering problem). Anyone who believes in the fundamental unity of Nature will be led to search for a quantum-theoretic counterpart of global, symplectic classical mechanics. This is the basic motivation behind the “geometric quantization” program.^{1–3}

My aim in this paper is as follows.

(1) I will show (in Sec. 2) that the “canonical” structure of the position and momentum operators in nonrelativistic quantum mechanics is a *fluke*; it is a degenerate form of relativistic invariance which, strictly speaking, breaks down in the relativistic theory. This insight into the canonical commutation relations (CCR) both explains their mystery and casts into doubt their value as a cornerstone of quantum theory. (These remarks are restricted to ordinary quantum mechanics; we make no claims in this paper concerning the canonical commutation and anticommutation relations of quantum field theory.)

(2) Since one of the great attractions of the CCR has been their formal relation to the classical symplectic structure, their critique cannot be considered entirely satisfactory until a substitute for this relation is found. This will be done in Sec. 3, where, summarizing Parts I and II of this series, a natural formulation of *relativistic* quantum theory in terms of functions over a *covariant extension of classical phase space* is given. As a by-product, two related, long-standing inconsistencies of relativistic quantum mechanics (in its

standard, space-time formulation) are resolved, namely the problems of *localization* and *covariant probabilistic interpretation*. Rather than being *sharply* localizable in *space*, particles in the new formulation are at best permitted to be *softly* localizable in *phase space*. This concept, which we call *microlocality*, turns out to give a consistent and satisfactory theory.

(3) Since strict localizability is untenable, the idea of local gauge invariance must be reexamined. In Sec. 4 we develop a “microlocal” gauge theory in phase space. The resulting gauge fields are shown to be formally similar to, but conceptually simpler than, the usual fields of Yang–Mills type.

2. QUANTIZATION AND RELATIVITY

To see how the CCR originate in Relativity, consider the invariance group \mathcal{P}^1_+ (restricted Poincaré group) of Minkowskian space-time, whose Lie algebra \mathfrak{p} is spanned by the generators T_k ($k = 1, 2, 3$) of spatial translations, T_0 of time translations, J_k of rotations, and K_k of pure Lorentz transformations. The Lie brackets of \mathfrak{p} are given by

$$\begin{aligned} [J_i, J_j] &= J_k, & [J_i, K_j] &= K_k, \\ [T_0, K_r] &= T_r, & [J_i, T_j] &= T_k, \\ [K_i, K_j] &= -c^{-2}J_k, & [T_r, K_s] &= c^{-2}\delta_{rs}T_0, \end{aligned} \quad (1)$$

where c is the speed of light, (i, j, k) is a cyclic permutation of $(1, 2, 3)$, $r, s = 1, 2, 3$, and all unspecified brackets vanish. The physical dimensions of the generators are as follows: T_0 is a reciprocal time, T_k is a reciprocal length, J_k is dimensionless (reciprocal angle), and K_k is a reciprocal velocity.

Note that so far, nothing has been said about quantum mechanics. \mathcal{P}^1_+ merely describes the geometry of classical, relativistic space-time. We now make our first assumption of quantal nature:

(Q): *The formalism of a relativistic quantum theory is based on a unitary (though possibly reducible) representation of \mathcal{P}^1_+ (or its universal covering group).*

Hardly anyone will dispute this statement, which is standard material in relativistic quantum theory. Unlike what is ordinarily called “quantization”, it is physically clear and mathematically unambiguous. Yet, (Q) *implies and, at the same time, supersedes the mysterious CCR!* To see this consider the nonrelativistic limit $c \rightarrow \infty$ of \mathfrak{p} . Letting $c^{-2}T_0 = M$ in Eq. (1), the Lie algebra \mathfrak{p} “contracts” to

$$\begin{aligned} [J_i, J_j] &= J_k, & [J_i, K_j] &= K_k, \\ [M, K_r] &= 0, & [J_i, T_j] &= T_k, \\ [K_i, K_j] &= 0, & [T_r, K_s] &= \delta_{rs}M, \end{aligned} \quad (2)$$

with all other brackets vanishing. Let us call this Lie algebra \mathfrak{g}_1 , and the corresponding (simply connected) Lie group \mathcal{G}_1 . Note that (a) M is a central element of \mathfrak{g}_1 and (b) M, T_k , and K_k generate an invariant subgroup \mathcal{W} (with Lie algebra \mathfrak{w}) of \mathcal{G}_1 . The remaining generators J_k give the rotation group or its double cover $SU(2)$, so \mathcal{G}_1 is the semidirect product of $SU(2)$ with \mathcal{W} :

$$\begin{aligned} \mathcal{G}_1 &= SU(2) \otimes \mathcal{W}, \\ \mathcal{W} &= \mathcal{G}_1 / SU(2). \end{aligned} \quad (3)$$

Now suppose that the unitary representation of \mathcal{P}^1_+ in assumption (Q) is *irreducible*. [Assumption (Q) means we are dealing with a quantum system, possibly a quantum field theory; it is the additional assumption of irreducibility which makes this system “elementary”—roughly, a *particle*.

Hence, the concept of *position*, discussed below, is only now admissible.] Assuming that the formal limit $c \rightarrow \infty$ of Lie algebras induces a rigorous limit on the representation level (and this is indeed the case, as proved in Refs. 8, 9), the central element M of \mathfrak{g}_1 will be represented by a purely imaginary constant in that limit: $M = -(i/\hbar)m$, where m is real. (Planck’s constant \hbar has been inserted for dimensional reasons, so that m can be interpreted as a *mass*.) Assume $m > 0$, and let

$$P_k = i\hbar T_k, \quad Q_k = -(i\hbar/m)K_k. \quad (4)$$

Then P_k and Q_k are represented by Hermitian operators which satisfy the CCR. In fact, \mathcal{W} of Eq. (3) is isomorphic to the *Weyl–Heisenberg group*,¹⁰ which is usually regarded to be at the heart of quantization theory!

How does it happen that classical relativistic geometry, represented by \mathcal{P}^1_+ , when combined with a simple assumption (Q), yields the mysterious CCR? To understand this, we first note that although (2) was obtained from (1) by taking $c \rightarrow \infty$, we cannot obtain the CCR so simply without invoking *Relativity*. For had we begun with Newtonian spacetime, we would have the Galilean group \mathcal{G} instead of \mathcal{P}^1_+ . Since Galilean boosts commute with spatial translations (time being absolute), the brackets between the corresponding generators vanish, hence no CCR! In the case of \mathcal{G} , the CCR are a remnant of relativistic invariance where, due to the nonabsolute nature of simultaneity, spatial translations do not commute with pure Lorentz transformations. Thus, the uncertainty principle originates, in some sense, in “classical” Relativity theory! The groups \mathcal{G} and \mathcal{G}_1 are related through a process called “central extension”^{11,12} which is mathematically complicated and physically every bit as obscure as the (closely related) process of canonical quantization. Thus we see that our conceptually simple derivation of the CCR really does depend on Relativity.

How does all this affect our ideas about quantization? The group \mathcal{W} , whose representation theory serves as a paradigm for the nonrelativistic quantization schemes, appears at first sight to have no counterpart in the relativistic theory. To get some insight, let us reflect for a moment on the physical significance of \mathcal{W} . In its usual form (where the generators are Q_k, P_k , and the identity operator), the group-manifold of \mathcal{W} has local coordinates of momentum p_k (generated by Q_k), position q_k (generated by P_k), and phase angle ϕ (generated by the identity). Thus \mathcal{W} is usually thought of as the product of *momentum phase space* with *phase angle*. However, our discussion above suggests a different interpretation: K_k generates velocity coordinates v_k , T_k generates position coordinates q_k as before, and M produces a degenerate form of relativistic “time.” Thus for a proper transition to Relativity, \mathcal{W} ought to be reinterpreted as the product of *velocity phase space* with “time.” Such an object is sometimes called a *state space*.

We are now in a position to go to Relativity. The counterpart of \mathcal{W} must clearly be the product \mathcal{C} of space-time with the velocity hyperboloid. Although this seven-dimensional manifold is no longer a group, it is directly related to \mathcal{P}_+^1 as a homogenous space:

$$\mathcal{C} = \mathcal{P}_+^1 / \text{SU}(2),$$

which is what survives of Eq. (3). (We are not distinguishing here between \mathcal{P}_+^1 and its twofold cover.) Incidentally, note that unlike phase space, the relativistic state space \mathcal{C} is a perfectly covariant object. In fact, we will see in Sec. 3 that \mathcal{C} carries a geometric structure which combines the space-time geometry with the phase space (symplectic) geometry in a natural way.

Our approach, which seeks to combine the basic structures of Relativity, quantum mechanics, and classical mechanics, will therefore be to construct unitary representations of \mathcal{P}_+^1 on spaces of functions over \mathcal{C} . (Actually, we will see that due to the positivity of the energy, \mathcal{C} can be imbedded in a four-dimensional complex manifold \mathcal{T} ; this turns out to be extremely useful, bringing in the methods of complex geometry and analysis.)

Thus \mathcal{W} should also be viewed, not as a subgroup of \mathcal{G}_1 , but as its homogeneous space. In retrospect, the group property of \mathcal{W} appears to be physically irrelevant. What is important is that the invariance group \mathcal{G}_1 of the nonrelativistic theory act upon the state space \mathcal{W} and that this action be transitive (otherwise we may as well look at orbits of \mathcal{G}_1 in \mathcal{W}). Viewed in this way, the relativistic and nonrelativistic theories link up smoothly (see Refs. 8,9).

Incidentally, the fact that \mathcal{W} is a group means that the formalism of nonrelativistic quantum mechanics could be based entirely on \mathcal{W} rather than \mathcal{G}_1 (this is, in fact, the usual practice). When that is done, the discovery of *spin* is a surprising empirical fact. Indeed, this was the historical path, for only later was it realized that the theory could be based on \mathcal{G}_1 (or, rather, on the extension of \mathcal{G}_1 by "dynamics"; see Ref. 11). By contrast, the relativistic theory necessarily implies the possibility of spin since \mathcal{C} is not a group, and we are forced to use \mathcal{P}_+^1 .

3. LOCALIZATION AND THE PHASE-SPACE APPROACH

The insight gained in the last section suggests a new point of departure for the study of the relation between classical and quantum mechanics: Rather than *lift* the position coordinates to the status of observables (namely, the generators Q_k of the group \mathcal{W}) in order to form a "quantized symplectic structure" such as the CCR, it appears far more natural to *lower* the momentum observables to the status of parameters (roughly, the velocity coordinates on \mathcal{C}) which then join the space-time variables x_μ to form a covariant version of classical state space. Indeed, while position operators can be defined for relativistic "elementary systems," their introduction is known to be fundamentally at odds with covariance. We shall now briefly review this situation, assuming for simplicity that we are dealing with a massive

scalar particle.

The first systematic, group-theoretical account was given by Newton and Wigner,¹³ who defined *localized states* as quantum states which, when arbitrarily displaced in space, become orthogonal to themselves and which have some additional, rather obvious, properties. They then showed that these properties uniquely determine localized states, there being one such state ψ_x for each space point x at time $t = 0$. These states, in turn, uniquely determine a set of commuting Hermitian operators Q_k (the Newton-Wigner position operators) of which they are simultaneous (nonnormalizable) eigenvectors with eigenvalues x_k .

And here begin the difficulties. First of all, the ψ_x are not covariant; viewed from a moving frame, a localized state is anything but "localized." In fact, the time-translation generator is $P_0 = (m^2 + \mathbf{P}^2)^{1/2} = (m^2 - \Delta)^{1/2}$ (Δ is the spatial Laplacian), which is a nonlocal operator. It follows that an arbitrarily small time after being "localized," ψ_x is spread all over the Universe! Thus, *assuming that states can be localized would seem to conflict with causality.* (In this connection, see also Ref. 14.) This puts into question the very concept of position as an observable. Since the Lorentz boosts involve P_0 as well, a similar catastrophe happens when, instead of getting evolved in time, ψ_x is boosted to a moving frame.

Various modifications of the Newton-Wigner idea have been attempted. By dropping the requirement that ψ_{x+a} be orthogonal to ψ_x for $a \neq 0$ and requiring covariance under \mathcal{P}_+^1 instead, a much more reasonable set of states is obtained which, however, cannot be used to define position operators.¹⁵ In spite of the vast industry devoted to the localization problem (see Refs. 16, 20 for comprehensive reviews), no fully satisfactory solution has been found.

These considerations further support our idea of abandoning position operators, and with them the CCR, as fundamental concepts in quantum mechanics. To see what is gained by doing so, let us first take stock of what is lost. In nonrelativistic quantum mechanics, the "canonical" symmetry between the Q 's and P 's gives rise to a great deal of freedom of choice between different but equivalent "representations" of the theory (that is, of \mathcal{W}). There is the *Q-representation*, in which the Q_k are diagonal and states are expressed as functions over configuration space. In the *P-representation*, the P_k are diagonal and states are expressed as functions over momentum space. The Plancherel theorem, equating the L^2 -norm of a function with that of its Fourier transform, tells us that the two representations are not only *mathematically* equivalent (i.e., related by a unitary transformation), but also *physically* equivalent: each has a standard probabilistic interpretation, and Nature has no preference between them. In addition to the Q - and P -representations, there is an infinite variety of (again, completely equivalent) *phase-space representations* where neither the Q 's nor the P 's are diagonal and states are represented by functions $f(\mathbf{q}, \mathbf{p})$ over *classical* phase space (see, e.g., Ref. 17 and the references therein). Again, the norms in these representations are L^2 -norms (giving once more a probabilistic interpretation) and are related to the norms in the Q - and P -representation by generalized "Plancherel" theorems.

We have seen that the transition to relativistic theory “breaks” the symmetry between Q ’s and P ’s. The P ’s still exist as natural quantum observables while the Q ’s become burdensome objects of questionable value. Let us see how this “symmetry-breaking” is reflected in the existence and nature of different “representations” of the theory, i.e., of \mathcal{P}^1_+ . The easiest to construct is the P -representation, where the energy-momentum P_μ is diagonal and states are functions over the momentum space, which is the upper mass hyperboloid $p_0 = (m^2 + p^2)^{1/2}$. The Hilbert-space norm is still an L^2 -norm, but not with respect to Lebesgue measure as in the nonrelativistic theory. Instead, the proper (i.e., \mathcal{P}^1_+ -invariant) measure is d^3p/p_0 . There is also a Q -representation, in which the Newton–Wigner operators Q_k are diagonal, states are functions over configuration space, and the norm is the L^2 -norm with respect to Lebesgue measure on R^3 . However, this representation is not natural from a relativistic point of view, a direct consequence of the noncovariance of the localized states ψ_x which in the P -representation have the form $\psi_x(p) = \sqrt{p_0} \exp(-ix \cdot p)$. For although the Q -representation has a non-negative candidate for a probability density [namely, the integrand $\rho(x,t) \equiv |\langle \psi_x | f_t \rangle|^2$ in the expression for $\langle f_t | f_t \rangle$, where f_t is the state at time t], there exists no conserved current whose time component is $\rho(x,t)$. See Ref. 13. Hence, from a physical point of view, the Q -representation is highly unsatisfactory. Finally, there is what I will call the *space–time representation*, in which neither the P ’s nor the Q ’s are diagonal but states are functions over space–time, given by solutions of the Klein–Gordon equation:

$$f(x) = \int_{\Omega} e^{-ixp} \hat{f}(p) d^3p/p_0, \quad (5)$$

where $xp = x_0 p_0 - \mathbf{x} \cdot \mathbf{p}$, Ω is the mass hyperboloid, and $\hat{f}(p)$ is square-integrable with respect to d^3p/p_0 . The inner product in the space–time representation is given as follows¹⁸: Choose a three-dimensional spacelike submanifold S of space–time (S is a generalized configuration space). Then the inner product of two solutions (which *a priori* depends on S) is given by

$$\langle f | g \rangle_S = i \int_S \left[\overline{f(x)} \frac{\partial g(x)}{\partial x^\mu} - \frac{\partial \overline{f(x)}}{\partial x^\mu} g(x) \right] \hat{d}\hat{x}^\mu, \quad (6)$$

where $\hat{d}\hat{x}^\mu$ is the Hodge dual^{5–7} of dx_μ (roughly, the 3-form $dx_0 \wedge dx_1 \wedge dx_2 \wedge dx_3$ with dx_μ missing) and summation over μ is implied. Two facts concerning (6) are worth noting.

(a) $\langle f | g \rangle_S$ is actually independent of the choice of S . This is due to the fact that

$$J_\mu(x) \equiv i \left[\overline{f(x)} \frac{\partial f(x)}{\partial x^\mu} - \frac{\partial \overline{f(x)}}{\partial x^\mu} f(x) \right], \quad (7)$$

is a “conserved current,” i.e.,

$$\frac{\partial J_\mu}{\partial x_\mu} = 0, \quad (8)$$

because $f(x)$ satisfies the Klein–Gordon equation. [There is no loss of generality in setting $g(x) = f(x)$ in (6), i.e., in consid-

ering $\|f\|_S^2 \equiv \langle f | f \rangle_S$ instead of $\langle f | g \rangle_S$, since we can always recover the latter by the “polarization identity.”] Thus the space–time representation, like the P -representation (but *not* the Q -representation) is *manifestly covariant*.

(b) $\|f\|_S^2$ is *not* an L^2 -norm. In fact, it even turns out that the integrand $J_\mu(x) dx^\mu$ need not be nonnegative, although the total integral is positive-definite! (This fact, which was only discovered around 1967, is discussed in Refs. 19 and 20.) This makes it impossible to give a probabilistic interpretation to the space–time representation.

We have described three relativistic representations: the P , Q , and space–time representations. Although these are all mathematically equivalent in the sense of being related by unitary transformations, they are certainly not physically equivalent. Thus, the space–time representation has a conserved current which cannot be interpreted as carrying probability, while the Q -representation has a probability density but no conserved current. The only satisfactory representation seems to be the P -representation which, however, makes no reference at all to space–time; hence *all sense of “locality” appears to be lost!*

What about the many “phase-space” representations of the nonrelativistic theory? Do they survive the transition to relativity? It turns out that the usual methods^{17,21} of constructing such representations do not work: they would define the norm by integrating over the time variable x_0 as well as other variables, resulting in divergent norms. (The reason is that the time-translation generator T_0 is not in the center of \mathcal{A} .) However, there is a simple, new construction that does work, both in the nonrelativistic and relativistic theories. It gives a “phase-space” representation of \mathcal{P}^1_+ which, in the limit $c \rightarrow \infty$, goes over smoothly to a phase-space representation of the central extension of the Galilean group.⁹ This method is based on the very condition which makes the problem of localization so difficult: the positivity of the energy, or, in covariant terms, the *spectral condition*.²²

(S): *Only those unitary representations of \mathcal{P}^1_+ are physically relevant for which the joint spectrum of the energy-momentum operators P_μ is contained in the closure \bar{V}_+ of the forward light cone V_+ .*

Like assumption (Q), of which it is a refinement, condition (S) is meant to apply to *any* isolated relativistic quantum system, hence is quite general. In fact, (Q) and (S) are two of the cornerstones of “axiomatic” quantum field theory.²² Some general elements of our “phase-space approach” can be derived from (Q) and (S) even before specializing to irreducible representations, i.e., “particles.” Namely, it follows immediately that the space–time manifold of the theory (regarded as a homogeneous space of \mathcal{P}^1_+) has a *canonical complexification* called the *forward tube*:

$$\mathcal{T} \equiv \{z = x - iy \in C^4 | y \in V_+\}. \quad (9)$$

For if $y \in V_+$, then (S) implies that the operator $yP \equiv y^\mu P_\mu$ is nonnegative. Hence the operator $\exp(-yP)$ is bounded, and we can extend the unitary group of spacetime translations $U(x) \equiv \exp(-ixP)$ to a (bounded, weakly holomorphic) semigroup

$$U(z) \equiv U(x - iy) \equiv \exp[-i(x - iy)P] \\ = \exp(-ixP) \exp(-yP). \quad (10)$$

Thus any objects in the theory which are functions over space-time (such as quantum fields or particle wave functions) can be extended to \mathcal{T} .

We now return to the simple case of a single, free massive scalar particle. (Some of the conclusions we shall reach apply also to much more general situations, such as interacting quantum fields; see Ref. 23.) Using (10), a solution $f(x)$ of the Klein-Gordon equation [given by (5)] can be extended to \mathcal{T} as

$$f(z) = \int_{\Omega} e^{-izp} \hat{f}(p) d^3p/p_0, \quad (11)$$

which is *holomorphic* (complex-analytic) in \mathcal{T} . For each $z \in \mathcal{T}$ define

$$e_z(p) = e^{ipz}, \quad (12)$$

where $\bar{z} = x + iy$ is the complex conjugate of z . If w is another point in \mathcal{T} , the inner product of e_z and e_w (regarded as states in the P -representation) is

$$\langle e_z | e_w \rangle = \int_{\Omega} e^{-iz - i\bar{w}} d^3p/p_0 \\ = -2i\Delta_+(z - \bar{w}), \quad (13)$$

where Δ_+ is the (analytic continuation to \mathcal{T} of the) Wightman two-point function for the free massive scalar field.²² Since $\Delta_+(z - \bar{z})$ is finite, it follows that each of the e_z 's is normalizable, and obviously $f(z) = \langle e_z | \hat{f} \rangle$, the inner product being that in the P -representation.

The states e_z turn out to have some very important properties, namely

(1) For any $z = x - iy$ in \mathcal{T} , e_z is an optimal wave-packet "focused" about the event x and traveling with expected energy-momentum $\langle P_{\mu} \rangle$ proportional to y_{μ} (see Refs. 8,9). The width of this wave packet in its rest frame at time x_0 (i.e., at the instant of maximal focus) is a monotone increasing function F of the parameter

$$\lambda \equiv (y_{\mu} y^{\mu})^{1/2}, \quad (14)$$

(see Fig. 1). This suggests that \mathcal{T} may be regarded as an *extended classical phase space for the particle* (extended, because it contains the time and "energy" dimensions as well as the position and momentum dimensions).

(2) The e_z 's are *covariant* under \mathcal{P}_+^1 . That is, if $U(a, A)$ is the unitary operator representing a Lorentz transformation A followed by a space-time translation a , then

$$U(a, A) e_z = e_{Az + a}. \quad (15)$$

Thus the e_z 's are covariant "localized states." Of course, they are not *sharply* localized in *space* but only *softly* localized in *phase space*, as mentioned in Sec. 1. (In technical terms, they do not give rise to a projection-valued measure but merely to a positive-operator valued measure.²⁴ In fact, in the limit $y \rightarrow 0$, e_z goes to $e_x = \exp(ixp)$, which coincides with the "Lorentz invariant localized states" obtained by Phillips.¹⁵ However, the e_z 's have numerous advantages over the e_x 's, such as normalizability and the possibility of defining a conserved probability current (see below). We

shall refer to objects which depend on z through e_z as *microlocal*.

(3) Let S be a three-dimensional submanifold of space-time R^4 (a prospective configuration space as in Eq. (6), except note that we are not assuming it to be spacelike!). For fixed $\lambda > 0$, let Ω_{λ} be the hyperboloid $y^2 \equiv y_{\mu} y^{\mu} = \lambda^2$ in V_+ (Ω_{λ} is something like a momentum or velocity space). Let σ be the six-dimensional submanifold of \mathcal{T} given by

$$\sigma = \{x - iy \in \mathcal{T} | x \in S, y \in \Omega_{\lambda}\}. \quad (16)$$

If S were spacelike, σ would be a kind of *classical phase space*. We wish to define a phase-space representation corresponding to σ . Thus we need a 6-dimensional measure $d\mu_{\sigma}(z)$ on σ such that

$$\|f\|_{\sigma}^2 = \int_{\sigma} |f(z)|^2 d\mu_{\sigma}(z) \quad (17)$$

defines a \mathcal{P}_+^1 -invariant norm on the space of all holomorphic solutions given by Eq. (11). A natural way to obtain $d\mu_{\sigma}$ is as follows⁹: Since we wish to think of σ as a classical phase space, we need a symplectic form⁵⁻⁷ α_{σ} on σ . To make everything covariant, we define α_{σ} by starting with an *invariant* 2-form α on \mathcal{T} and restricting it to σ . Without essential loss of generality,^{8,9} we may take

$$\alpha = dy_{\mu} \wedge dx^{\mu}. \quad (18)$$

Then it turns out that α_{σ} is a symplectic form on σ if and only if S is *space-or-lightlike*, i.e., its normal $n(x)$ satisfies $n^2 \geq 0$. That is, σ is a phase space if and only if S is a (generalized) configuration space! This result is interesting in itself, since usually the geometry of space-time is considered not to be naturally compatible with the geometry of phase space. (Note that σ is essentially a "slice" of the relativistic state space \mathcal{C} introduced in Sec. 2.) Assuming S to be space-or-lightlike, we now have a full-fledged phase space $(\sigma, \alpha_{\sigma})$. Furthermore, everything is covariant: Poincaré transformations

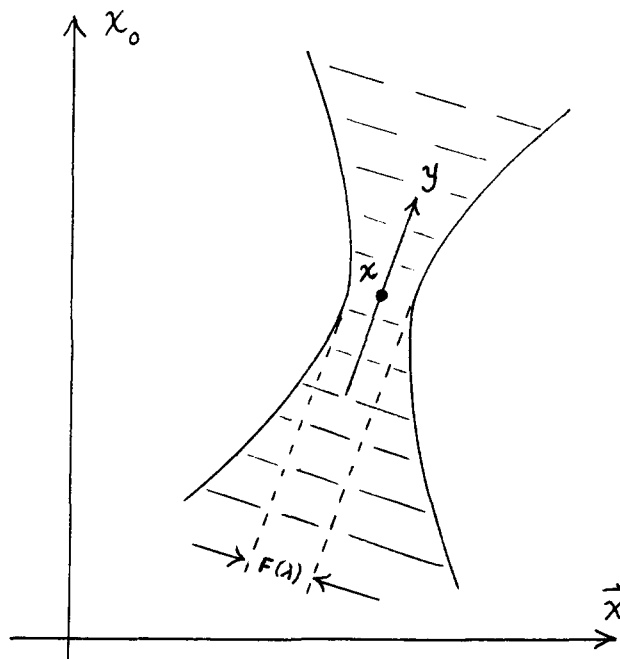


FIG. 1. Schematic diagram of e_z .

transformations.^{8,9} We now define $d\mu_\sigma$ as the Liouville measure of σ , i.e., $d\mu_\sigma$ is the measure defined by the volume form $\alpha_\sigma \wedge \alpha_\sigma \wedge \alpha_\sigma$ on σ . Equivalently, $d\mu_\sigma$ is the restriction to σ of the 6-form $\alpha \wedge \alpha \wedge \alpha$ on \mathcal{T} . But (18) implies that

$$\alpha \wedge \alpha \wedge \alpha = 3! \hat{d}y_\mu \wedge \hat{d}x^\mu, \quad (19)$$

where, as in Eq. (6), $\hat{d}y_\mu$ and $\hat{d}x^\mu$ are Hodge duals (with respect to $g_{\mu\nu}$) of dy^μ and dx_μ , respectively. Thus (17) becomes (modulo a constant factor)

$$\begin{aligned} \|f\|_\sigma^2 &= \int_{S \times \Omega_\lambda} |f(x-iy)|^2 \hat{d}y_\mu \wedge \hat{d}x^\mu \\ &= \int_S \left[\int_{\Omega_\lambda} |f(x-iy)|^2 \hat{d}y_\mu \right] \hat{d}x^\mu \\ &\equiv \int_S \tilde{J}_\mu(x) \hat{d}x^\mu, \end{aligned} \quad (20)$$

automatically giving $\|f\|_\sigma^2$ as the total flux through S of a space-time current $\tilde{J}_\mu(x)$. This current is actually conserved, so that $\|f\|_\sigma^2$ is independent of the choice of S as desired. To see this, let $B_\lambda = \{y \in V_+ | |y_\mu y^\mu| > \lambda^2\}$, so that the oriented boundary of B_λ is $\partial B_\lambda = -\Omega_\lambda$. Then, by Stokes' theorem,

$$\begin{aligned} \tilde{J}_\mu(x) &= - \int_{\partial B_\lambda} |f(x-iy)|^2 \hat{d}y_\mu \\ &= - \int_{B_\lambda} \frac{\partial |f(x-iy)|^2}{\partial y^\mu} d^4y \\ &\equiv \int_{B_\lambda} j_\mu(x-iy) d^4y. \end{aligned} \quad (21)$$

Thus

$$\frac{\partial \tilde{J}_\mu(x)}{\partial x_\mu} = \int_{B_\lambda} \frac{\partial j_\mu(x-iy)}{\partial x_\mu} d^4y.$$

Using the notation

$$\partial_\mu \equiv \frac{\partial}{\partial z^\mu} = \frac{1}{2} \left(\frac{\partial}{\partial x^\mu} + i \frac{\partial}{\partial y^\mu} \right) \quad (22)$$

$$\bar{\partial}_\mu \equiv \frac{\partial}{\partial \bar{z}^\mu} = \frac{1}{2} \left(\frac{\partial}{\partial x^\mu} - i \frac{\partial}{\partial y^\mu} \right)$$

we have

$$\begin{aligned} \frac{\partial j_\mu}{\partial x_\mu} &= - \frac{\partial^2 |f|^2}{\partial x_\mu \partial y^\mu} = i(\partial_\mu + \bar{\partial}_\mu)(\partial^\mu - \bar{\partial}^\mu)(\bar{f}f) \\ &= i(\square - \bar{\square})(\bar{f}f) \\ &= i\bar{f}\square f - i(\square f)\bar{f} \\ &= -i\bar{f}m^2 f + im^2 \bar{f}f = 0. \end{aligned} \quad (23)$$

Here $\square \equiv \partial_\mu \partial^\mu$ is the complex d'Alambertian, and we have used the facts that (a) f satisfies the Klein-Gordon equation and (b) f is holomorphic in \mathcal{T} . As we have seen, the second condition comes from (S), i.e., the positivity of the energy.

Note also that $j_\mu(z)$ is a microlocal current which, by (23), is conserved with respect to space-time variations. By the holomorphy of f ,

$$\begin{aligned} j_\mu(z) &= - \frac{\partial}{\partial y^\mu} |f|^2 = i(\partial_\mu - \bar{\partial}_\mu)(\bar{f}f) \\ &= i\bar{f} \frac{\partial f}{\partial z^\mu} - i \frac{\partial \bar{f}}{\partial \bar{z}^\mu} f \\ &= i \left[\bar{f} \frac{\partial f}{\partial x^\mu} - \frac{\partial \bar{f}}{\partial x^\mu} f \right]; \end{aligned}$$

hence Eq. (21) shows that our current $\tilde{J}_\mu(x)$ is a "regularized" version of the usual current $J_\mu(x)$ [Eq. (7)].

Thus we have arrived at a large family of equivalent phase-space representations of \mathcal{P}_+ , one for each σ . Moreover, these representations have a certain advantage over the P -, Q -, and space-time representations discussed earlier. Namely, they are manifestly covariant (unlike the Q -representation), admit a probability interpretation (unlike the space-time representation), and retain some notion of locality (unlike the P -representation). Although these representations describe scalar particles, they easily generalize to particles with spin: simply interpret $|f|^2$ as including a summation over spin indices.

We summarize the phase-space approach by noting^{8,9} that the covariant wavepackets e_z provide a conceptual bridge between classical and quantum mechanics and give rise to a "continuous resolution of the identity"

$$\int_\sigma |e_z\rangle \langle e_z| d\mu_\sigma(z) = 1, \quad (24)$$

for each classical phase space σ .

4. HOLOMORPHIC GAUGE THEORY

We have seen that strict localizability of relativistic quantum particles is untenable, and that microlocality promises to be an adequate substitute. In this section we explore one consequence of this hypothesis, namely the modifications it suggests in the notion of local gauge invariance.

Consider, once again, a massive scalar particle in a state $f(z)$ [Eq. (11)]. Since the probability density [with respect to the Liouville measure $d\mu_\sigma(z)$] of finding it "at" $z \in \sigma$ is

$$\rho(z) = |f(z)|^2, \quad (25)$$

it follows that $f(z)$ and $\exp(i\phi) f(z)$ define the same physical state, for any real constant ϕ . Let us now imitate the idea of local gauge invariance²⁵ by allowing ϕ to be a function of z . We immediately run into trouble. If $\phi(z)$ is real (but not constant), it cannot be holomorphic; hence the holomorphy of $f(z)$ is destroyed. On the other hand, if $\phi(z)$ is holomorphic, it cannot be everywhere real (unless it is constant); thus $|f(z)|^2$ is not preserved. The solution to this dilemma is to modify $\rho(z)$ by introducing a positive weight function $h(z)$ which transforms so as to compensate for the change in $|f(z)|^2$. Define

$$\rho(z) = \overline{f(z)} h(z) f(z), \quad (26)$$

which is invariant under the transformation

$$\begin{aligned} f'(z) &= e^{i\phi(z)} f(z), \\ h'(z) &= e^{2\text{Im}\phi(z)} h(z), \end{aligned} \quad (27)$$

with $\phi(z)$ holomorphic.

More generally, let $f: \mathcal{T} \rightarrow C^n$ be a vector-valued holomorphic function (representing the state of a particle with internal symmetry) and $h(z)$ be a C^∞ function on \mathcal{T} whose values are $n \times n$ positive-definite matrices. (f is a section of $\mathcal{T} \times C^n$ and h is a fiber metric.²⁶) Define the scalar function

$$\rho(z) = f(z)^* h(z) f(z), \quad (28)$$

which is supposed to represent the probability density, relative to $d\mu_\sigma(z)$, of finding the particle "at" $z \in \sigma$. We define a holomorphic gauge transformation as a holomorphic change of frame, i.e.,

$$\begin{aligned} f'(z) &= \chi(z)^{-1} f(z), \\ h'(z) &= \chi(z)^* h(z) \chi(z), \end{aligned} \quad (29)$$

where $\chi(z)$ is a holomorphic function on \mathcal{T} whose values are invertible $n \times n$ matrices. Since $\rho(z)$ is invariant under (29), so is the norm

$$\|f\|_\sigma^2 \equiv \int_\sigma \rho(z) d\mu_\sigma(z), \quad (30)$$

for any fixed phase space σ . To obtain dynamics, we now require that $\|f\|_\sigma^2$ be independent of the configuration space S . Equations (17)–(21), followed in reverse order, show that this will be the case, provided that

$$\frac{\partial^2 \rho}{\partial x_\mu \partial y^\mu} = 0. \quad (31)$$

By (22) and (23), this means that

$$(\bar{\square} - \square)(f^* h f) = \bar{\square}(f^* h) f - f^* \square(h f) = 0. \quad (32)$$

Thus $\|f\|_\sigma$ will be independent of S if we assume that f satisfies the "Klein-Gordon" equation

$$\square(h f) = G f \quad (33)$$

for some Hermitian matrix-valued function $G(z)$. We will determine $G(z)$ by requiring agreement with the usual, space-time theory. Equation (33) can be rewritten as

$$(\partial_\mu + \theta_\mu)(\partial^\mu + \theta^\mu) f = h^{-1} G f, \quad (34)$$

where $\partial_\mu \equiv \partial / \partial z^\mu$ and

$$\theta_\mu = h^{-1} \partial_\mu h = \frac{1}{2} h^{-1} \left(\frac{\partial h}{\partial x^\mu} + i \frac{\partial h}{\partial y^\mu} \right). \quad (35)$$

We will now show that with a proper choice of $G(z)$, Eq. (34) corresponds closely to the standard Klein-Gordon equation for a particle in space-time coupled to a Yang-Mills field. To begin with, since no fiber metric appears in the standard formalism, we "hide" h by choosing a nonsingular matrix $k(z)$ such that

$$h(z) = k(z)^* k(z), \quad (36)$$

and writing

$$\tilde{f}(z) = k(z) f(z), \quad (37)$$

so that $\rho = \tilde{f}^* \tilde{f}$. (This can be done since h is positive-definite.) Using the notation

$$d_\mu = \frac{\partial}{\partial x^\mu}, \quad \bar{d}_\mu = \frac{\partial}{\partial y^\mu},$$

$$\begin{aligned} \partial_\mu &= \frac{\partial}{\partial z^\mu} = \frac{1}{2}(d_\mu + i\bar{d}_\mu), \\ \bar{\partial}_\mu &= \frac{\partial}{\partial \bar{z}^\mu} = \frac{1}{2}(d_\mu - i\bar{d}_\mu), \end{aligned} \quad (38)$$

we have, for any differentiable function $F(z)$,

$$\begin{aligned} h(d_\mu + \theta_\mu)F &= h d_\mu F + (\partial_\mu h)F \\ &= d_\mu(hF) - (d_\mu h)F + (\partial_\mu h)F \\ &= d_\mu(hF) - (\bar{\partial}_\mu h)F \\ &= k^* d_\mu(kF) + (d_\mu k^*)kF - (\bar{\partial}_\mu k^*)kF \\ &\quad - (k^* \bar{\partial}_\mu k)F \\ &= k^* d_\mu(kF) + (\partial_\mu k^*)kF - (k^* \bar{\partial}_\mu k)F, \end{aligned}$$

so that

$$(d_\mu + \theta_\mu)F = k^{-1} [d_\mu + (k^*)^{-1} \partial_\mu k^* - \bar{\partial}_\mu k \cdot k^{-1}] (kF), \quad (39)$$

or

$$\mathcal{D}_\mu F = k^{-1} D_\mu(kF), \quad (40)$$

where

$$\mathcal{D}_\mu = \frac{\partial}{\partial x^\mu} + \theta_\mu, \quad (41)$$

$$D_\mu = \frac{\partial}{\partial x^\mu} + iA_\mu(z),$$

with

$$A_\mu(z) = i\bar{\partial}_\mu k \cdot k^{-1} - i(k^*)^{-1} \partial_\mu k^* \quad (42)$$

Hermitian. Thus, using $\bar{\partial}_\mu f = 0$ and (40),

$$\begin{aligned} (\partial_\mu + \theta_\mu)(\partial^\mu + \theta^\mu) f &= (\mathcal{D}_\mu - \bar{\partial}_\mu)(\mathcal{D}^\mu - \bar{\partial}^\mu) f \\ &= \mathcal{D}_\mu \mathcal{D}^\mu f - \bar{\partial}_\mu (\mathcal{D}^\mu f) \\ &= k^{-1} D_\mu D^\mu \tilde{f} - (\bar{\partial}_\mu \theta^\mu) f. \end{aligned} \quad (43)$$

Since we wish $\tilde{f}(z)$ to correspond to the usual space-time wavefunction, assume

$$D_\mu D^\mu \tilde{f} = -m^2 \tilde{f}. \quad (44)$$

Then (34) determines that

$$\begin{aligned} G(z) &= -h(m^2 + \bar{\partial}_\mu \theta^\mu) \\ &= -m^2 h - h [(\bar{\partial}_\mu h^{-1}) \partial^\mu h + h^{-1} \bar{\partial}_\mu \partial^\mu h] \\ &= -m^2 h + (\bar{\partial}_\mu h) h^{-1} (\partial^\mu h) - \bar{\partial}_\mu \partial^\mu h, \end{aligned} \quad (45)$$

which is clearly Hermitian as required. Hence by (40) and (44), our "Klein-Gordon" equation (34) finally becomes

$$\mathcal{D}_\mu \mathcal{D}^\mu f = -m^2 f, \quad (46)$$

with \mathcal{D}_μ given by (41).

We can now give the exact correspondence of our theory with the usual one: Holomorphic solutions $f(z)$ of (46) [if such exist; see Remark (3) in Sec. 5] correspond to space-time solutions $\tilde{f}(x)$ for a Klein-Gordon particle coupled to a space-time Yang-Mills field with potentials $A_\mu(x)$, where

$$\tilde{f}(x) = \lim_{y \rightarrow 0} \tilde{f}(z) = k(x) f(x), \quad (47)$$

$$A_\mu(x) = \lim_{y \rightarrow 0} A_\mu(z),$$

$A_\mu(z)$ being given by (42). The norm (30) can be rewritten as

$$\begin{aligned} \|f\|_\sigma^2 &= \int_\sigma \rho(z) d\mu_\sigma(z) \\ &= - \int_S \left[\int_{\partial B_\lambda} \rho(z) \widehat{d}y_\mu \right] \widehat{d}x^\mu \\ &= - \int_S \left[\int_{B_\lambda} \frac{\partial \rho}{\partial y^\mu} d^4y \right] \widehat{d}x^\mu \\ &\equiv \int_{S \times B_\lambda} j_\mu(z) d^4y \widehat{d}x^\mu, \end{aligned} \quad (48)$$

where $j_\mu(z)$ is a *microlocal*, space-time-conserved current, i.e.,

$$\frac{\partial j_\mu(z)}{\partial x_\mu} = 0, \quad (49)$$

by (31). Now since f is holomorphic we have $\partial_\mu f = d_\mu f$; thus (40) gives

$$\begin{aligned} j_\mu(z) &= - \frac{\partial}{\partial y^\mu} (f^* h f) \\ &= i(\partial_\mu - \bar{\partial}_\mu)(f^* h f) \\ &= i f^* \partial_\mu (h f) - i \bar{\partial}_\mu (f^* h) \cdot f \\ &= i f^* h \mathcal{D}_\mu f - i (\mathcal{D}_\mu f)^* \cdot h f \\ &= i \tilde{f}^* D_\mu \tilde{f} - i (D_\mu \tilde{f})^* \cdot \tilde{f} \\ &\equiv i \tilde{f}^* \vec{D}_\mu \tilde{f}; \end{aligned} \quad (50)$$

hence

$$\|f\|_\sigma^2 = \int_S \tilde{J}_\mu(x) \widehat{d}x^\mu, \quad (51)$$

where

$$\tilde{J}_\mu(x) = i \int_{B_\lambda} \tilde{f}(z)^* \vec{D}_\mu \tilde{f}(z) d^4y, \quad (52)$$

which is a “regularized” version of the usual current

$$J_\mu(x) = i \tilde{f}(x)^* \vec{D}_\mu \tilde{f}(x). \quad (53)$$

Let us briefly summarize what we have done in vector bundle terms.^{26,28} (See also Ref. 27.) $f(z)$ is a holomorphic section of the trivial vector bundle $\mathcal{F} \times C^n$, $h(z)$ is a fiber metric, and the fiberwise inner product of $f(z)$ with itself is the probability density:

$$\rho(z) = f^* h f \equiv \langle f, f \rangle(z). \quad (54)$$

The exterior derivative splits into two parts of type (1,0) and (0,1):

$$\begin{aligned} d &= dx^\mu \frac{\partial}{\partial x^\mu} + dy^\mu \frac{\partial}{\partial y^\mu} \\ &= dz^\mu \frac{\partial}{\partial z^\mu} + d\bar{z}^\mu \frac{\partial}{\partial \bar{z}^\mu} \end{aligned}$$

$$\equiv \partial + \bar{\partial}, \quad (55)$$

with $\partial^2 = \bar{\partial}^2 = \partial \bar{\partial} + \bar{\partial} \partial = 0$. Thus

$$\begin{aligned} d \langle f, f \rangle &= (\bar{\partial} + \partial)(f^* h f) \\ &= \bar{\partial}(f^* h) \cdot f + f^* \partial(h f) \\ &= (\mathcal{D} f)^* h f + f^* h \mathcal{D} f \\ &= \langle \mathcal{D} f, f \rangle + \langle f, \mathcal{D} f \rangle, \end{aligned} \quad (56)$$

where, since $\bar{\partial} f = 0$,

$$\mathcal{D} f \equiv (d + \theta) f = (\partial + \theta) f = h^{-1} \partial(h f), \quad (57)$$

with

$$\theta = h^{-1} \partial h = \theta_\mu dz^\mu. \quad (58)$$

θ is called the *canonical connection* determined by h , and \mathcal{D} is the *covariant derivative* with respect to θ . Equation (56) shows that θ is metric-compatible with h . The *curvature* of θ is

$$\begin{aligned} \Theta &\equiv d\theta + \theta \wedge \theta \\ &= \bar{\partial}\theta + \partial\theta + \theta \wedge \theta \\ &= \bar{\partial}\theta, \end{aligned} \quad (59)$$

since the last two terms cancel due to the “integrability condition”

$$\begin{aligned} \partial\theta &= \partial(h^{-1} \partial h) = \partial(h^{-1}) \wedge \partial h \\ &= -(h^{-1} \partial h \cdot h^{-1}) \wedge \partial h \\ &= -\theta \wedge \theta. \end{aligned} \quad (60)$$

Since Θ and θ are our versions of the Yang-Mills field and potential respectively (see below), Eq. (59) shows that *in the microlocal theory, the field is linear in the potential, even for nonabelian gauge groups ($n > 1$)!* This is a remarkable fact which may prove useful, since the usual nonlinearity (whose source we will see below) makes solutions very difficult to find. Under the holomorphic gauge transformation (29), we have

$$\begin{aligned} \theta' &= (h')^{-1} \partial h' = \chi^{-1} h^{-1} (\chi^*)^{-1} \chi^* \partial(h\chi) \\ &= \chi^{-1} h^{-1} (\partial h \cdot \chi + h \cdot \partial \chi) \\ &= \chi^{-1} \theta \chi + \chi^{-1} \partial \chi, \end{aligned} \quad (61)$$

hence $\Theta = \bar{\partial}\theta$ is invariant, by the holomorphy of χ . The *Bianchi identity* reads

$$\begin{aligned} d\Theta &= d(d\theta + \theta \wedge \theta) = d\theta \wedge \theta - \theta \wedge d\theta \\ &= (\Theta - \theta \wedge \theta) \wedge \theta - \theta \wedge (\Theta - \theta \wedge \theta) \\ &= \Theta \wedge \theta - \theta \wedge \Theta \\ &\equiv [\Theta, \theta], \end{aligned} \quad (62)$$

and $\bar{\partial}\Theta = \bar{\partial}^2\theta = 0$ implies $\partial\Theta = [\Theta, \theta]$.

In vector bundle terms, the correspondence with the usual theory is obtained as follows: Since $\chi(z)$ in Eq. (29) is holomorphic, it cannot be everywhere unitary unless it is constant. Hence the structure group corresponding to holomorphic gauge transformations is $GL(n, C)$. As $k(z)$ is non-singular, the transformation $f \rightarrow \tilde{f} = kf$ may be regarded as a

nonholomorphic gauge transformation which amounts to a reduction²⁸ of the structure group to the usual one, $U(n)$, since $\rho = \tilde{f}^* \tilde{f}$ is merely invariant under

$$\tilde{f}'(z) = U(z) \tilde{f}(z), \quad (63)$$

with $U(z)$ unitary ($k' = Uk$). Under $f \rightarrow \tilde{f}$, the connection θ reduces to

$$\begin{aligned} iA(z) &= k^{-1} \partial k - \bar{\partial} k k^{-1}, \\ &\equiv \gamma^* - \gamma, \end{aligned} \quad (64)$$

where $\gamma = \bar{\partial} k k^{-1}$. The curvature of A corresponds to the extension to \mathcal{S} of the usual, space-time Yang-Mills field, and is given by

$$\begin{aligned} iF(z) &= d(iA) + (iA) \wedge (iA) \\ &= i(dA + iA \wedge A) \\ &= (d\gamma^* + \gamma^* \wedge \gamma^*) \\ &\quad - (d\gamma - \gamma \wedge \gamma) - \gamma^* \wedge \gamma - \gamma \wedge \gamma^*, \end{aligned} \quad (65)$$

But γ and γ^* satisfy integrability conditions similar to that satisfied by θ [Eq. (60)]:

$$\begin{aligned} \bar{\partial} \gamma - \gamma \wedge \gamma &= 0, \\ \partial \gamma^* + \gamma^* \wedge \gamma^* &= 0. \end{aligned} \quad (66)$$

Hence

$$iF = \bar{\partial} \gamma^* - \partial \gamma - \gamma^* \wedge \gamma - \gamma \wedge \gamma^*. \quad (67)$$

Note that the usual nonlinearity between the Yang-Mills connection and field is a direct result of the reduction from holomorphic to unitary gauge freedom. That is, F is no longer linear in A or γ [Eqs. (65) and (67)].

The relation between Θ and F is

$$iF = k \Theta k^{-1}, \quad (68)$$

so that

$$\langle f, \Theta f \rangle = f^* h \Theta f = i \tilde{f}^* F \tilde{f}. \quad (69)$$

Now in components,

$$\begin{aligned} \Theta(z) &= \bar{\partial}(h^{-1} \partial h) = \bar{\partial} h^{-1} \wedge \partial h + h^{-1} \bar{\partial} \partial h \\ &= (\bar{\partial}_\mu h^{-1} \partial_\nu h + h^{-1} \bar{\partial}_\mu \partial_\nu h) d\bar{z}^\mu \wedge dz^\nu \\ &= \Theta_{\bar{\mu}\nu} d\bar{z}^\mu \wedge dz^\nu \\ &= \Theta_{\bar{\mu}\nu} (dx^\mu \wedge dx^\nu + dy^\mu \wedge dy^\nu + idy^\mu \wedge dx^\nu \\ &\quad + idy^\nu \wedge dx^\mu) \\ &= \frac{1}{2} (\Theta_{\bar{\mu}\nu} - \Theta_{\bar{\nu}\mu}) (dx^\mu \wedge dx^\nu + dy^\mu \wedge dy^\nu) \\ &\quad + \frac{1}{2} i (\Theta_{\bar{\mu}\nu} + \Theta_{\bar{\nu}\mu}) (dy^\mu \wedge dx^\nu + dy^\nu \wedge dx^\mu), \end{aligned} \quad (70)$$

which is seen to have a symmetric part as well as an antisymmetric part.

The usual (space-time) Yang-Mills field $F(x)$ is the restriction (pull-back) of $F(z)$ to R^4 (let $y \rightarrow 0$ and $dy_\mu \rightarrow 0$).

Hence

$$\begin{aligned} F(x) &= -ik(x) \Theta(x) k(x)^{-1} \\ &= -\frac{1}{2} ik(x) [\Theta_{\bar{\mu}\nu}(x) - \Theta_{\bar{\nu}\mu}(x)] k(x)^{-1} dx^\mu \wedge dx^\nu, \end{aligned} \quad (71)$$

so that only the antisymmetric part shows up in pure space-time. It would obviously be of interest to explore what phys-

ical significance the symmetric part might have. If one can be found, it may represent a concrete dividend of the approach advocated here.

5. CONCLUDING REMARKS

(1) The view held by many physicists, that in quantum mechanics "everything" becomes an operator, does not appear to be consistent with relativity theory. The alternative proposed here is to introduce the quantum formalism for a free system through the theory of unitary representations of \mathcal{P}_+^1 or related groups, then add interactions through microlocal gauge freedom, which necessitates the introduction of a fiber metric. In this way, the energy and momentum have a natural place as operators in the theory, and potentials are represented geometrically as the components of a connection determined by the fiber metric, to be regarded as functions of the underlying phase space variables rather than the position operators. The desired "canonical" symmetry between positions and momenta survives in its "classical" form: together with the Lorentz metric, it defines the geometric structure of the extended phase space on which the theory is based [see also Ref. 27, Sec. 4(c)]. The "quantized" version of the canonical structure, in the form of the canonical commutation relations, is seen to be both unnecessary and in conflict with relativistic covariance.

(2) In addition to resolving the inconsistencies of the usual theory, the new theory deals directly and covariantly with extended particles rather than point particles. Since microlocality implies nonlocality in space, it may be hoped that the microlocal theory provides a natural framework for the description of hadrons and other extended particles without the necessity to resort to such difficult (and perhaps at times ad hoc) methods as integro-differential equations.

(3) Note that the requirement that Eq. (46) possess holomorphic solutions $f(z)$ puts considerable restrictions on the choice of $h(z)$. In fact, (46) should perhaps be regarded, together with the Cauchy-Riemann equations for f , as a coupled system of equations for f and h . These equations may have to be supplemented with some counterpart of the Yang-Mills equations, to determine the reaction of $h(z)$ to the presence of matter [represented by $f(z)$]. For this, it might be helpful to develop a Lagrangian approach to the microlocal theory, which is still missing.

¹D. J. Simms and N. M. J. Woodhouse, *Lectures on Geometric Quantization*, Springer Lecture Notes in Physics No 53 (Springer-Verlag, Berlin, 1976).

²J. Sniatycki, *Geometric Quantization and Quantum Mechanics* (Springer-Verlag, Berlin, 1980).

³V. Guillemin and S. Sternberg, *Geometric Asymptotics* (Am. Math. Soc., Providence, R.I., 1977).

⁴B. Simon, *Commun. Math. Phys.* **71**, 247 (1980).

⁵R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin,

- New York, 1978), 2nd ed.
- ⁶V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer-Verlag, Berlin, 1978).
- ⁷W. Thirring, *Classical Dynamical Systems* (Springer-Verlag, Berlin, 1978).
- ⁸G. Kaiser, "Relativistic Coherent-State Representations," in *Group-Theoretical Methods in Physics*, edited by R. T. Sharp and B. Kolman (Academic, New York, 1977); thesis, Univ. of Toronto, 1977.
- ⁹G. Kaiser, *J. Math. Phys.* **18**, 952 (1977); **19**, 502 (1978).
- ¹⁰H. Weyl, *The Theory of Groups and Quantum Mechanics* (Dover, New York, 1950).
- ¹¹V. S. Varadarajan, *Geometry of Quantum Theory* (Van Nostrand, Princeton, N. J. 1970), Vol. 2.
- ¹²P. Roman and J. P. Leveille, *J. Math. Phys.* **15**, 1760 (1974); **15**, 2053 (1974).
- ¹³T. D. Newton and E. P. Wigner, *Rev. Mod. Phys.* **21**, 400 (1949).
- ¹⁴G. C. Hegerfeldt and S. N. M. Ruijsenaars, "Remarks on Causality, Localization and Spreading of Wave Packets," Princeton Univ. preprint, 1980.
- ¹⁵T. O. Phillips, *Phys. Rev.* **136**, B893 (1964).
- ¹⁶A. J. Kalnay, "The Localization Problem," in *Problems in the Foundations of Physics*, edited by M. Bunge (Springer-Verlag, Berlin, 1971).
- ¹⁷J. R. Klauder, *J. Math. Phys.* **4**, 1055, 1058 (1963); also **5**, 177 (1964).
- ¹⁸S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Row, Peterson, Evanston, Ill., 1961).
- ¹⁹B. Gerlach, D. Gromes, and J. Petzold, *Z. Phys.* **202**, 401 (1967).
- ²⁰A. O. Barut and S. Malin, *Rev. Mod. Phys.* **40**, 632 (1968).
- ²¹A. M. Perelomov, *Commun. Math. Phys.* **26**, 222 (1972).
- ²²R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics and All That* (Benjamin, New York, 1964).
- ²³G. Kaiser, "Relativistic Quantum Theory in Complex Spacetime," to appear in *Differential-Geometrical Methods in Mathematical Physics*, edited by J.-M. Souriau, P. L. Garcia, and A. Perez-Rendon, Springer Lecture Notes in Mathematics (Springer-Verlag, Berlin, 1981).
- ²⁴E. B. Davies, *Quantum Theory of Open Systems* (Academic, New York, 1976).
- ²⁵E. S. Abers and B. W. Lee, "Gauge Theories," *Phys. Reports* **9**, No. 1 (1973).
- ²⁶R. O. Wells, *Differential Analysis on Complex Manifolds* (Springer-Verlag, Berlin, 1980), 2nd ed.
- ²⁷G. Kaiser, "Holomorphic Gauge Theory," in *Geometric Methods in Mathematical Physics*, edited by G. Kaiser and J. E. Marsden, Springer Lecture Notes in Mathematics No. 775 (Springer-Verlag, Berlin, 1980).
- ²⁸S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963 and 1969), Vols. I and II.

Path integrals with a periodic constraint: The Aharonov–Bohm effect

Christopher C. Bernido and Akira Inomata

Department of Physics, State University of New York at Albany, Albany, New York 12222

(Received 3 January 1980; accepted for publication 28 March 1980)

The Aharonov–Bohm effect is formulated in terms of a constrained path integral. The path integral is explicitly evaluated in the covering space of the physical background to express the propagator as a sum of partial propagators corresponding to homotopically different paths. The interference terms are also calculated for an infinitely thin solenoid, which are found to contain the usual flux dependent shift as the dominant observable effect and an additional topological shift unnoticeable in the two slit interference experiment.

PACS numbers: 03.65Db, 03.70.+k

I. INTRODUCTION

The nonintegrable phase factor of a wavefunction appears to assume a new role in quantum physics. In analyzing the Aharonov–Bohm effect, Wu and Yang¹ have pointed out that electromagnetism is underscribed by the field strength but overdescribed by the loop integral of the vector potential, and concluded that electromagnetism is the gauge-invariant manifestation of nonintegrable phase factor.

The Aharonov–Bohm effect² (or the AB effect) is an observable quantum phenomenon³ in which the role of the vector potential is conspicuous. The effect has been well studied and well confirmed, but not necessarily well understood.⁴ One interpretation is that the vector potential remembers what is going on elsewhere, while the other claims that the physical consequences of this effect can be obtained without the use of any potential inasmuch as the space is multiply connected. In fact, the topological nature of the AB effect has been long recognized, but no quantitative prescription has been given for linking the nonintegrable phase factor and the topological feature of the background space. A qualitative framework has been suggested by Schulman⁵ to deal with the effect in a multiply connected space. The mathematical object to be computed in this framework is a propagator expressed as a path integral in the covering space of the background physical space. In a recent paper,⁶ we have developed a method to evaluate path integrals under a periodic constraint and indicated its possible application to the AB effect.

The purpose of the present paper is to provide an explicit formulation of the propagator for the AB effect in a multiply connected space. First we analyze in Sec. II an idealized AB experiment in connection with propagators. In Sec. III, we formulate the AB effect in terms of a constrained path integral and calculate the corresponding propagator in the covering space. The propagator turns out to be a sum of partial propagators belonging to homotopically inequivalent paths. Section IV deals with an infinitely thin solenoid by which we elaborate the interference shift. The standard flux dependent shift is found as the dominant observable effect. In addition, there is a topological effect which is not noticeable in the two slit interference experiment.

II. IDEALIZED AHARONOV–BOHM EXPERIMENT

The setup for the measurement of the Aharonov–Bohm

effect may be idealized in two dimensions as shown in Fig. 1. It consists of the particle source S, the detector D, and the circular cross section of an impenetrable solenoid confining a magnetic field B inside. Charged particles emitted from S are to arrive at D via field-free regions around the solenoid. There are various paths that link S and D. In the quantum aspect, some of paths are equivalent and some are not. Path 1 and path 2 in Fig. 1 typically indicate two inequivalent paths. Although the space outside the solenoid is free from the B -field, the vector potential A is not zero inside or outside insofar as the B -field inside the solenoid remains nonvanishing. It is possible to think of an impenetrable solenoid which contains no flux but functions to separate paths into inequivalent classes. To isolate the pure electromagnetic effect, however, we stipulate that $A \neq 0$ means the presence of the solenoid packed with the flux between S and D, and that $A = 0$ means no flux and no solenoid.

The solution of Schrödinger's equation for a charged particle in a vector potential A is given in the path-dependent form

$$\psi_\alpha(\mathbf{r}) = \psi_0(\mathbf{r}) \exp\left\{\frac{ie}{\hbar c} \int_{\text{path}\alpha} \mathbf{A} \cdot d\mathbf{r}\right\}, \quad (2.1)$$

where $\psi_0(\mathbf{r})$ is the potential-free solution.⁷ The wavefunction $\Psi(\mathbf{r})$ effective to the measurement at D is the sum of solutions corresponding to inequivalent paths,

$$\psi(\mathbf{r}) = \sum_\alpha \psi_\alpha(\mathbf{r}). \quad (2.2)$$

Observable are interference patterns that depend on the flux in the solenoid. What we wish to achieve in this paper is to describe the same effect in terms of propagators expressed as constrained path integrals.

The propagator $K(\mathbf{r}'', \mathbf{r}'; \tau)$ is the kernel of the integral equation

$$\psi(\mathbf{r}'', \tau) = \int K(\mathbf{r}'', \mathbf{r}'; \tau) \psi(\mathbf{r}'; 0) d\mathbf{r}'. \quad (2.3)$$

For a particle free from the vector potential, we write

$$\psi_0(\mathbf{r}'', \tau) = \int K_0(\mathbf{r}'', \mathbf{r}'; \tau) \psi_0(\mathbf{r}'; 0) d\mathbf{r}'. \quad (2.4)$$

Multiplying both sides of (2.4) by the same path-dependent phase factor as in (2.1), we can formally construct a solution in the presence of the potential

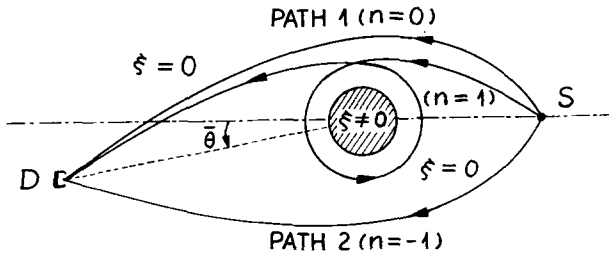


FIG. 1. Idealized Aharonov-Bohm setup.

$$\psi_\alpha(\mathbf{r}'', \tau) = \sum_\beta \int K_{\alpha\beta}(\mathbf{r}'', \mathbf{r}'; \tau) \psi_\beta(\mathbf{r}', 0) d\mathbf{r}', \quad (2.5)$$

where

$$K_{\alpha\beta}(\mathbf{r}'', \mathbf{r}'; \tau) = K_0(\mathbf{r}'', \mathbf{r}'; \tau) \exp \left\{ \frac{ie}{\hbar c} \left(\int_\alpha^{\mathbf{r}''} - \int_\beta^{\mathbf{r}'} \right) \mathbf{A} \cdot d\mathbf{r} \right\}. \quad (2.6)$$

The summation over β in (2.5) is necessarily associated with the integration over \mathbf{r}' . Now, substitution of (2.5) into (2.2) will lead us to the formal expression (2.3) provided that the double sum over α and β is separable as

$$\sum_{\alpha, \beta} K_{\alpha\beta} \psi_\beta = K \sum_\beta \psi_\beta. \quad (2.7)$$

Actually the propagator $K_{\alpha\beta}$ in (2.6) has the following path-dependence. To be consistent with the paths of ψ_α and ψ_β , it should follow path β back to infinity and switch over to path α at infinity to proceed to point \mathbf{r}'' . Therefore, $K_{\alpha\beta}$ may be written as $K_{\alpha-\beta}$. If α and β range from $-\infty$ to ∞ , then the separation (2.7) is possible with

$$K(\mathbf{r}'', \mathbf{r}'; \tau) = \sum_{n=-\infty}^{\infty} K_n(\mathbf{r}'', \mathbf{r}'; \tau). \quad (2.8)$$

As the wavefunction (2.2) results in the flux-dependent interference patterns, the propagator (2.8) should contain all information concerning the observable effect.

Since the solenoid is assumed impenetrable, the space of the particle motion M is the plane of the idealized Aharonov-Bohm experiment minus the cross section of the solenoid. Everywhere in M , $\nabla \times \mathbf{A} = 0$ and hence $\mathbf{A} = \nabla \Lambda$ where Λ is an arbitrary scalar function of \mathbf{r} . If a cut is made on M from the side of the solenoid to infinity, then the path-dependent integrals in (2.6) become integrable on the resultant singly connected patch. It is therefore clear that the path dependence of the phase factor in (2.6) is wholly of topological origin. Thus the Aharonov-Bohm problem is reduced to showing that the full propagator can be expressed as a sum of partial propagators belonging to all topologically inequivalent paths as given by (2.8).

III. PATH INTEGRAL APPROACH

In Feynmann's prescription,⁸ the propagator is given as a sum over all histories

$$K(\mathbf{r}'', \mathbf{r}'; \tau) = \lim_{N \rightarrow \infty} A_N \int \exp \left\{ \frac{i}{\hbar} \sum_{j=1}^N S_j \right\} \prod_{j=1}^{N-1} d\mathbf{r}_j, \quad (3.1)$$

where $\mathbf{r}' = \mathbf{r}_0$, $\mathbf{r}'' = \mathbf{r}_N$, and the segmental action

$$S_j = \int_{t_{j-1}}^{t_j} L(\mathbf{r}, \dot{\mathbf{r}}) dt, \quad (3.2)$$

is defined for a given classical Lagrangian L . Since the back-

ground space M is multiply connected, special care has to be taken in summing over all paths. As Schulman has suggested,⁵ we can most conveniently go over to the universal covering space M^* of M to perform the path integral (3.1). In what follows, we shall explicitly calculate the propagator (3.1) for the AB effect in M^* by using the technique developed earlier.⁶

The Lagrangian for a particle of mass μ and charge e moving in the potential \mathbf{A} is

$$L = \frac{1}{2} \mu \dot{\mathbf{r}}^2 + (e/c) \mathbf{A} \cdot \dot{\mathbf{r}}. \quad (3.3)$$

To be more specific, we take the origin of the coordinates at the center of the solenoid and assume the vector potential of the form

$$\mathbf{A} = -(\phi_0/2\pi)(y\mathbf{i} - x\mathbf{j})/r^2 \quad (r < \bar{r}), \quad (3.4)$$

$$\mathbf{A} = -(\phi_0/2\pi)(y\mathbf{i} - x\mathbf{j})/r^2 \quad (r > \bar{r}), \quad (3.5)$$

where $\phi_0 = \pi \bar{r}^2 B$ and $r^2 = x^2 + y^2$. In addition, we require that the particle cannot penetrate the circular region of radius \bar{r} around the origin unless $\mathbf{A} = 0$. This requirement would effectively introduce an additional constraint potential V^c into the Lagrangian (3.3). The effect of the constraint via the action is such that

$$\exp \left\{ \frac{i}{\hbar} S_j^c \right\} = \Theta(r_j - \bar{r}) \quad (3.6)$$

where $S^c = -\int V^c dt$ and $\Theta(x)$ is the step function which has two values, unity for $x > 0$ and zero for $x \leq 0$. Since the potential outside the solenoid (3.5) can be written in polar coordinates as

$$\mathbf{A} = (\phi_0/2\pi) \nabla \theta \quad (r > \bar{r}), \quad (3.7)$$

the segmental action corresponding to the Lagrangian (3.3) is given by

$$S_j = \frac{1}{2} \mu \int_{t_{j-1}}^{t_j} \dot{r}^2 dt + \xi \hbar \int_{t_{j-1}}^{t_j} \dot{\theta} dt, \quad (3.8)$$

which may be approximated by

$$S_j = \frac{1}{2} \mu (r_j^2 + r_{j-1}^2)/\epsilon - \mu(r_j r_{j-1})/\epsilon \cos \Delta\theta_j + \xi \hbar \Delta\theta_j, \quad (3.9)$$

where $\epsilon = t_j - t_{j-1} = \tau/N$, $\xi = e\phi_0/(2\pi\hbar c)$, and

$$\Delta\theta_j = \int_{t_{j-1}}^{t_j} \dot{\theta} dt. \quad (3.10)$$

The angular variable θ varies from 0 to 2π , and hence we usually assume $0 \leq \Delta\theta_j \leq 2\pi$ for all j in (3.9). However, the path can loop around the impenetrable region many times, requiring the integral $\Delta\theta_j$ to vary from $-\infty$ to ∞ . Therefore, the path integral (3.1), if calculated with the usual assumption, gives only a partial propagator, which belongs to a class of paths topologically constrained by $0 \leq \Delta\theta_j \leq 2\pi$. For a full account, we have to consider the contributions from paths belonging to all homotopically different classes. In performing the path integration over the angular variable, we can either remain in the physical space M by assuming

$$\Delta\theta_j = \theta_j - \theta_{j-1} + 2\pi n, \quad \text{with } 0 \leq \theta_j \leq 2\pi, \quad (3.11)$$

or go over to the covering space M^* by taking

$$\Delta\theta_j = \theta_j - \theta_{j-1}, \quad \text{with } -\infty < \theta_j < \infty. \quad (3.12)$$

The partial propagator belonging to a class of homotopically equivalent paths which entangle around the solenoid by an angle φ ($-\infty < \varphi < \infty$) can be expressed as

$$K_\varphi(\mathbf{r}'', \mathbf{r}'; \tau) = \lim_{N \rightarrow \infty} A_N \int \prod_{j=1}^N \left\{ \delta(\varphi - \Delta\theta_j) \Theta(r_j - \bar{r}) \times \exp\left[\frac{i}{\hbar} S_j\right] \right\} \prod_{j=1}^{N-1} d\mathbf{r}_j \quad (3.13)$$

or

$$K_\varphi(\mathbf{r}'', \mathbf{r}'; \tau) = (2\pi)^{-1} \lim_{N \rightarrow \infty} A_N \int \int e^{i\lambda\varphi} \times \prod_{j=1}^N \left\{ \Theta(r_j - \bar{r}) \exp\left[\frac{i}{\hbar} \hat{S}_j\right] \right\} \prod_{j=1}^{N-1} d\mathbf{r}_j d\lambda \quad (3.14)$$

where

$$\hat{S}_j = \frac{1}{2} \mu(r_j^2 + r_{j-1}^2) / \epsilon - \mu(r_j r_{j-1} / \epsilon) \cos \Delta\theta_j - \lambda' \hbar \Delta\theta_j, \quad (3.15)$$

with $\lambda' = \lambda - \xi$. It is obvious that the full propagator is obtained by integrating (3.13) over φ ,

$$K(\mathbf{r}'', \mathbf{r}'; \tau) = \int K_\varphi(\mathbf{r}'', \mathbf{r}'; \tau) d\varphi. \quad (3.16)$$

To carry out the path integrations in (3.14), we first rewrite (3.15) in the form

$$\hat{S}_j = \frac{1}{2} \mu(r_j^2 + r_{j-1}^2) / \epsilon - \mu(r_j r_{j-1} / \epsilon) \times \cos(\theta_j - \theta_{j-1} - \lambda' \hbar \epsilon / \mu r_j r_{j-1}) - \lambda'^2 \hbar^2 \epsilon / 2\mu r_j r_{j-1}. \quad (3.17)$$

Then we use the following asymptotic relation with the modified Bessel function $I_\nu(z)$ for large $|z|$ and $|\arg(z)| < \pi/2$

$$\exp\{z \cos[\theta + i(\lambda'/z)] - \lambda'^2/(2z)\} \sim \sum_{m=-\infty}^{\infty} e^{im\theta} I_{m+\lambda'}(z) \quad (3.18)$$

to find

$$\exp\left[\frac{i}{\hbar} \sum_{j=1}^N \hat{S}_j\right] = \prod_{j=1}^N \left\{ \sum_{m_j=-\infty}^{\infty} \exp[im_j(\theta_j - \theta_{j-1})] \times R_{m_j+\lambda'}(r_j, r_{j-1}; \epsilon) \right\}, \quad (3.19)$$

where

$$R_\nu(r, r'; \epsilon) = \exp[i\mu(r^2 + r'^2)/2\hbar\epsilon] I_{|\nu|}(\mu r r' / i\hbar\epsilon). \quad (3.20)$$

Interchanging the multiplications and summations on the right-hand side of (3.19), and substituting the result into (3.14), we complete the angular integrations of (3.14)

$$K_\varphi(\mathbf{r}'', \theta''; \mathbf{r}', \theta'; \tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \exp[im(\theta'' - \theta') + i\lambda\varphi] \times Q_{m+\lambda-\xi} d\lambda, \quad (3.21)$$

where

$$Q_\nu(\mathbf{r}'', \mathbf{r}'; \tau) = \lim_{N \rightarrow \infty} (2\pi)^{N-1} A_N \times \int \prod_{j=1}^N R_\nu(r_j, r_{j-1}; \tau) \Theta(r_j - \bar{r}) \prod_{j=1}^{N-1} (r_j dr_j). \quad (3.22)$$

Use of (3.21) in (3.16) yields the full propagator in the form

$$K(\mathbf{r}'', \mathbf{r}'; \tau) = \sum_{m=-\infty}^{\infty} \exp[im(\theta'' - \theta')] Q_{m-\xi}(\mathbf{r}'', \mathbf{r}'; \tau). \quad (3.23)$$

The integration variable in (3.21) may be changed from λ to $\lambda - m + \xi$, so that

$$K_\varphi = \frac{1}{2\pi} \int_{-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \exp[im(\theta'' - \theta' - \varphi) + i(\lambda + \xi)\varphi] Q_\lambda d\lambda. \quad (3.24)$$

Furthermore, the use of the Poisson sum formula

$$\sum_{m=-\infty}^{\infty} \exp(im\theta) = 2\pi \sum_{n=-\infty}^{\infty} \delta(\theta + 2\pi n) \quad (3.25)$$

enables us to write (3.24) in the form

$$K_\varphi(\mathbf{r}'', \theta''; \mathbf{r}', \theta'; \tau) = \sum_{n=-\infty}^{\infty} \delta(\theta'' - \theta' - \varphi + 2\pi n) \times e^{i\xi\varphi} \int_{-\infty}^{\infty} e^{i\lambda\varphi} Q_\lambda(\mathbf{r}'', \mathbf{r}'; \tau) d\lambda. \quad (3.26)$$

The δ -function appearing in the above expression implies that the contributions to K_φ will be only from the terms corresponding to the angle

$$\varphi = \theta'' - \theta' + 2\pi n. \quad (3.27)$$

If we choose θ' and θ'' so that $0 \leq \theta' < \theta'' \leq 2\pi$, then the integral number n signifies, if positive, an n , times counterclockwise entanglement of the particle's path around the solenoid, and if negative, an $|n| - 1$ times clockwise entanglement. Corresponding to $n = 0$ and $n = -1$, respectively, are path 1 and path 2 in Fig. 1 for which no entanglements occur. The number n , therefore, classifies all homotopically inequivalent paths. The full propagator is now expressed via (3.16) and (3.26) as the sum of the partial propagators, each of which belongs to a class of equivalent paths

$$K(\mathbf{r}'', \theta''; \mathbf{r}', \theta'; \tau) = \sum_{n=-\infty}^{\infty} K_n(\mathbf{r}'', \theta''; \mathbf{r}', \theta'; \tau), \quad (3.28)$$

where

$$K_n = \exp[i\xi(\theta'' - \theta' + 2\pi n)] \times \int_{-\infty}^{\infty} \exp[i\lambda(\theta'' - \theta' + 2\pi n)] Q_\lambda(\mathbf{r}'', \mathbf{r}'; \tau) d\lambda. \quad (3.29)$$

This result potentially assures the occurrence of the AB effect as has been discussed in Sec. II.

IV. PATHS OF DOMINANT CONTRIBUTIONS

For further elaboration, let us consider the following two limiting situations where the radius \bar{r} of the solenoid becomes commonly zero. The first case is the trivial one for which the solenoid disappears together with the flux so that $\xi = 0$. In the second case, the solenoid, being infinitely thin, maintains a nonvanishing flux $\phi_0 \neq 0$ at $r = 0$. For both cases, there will no longer be a domain in which the interior potential (3.4) is meaningful, but the exterior potential (3.5) will prevail over the entire space.

In the limit $\bar{r} \rightarrow 0$, the integrations in (3.22) can be carried out with the help of the formula⁹

$$\int_0^\infty \exp(iar^2) I_\nu(-iar) I_\nu(-ibr) r dr$$

$$= (i/2\alpha) \exp[-i(a^2 + b^2)/4\alpha] I_\nu(-iab/2\alpha), \quad (4.1)$$

valid for $\text{Re}(\nu) > -1$ and $\text{Re}(\alpha) > 0$, the result being

$$Q_\lambda(r'', r'; \tau) = (\mu/2\pi i \hbar \tau) \exp[i(r'^2 + r''^2)\mu/2\hbar\tau] \\ \times I_{|\lambda|}(r'r''\mu/i\hbar\tau), \quad (4.2)$$

where $A_N = (2\pi i \epsilon \hbar / \mu)^{-N}$ have been used. Upon substitution of (4.2), the full propagator (3.23) takes the form

$$K = (\mu/2\pi i \hbar \tau) \exp[i(r'^2 + r''^2)\mu/2\hbar\tau] \\ \times \sum_{m=-\infty}^{\infty} \exp[im(\theta'' - \theta')] I_{|m-\xi|}(r'r''\mu/i\hbar\tau), \quad (4.3)$$

and the partial propagator (3.29) reads

$$K_n = (\mu/2\pi i \hbar \tau) \exp[i(r'^2 + r''^2)\mu/2\hbar\tau] \\ \times \exp[i\xi(\theta'' - \theta' + 2\pi n)] \\ \times \int_{-\infty}^{\infty} \exp[i\lambda(\theta'' - \theta' + 2\pi n)] I_{|\lambda|}(r'r''\mu/i\hbar\tau) d\lambda. \quad (4.4)$$

If $\xi = 0$ ($\phi_0 = 0$), the full propagator (4.3) becomes, as expected, the free particle propagator

$$K(r'', r'; \tau) = (\mu/2\pi i \hbar \tau) \exp[i\mu(r'' - r')^2/2\hbar\tau]. \quad (4.5)$$

In the case where $\xi \neq 0$ ($\phi_0 \neq 0$), it is difficult to compute the integral in (4.4). Since r' and r'' may be chosen so that $r'r''\mu \gg \hbar\tau$, we approximate the modified Bessel function in (4.4) by its asymptotic form for large $|z|$

$$I_{|\lambda|}(z) \approx (2\pi z)^{-1/2} \exp\{z - \frac{1}{2}(\lambda^2 - \frac{1}{4})/z\}, \quad (4.6)$$

where $z = r'r''\mu/i\hbar\tau$, and evaluate the integral in (4.4). Namely, for $|z|$ large

$$\int_{-\infty}^{\infty} e^{i\lambda\theta} I_{|\lambda|}(z) d\lambda \approx \exp\{z + (8z)^{-1} - \frac{1}{2}z\theta^2\}. \quad (4.7)$$

Thus, for $r'r''\mu \gg \hbar\tau$, the partial propagator (4.4) can be expressed as

$$K_n = (\mu/2\pi i \hbar \tau) \exp[i(r'' - r')^2\mu/(2\hbar\tau) + i\hbar\tau/(8r'r''\mu)] \\ \times \exp[i\xi(\theta'' - \theta' + 2\pi n) \\ + \frac{1}{2}i(r'r''\mu/\hbar\tau)(\theta'' - \theta' + 2\pi n)^2]. \quad (4.8)$$

From this immediately follow the interference terms ($n \neq m$),

$$K^*_n K_m + K^*_m K_n \\ = 2(\mu/2\pi \hbar \tau)^2 \cos[2\pi(m-n)\{\xi + (r'r''\mu/\hbar\tau)\bar{\theta}\} \\ + 2\pi^2(r'r''\mu/\hbar\tau)(m-n)(m+n+1)], \quad (4.9)$$

where we have set $\theta'' - \theta' = \bar{\theta} + \pi$. Apparently, the interference patterns depend not only on the relative position of the detector and the flux enclosed, but also on the winding numbers. The flux dependent shift is the proper AB effect. In addition, there is an interference shift due to the winding of paths about the solenoid. Even if $\xi = 0$, the winding shift appears to remain effective. However, the total propagator for $\xi = 0$ becomes the free propagator (4.5) and such shifts

cancel out. When $\xi \neq 0$, the winding shift disappears only if $m + n + 1 = 0$. For instance, the interference via path 1 ($m = 0$) and path 2 ($n = -1$) in Fig. 1 is not affected. The winding dependence is certainly a pure topological effect, which is not noticeable in an ideal two-slit interference experiment. This seems to suggest a contradiction to Schulman's observation⁵ that the physical consequences of the Aharonov-Bohm effect are obtainable without the use of an electromagnetic potential.

The maximum contribution to the AB effect comes from the smallest value of $|m - n| > 0$, that is, from $|m - n| = 1$. Therefore, the maximum effect free from the smearing of the winding shifts occurs when $m = 0$ and $n = -1$, or when $m = -1$ and $n = 0$. Namely, the AB effect at $\bar{\theta} = 0$ is to be dominated by the term

$$K^*_0 K_{-1} + K^*_{-1} K_0 = 2(\mu/2\pi \hbar \tau)^2 \cos(2\pi\xi), \quad (4.10)$$

with $\xi = e\phi_0/(2\pi\hbar c)$. This corresponds to the standard result⁷ obtained from the two wavefunctions taking path 1 and path 2 in Fig. 1. An appropriate device, such as an electrical biprism, may be used to lower the contributions other than those from path 1 and path 2. How one can select other pairs of inequivalent paths to detect the topological shift is yet to be answered.

Finally, it must be remarked that the asymptotic form (4.6) is valid provided that $\text{Re}(z) > 0$. What we have for z in (4.3) is, however, $-ir'r''\mu/(\hbar\tau)$, which is a pure imaginary number. To circumvent this difficulty, as proposed earlier,¹⁰ we assume a complex mass $\mu = \mu_R + i\mu_I$ ($\mu_I > 0$), and take the limit $\mu_I \rightarrow 0$ after the integration in (4.7). In fact, a similar limiting procedure has already been implicitly assumed in the calculation of (3.19). Furthermore, we notice that the sum of the noninterference terms ($n = m$) blows up if evaluated directly from (4.8). A finite result for the sum can be obtained from (4.4) via the above limiting procedure.

ACKNOWLEDGMENT

One of the authors (A.I.) would like to thank Professor H. Sato for his hospitality at Research Institute for Fundamental Physics, Kyoto University, where part of this work was done.

¹T. T. Wu and C. N. Yang, Phys. Rev. D **12**, 3845 (1975); D **14**, 437 (1976).

²Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485 (1959); **123**, 1511 (1961).

³F. G. Werner and D. R. Brill, Phys. Rev. Lett. **4**, 349 (1960); R. G. Chambers, Phys. Rev. Lett. **5**, 3 (1960).

⁴The controversy still persists; see P. Bocchieri and A. Loinger, Nuovo Cimento **47 A**, 475 (1978); D. Bohm and B. J. Hiley, Nuovo Cimento **52 A**, 295 (1979).

⁵L. S. Schulman, J. Math. Phys. **12**, 304 (1971). See also M. G. G. Laidlaw and C. M. DeWitt, Phys. Rev. D **3**, 1375 (1971) and J. S. Dowker, J. Phys. A **5**, 936 (1972).

⁶A. Inomata and V. A. Singh, J. Math. Phys. **19**, 2318 (1978).

⁷See, e.g., J. J. Sakurai, *Advanced Quantum Mechanics* (Addison Wesley, Reading, MA, 1967), p. 16.

⁸R. P. Feynman, Rev. Mod. Phys. **20**, 367 (1948); R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).

⁹D. Peak and A. Inomata, J. Math. Phys. **10**, 1422 (1969).

¹⁰W. Langguth and A. Inomata, J. Math. Phys. **20**, 499 (1979).

A new set of coherent states for the isotropic harmonic oscillator: Coherent angular momentum states

A. J. Bracken and H. I. Leemon

Department of Mathematics, University of Queensland, St. Lucia, Queensland 4067, Australia

(Received 27 November 1979; accepted for publication 3 March 1980)

The Hamiltonian for the oscillator has earlier been written in the form

$$H = \hbar\omega(2\nu^\dagger\nu + \lambda^\dagger\cdot\lambda + \frac{3}{2}),$$

where ν^\dagger and ν are raising and lowering operators for $\nu^\dagger\nu$, which has eigenvalues k (the “radial” quantum number), and λ^\dagger and λ are raising and lowering 3-vector operators for $\lambda^\dagger\cdot\lambda$, which has eigenvalues l (the total angular momentum quantum number). A new set of coherent states for the oscillator is now defined by diagonalizing ν and λ . These states bear a similar relation to the commuting operators H , L^2 , and L_3 (where L is the angular momentum of the system) as the usual coherent states do to the commuting number operators N_1 , N_2 , and N_3 . It is proposed to call them coherent angular momentum states. They are shown to be minimum-uncertainty states for the variables ν , ν^\dagger , λ , and λ^\dagger , and to provide a new quasiclassical description of the oscillator. This description coincides with that provided by the usual coherent states only in the special case that the corresponding classical motion is circular, rather than elliptical; and, in general, the uncertainty in the angular momentum of the system is smaller in the new description. The probabilities of obtaining particular values for k and l in one of the new states follow independent Poisson distributions. The new states are overcomplete, and lead to a new representation of the Hilbert space for the oscillator, in terms of analytic functions on $\mathbb{C}\times\mathbb{K}_3$, where \mathbb{K}_3 is the three-dimensional complex cone. This space is related to one introduced recently by Bargmann and Todorov, and carries a very simple realization of all the representations of the rotation group.

PACS numbers: 03.65.Fd, 03.65.Ca, 03.65.Ge

1. INTRODUCTION

The states of the isotropic harmonic oscillator with Hamiltonian

$$H = \frac{\mathbf{p}^2}{2M} + \frac{1}{2}M\omega^2 \mathbf{x}^2 \quad (1)$$

are frequently described in terms of the basis vectors $|n_1, n_2, n_3\rangle$ which are eigenstates of H and also of the number operators

$$N_i = a_i^\dagger a_i, \quad i = 1, 2, 3 \quad (\text{no sum}), \quad (2)$$

where

$$a_i = (2M\hbar\omega)^{-1/2}(i p_i + M\omega x_i). \quad (3)$$

The occupation numbers n_i independently run over the non-negative integers and a_i is a shift operator for N_i , lowering the corresponding eigenvalue n_i by 1.

An alternative approach¹⁻⁵ introduces the “coherent states” $|z\rangle$, which are eigenvectors of the lowering operators

$$a_i |z\rangle = z_i |z\rangle, \quad z_i \in \mathbb{C}. \quad (4)$$

These vectors have many attractive properties. In particular, there is a well-defined sense in which one can say that when the system is progressing through a succession of coherent states its behavior is as close as possible to the behavior of its classical counterpart. The coherent states are therefore justifiably called “quasiclassical” states of the oscillator.

For any three dimensional system, the total angular momentum quantum number, denoted j in general, can take the values $0, 1, 2, \dots$ or the values $\frac{1}{2}, \frac{3}{2}, \dots$. From this we are led to observe that it may be possible and useful to define, by

diagonalizing suitable lowering operators for j , coherent angular momentum states for a variety of systems, including the isotropic oscillator. The latter is the subject of the present work. Several authors⁶⁻¹² have defined and discussed coherent angular momentum states, in particular for systems (such as the rigid rotor) for which the angular momentum operators \mathbf{J}^2 and J_3 provide a complete set of commuting operators. (These states, and those we define in this paper, are not to be confused with the so-called “coherent spin states” which have been widely discussed since their introduction by Radcliffe,¹³ and which are superpositions of eigenvectors of J_3 for a fixed value of \mathbf{J}^2 .) Atkins and Dobson⁷ defined states by exploiting the Schwinger¹⁴ boson calculus for $SU(2)$, and diagonalizing the associated pair of boson annihilation operators. A difficulty here is that the states obtained are superpositions of states of integral and half-integral j , because the boson operators lower j by $\frac{1}{2}$ rather than 1. In order to obtain states which might apply to some physical system with only rotational degrees of freedom, the states with half-integral j had to be rather arbitrarily deleted from the superpositions. Bhaumik, Nag, and Dutta-Roy⁸ avoided this difficulty by constructing, within the Schwinger calculus, two operators quadratic in the boson operators, which lower the value of j by 1. These operators were diagonalized to define coherent angular momentum states, which could then be identified as possible coherent angular momentum states for a system, again having only rotational degrees of freedom. The operators of Bhaumik *et al.* have algebraic properties similar to two components of the 3-vector operator λ we introduce below. However, the operator λ

acts within an entirely different space, namely that within which the oscillator boson operators a_i and a_i^\dagger act. The operators a_i , which form a 3-vector, should not be confused with the boson annihilation operators in the Schwinger calculus, which form a 2-spinor. At no stage in what follows do we work with the Schwinger calculus.

A suitable definition of coherent angular momentum states for a system such as the isotropic oscillator, which possesses translational as well as rotational degrees of freedom, is more difficult than for a system possessing only rotational degrees of freedom, because the dynamics will, in general, couple the degrees of freedom in the former case. Nevertheless, we may hope with the authors mentioned above that coherent angular momentum states can be defined in a suitable way for some systems, and that by analogy with the properties of the usual coherent states, these new states will have one or more of several nice properties. They may exhibit "quasiclassical" behavior, at least in their angular dependence, and they may be useful in examining the behavior of the angular momentum of the system as the classical limit is approached. They may also represent states of "minimum uncertainty" for certain noncommuting variables associated with the angular dependence in the problem and, as they will most likely be overcomplete, they may permit the construction of a representation in which, at the least, the angular dependence of the density matrix for the system can be put in a diagonal form.

With this motivation, we define in this paper a new set of quasiclassical states for the isotropic oscillator and call them "coherent angular momentum states." They bear a similar relation to the commuting operators H , L^2 , and L_3 (where $L = \mathbf{x} \times \mathbf{p}$ is the angular momentum of the system) as the usual coherent states do to the commuting operators H , N_1 , N_2 , and N_3 . In particular, these new states are eigenvectors of a 3-vector operator which lowers the value of the total angular momentum quantum number. We emphasize that the problem is *not* the straightforward one of expressing the usual coherent states for the three-dimensional oscillator as superpositions of the common eigenvectors of H , L^2 , and L_3 , instead of superpositions of the vectors $|n_1, n_2, n_3\rangle$. Such expressions have been obtained and discussed by Mikhailov,¹⁵ but those coherent states are not eigenvectors of any lowering operator for j (or, rather, l in this case). The states we shall define below are in general quite distinct from the usual coherent states, as we shall see. We shall demonstrate that they do have some of the attractive properties mentioned above.

We have shown in an earlier publication¹⁶ (henceforth referred to as BL) that the operator H of Eq. (1) can also be written in the form

$$H = \hbar\omega(2\nu^\dagger\nu + \lambda^\dagger\lambda + \frac{3}{2}), \quad (5)$$

where ν^\dagger and ν are (boson) raising and lowering operators for $\nu^\dagger\nu$ (which we also write as K), while λ^\dagger and λ are raising and lowering operators for $\lambda^\dagger\lambda$ (which we also write as L). The eigenvalues k and l of K and L run over the nonnegative integers independently, and the eigenvalues of H appear in the form $\hbar\omega(2k + l + \frac{3}{2})$. Here k is the "radial" quantum number and l is the total angular momentum quantum number.

Both are familiar from the treatment in the coordinate representation of the eigenvalue problem for H , L^2 , and L_3 . Here we adopt no particular representation.

The basic algebraic relations satisfied by the operators ν , ν^\dagger , λ , and λ^\dagger are (see BL)

$$\begin{aligned} [\nu, \nu^\dagger] &= 1, \\ [\lambda_i, \nu] &= 0 = [\lambda_i^\dagger, \nu^\dagger], \\ [\lambda_i, \nu^\dagger] &= 0 = [\lambda_i^\dagger, \nu], \\ [\lambda_i, \lambda_j] &= 0 = [\lambda_i^\dagger, \lambda_j^\dagger], \\ (2\lambda^\dagger\lambda + 1)[\lambda_i, \lambda_j^\dagger] &= (2\lambda^\dagger\lambda + 1)\delta_{ij} - 2\lambda_i^\dagger\lambda_j, \\ \lambda\cdot\lambda &= 0 = \lambda^\dagger\lambda^\dagger, \\ L_i &= -i\hbar\epsilon_{ijk}\lambda_j^\dagger\lambda_k. \end{aligned} \quad (6)$$

With $K = \nu^\dagger\nu$ and $L = \lambda^\dagger\lambda$, it follows that

$$\begin{aligned} L\lambda &= \lambda(L - 1), \quad L\lambda^\dagger = \lambda^\dagger(L + 1), \\ [L, \nu] &= 0 = [L, \nu^\dagger], \\ K\nu &= \nu(K - 1), \quad K\nu^\dagger = \nu^\dagger(K + 1), \\ [K, \lambda] &= 0 = [K, \lambda^\dagger], \end{aligned} \quad (7)$$

and also that

$$L^2 = L(L + 1)\hbar^2, \quad (8)$$

so that when L has the eigenvalue l , L^2 has the eigenvalue $l(l + 1)\hbar^2$. The two alternative sets of dynamical variables for the isotropic oscillator $\{\nu, \nu^\dagger, \lambda, \lambda^\dagger\}$ and $\{\mathbf{a}, \mathbf{a}^\dagger\}$, are related by the equations

$$\begin{aligned} \nu &= (\mathbf{a}\cdot\mathbf{a})(4K + 4L + 2)^{-1/2}, \\ \lambda_i &= (a_i L - i\hbar^{-1}\epsilon_{ijk}a_j L_k) \\ &\quad \times [(2L + 1)(2K + 2L + 1)]^{-1/2}, \\ a_i &= \lambda_i [(2K + 2L + 1)/(2L + 1)]^{1/2} \\ &\quad + \lambda_i^\dagger \nu [2/(2L + 3)]^{1/2}, \end{aligned} \quad (9)$$

and their conjugates.

As the four lowering operators ν and λ_i commute, we define the coherent angular momentum states as their common eigenvectors. Thus we seek vectors $|z, \xi\rangle$ satisfying

$$\begin{aligned} \nu|z, \xi\rangle &= z|z, \xi\rangle, \\ \lambda_i|z, \xi\rangle &= \xi_i|z, \xi\rangle, \end{aligned} \quad (10)$$

where the eigenvalues z and ξ_i may be expected to be complex since ν and λ_i are not Hermitian. Noting from Eqs. (6) that $\lambda^2 = 0$, we see that ξ is confined to a complex cone

$$\xi^2 = 0. \quad (11)$$

The coherent angular momentum states will therefore be labelled by the four complex numbers z and ξ_i , of which only three are independent, whereas the usual coherent states are labelled by three complex numbers z_i .

Let us remark at this stage that although there is a certain $\text{so}(2, 1) \oplus \text{so}(3, 2)$ Lie algebra underlying the algebra of operators which we use for this system (see BL), the operators ν and λ are not actually in (the complexification of) this Lie algebra, so that the states we define are not coherent states for a Lie algebra or group in the sense of Barut and Girardello¹⁷ or Perelomov,¹⁸ although they are closely relat-

ed to such states. We make some comments on this at the end of Sec. 5.

In Sec. 2 we find, for arbitrary complex z and arbitrary complex ζ satisfying Eq. (11), a nondegenerate normalized vector $|z, \zeta\rangle$ satisfying Eqs. (10), in the form

$$|z, \zeta\rangle = \exp\left(-\frac{1}{2}|z|^2 - \frac{1}{2}|\zeta|^2\right) \times \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \sum_{m=-l}^l a_{klm}(z, \zeta) |klm\rangle, \quad (12)$$

with

$$a_{klm}(z, \zeta) = z^k (\zeta_3)^{l-|m|} (-\epsilon \zeta_- \epsilon)^{|m|} \times [(2l)!/k!2^l l!(l+m)!(l-m)!]^{1/2}, \quad (13)$$

where ϵ is the sign of m and $\zeta_{\pm} = \zeta_1 \pm i\zeta_2$. Here $|klm\rangle$ is the nondegenerate normalized common eigenvector of K, L , and L_3 , as constructed in BL, which in the coordinate representation has the familiar form¹⁹

$$|klm\rangle = (-1)^k \left[\frac{2a^3 k!}{\Gamma(k+l+\frac{3}{2})} \right]^{1/2} \xi^l e^{-\frac{1}{2}\xi^2} \times L_k^{(l+\frac{1}{2})}(\xi^2) Y_{lm}(\theta, \phi), \quad (14)$$

where $a = (M\omega/\hbar)^{1/2}$ and $\xi = ar$ (r, θ , and ϕ are the usual spherical polar coordinates), $L_k^{(l+\frac{1}{2})}$ is the generalized Laguerre polynomial defined as in Ref. (20), and the spherical harmonic Y_{lm} is defined as in Ref. (21). From Eqs. (12) and (14) one can deduce (see Appendix A) that in the coordinate representation

$$|z, \zeta\rangle = \left[\frac{a}{\sqrt{\pi}} \right]^{3/2} \exp\left(-\frac{1}{2}|z|^2 - \frac{1}{2}|\zeta|^2 - \frac{1}{2}a^2|\mathbf{x}|^2\right) \times \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \left[\frac{\Gamma(l+\frac{3}{2})}{\Gamma(k+l+\frac{3}{2})} \right]^{1/2} (-z)^k \times L_k^{(l+\frac{1}{2})}(a^2|\mathbf{x}|^2) \frac{(\sqrt{2ax\cdot\zeta})^l}{l!}, \quad (15)$$

a result we have not been able to express more simply, except in the special case $z = 0$, when it becomes

$$|0, \zeta\rangle = \left[\frac{a}{\sqrt{\pi}} \right]^{3/2} \exp\left(-\frac{1}{2}|\zeta|^2 - \frac{1}{2}a^2|\mathbf{x}|^2 + \sqrt{2ax\cdot\zeta}\right). \quad (16)$$

In Sec. 3 we examine the expectation values of the dynamical variables when the system is in the state $|z, \zeta\rangle$, and find in particular that the probability of obtaining the value l on measurement of the total angular momentum follows a Poisson distribution.

We go on to consider the sense in which the state $|z, \zeta\rangle$ is a "minimum-uncertainty" state for the hermitian variables σ, τ, α , and β , where

$$\sqrt{2\hbar} \nu = \sigma + i\tau$$

and

$$\sqrt{2\hbar} \lambda = \alpha + i\beta. \quad (17)$$

Letting $\langle A \rangle$ denote the expectation value of any observable A for a given state of the system, and defining the dispersions of σ and α for that state by

$$\Delta\sigma = [\langle\sigma^2\rangle - \langle\sigma\rangle^2]^{1/2}$$

and

$$\Delta\alpha = [\langle\alpha\cdot\alpha\rangle - \langle\alpha\rangle\cdot\langle\alpha\rangle]^{1/2}, \quad (18)$$

with similar definitions for $\Delta\tau$ and $\Delta\beta$, we find that, in general,

$$\Delta\sigma\Delta\tau \geq \frac{1}{2}\hbar \quad (19a)$$

and

$$\Delta\alpha\Delta\beta \geq \hbar(1 + \frac{1}{2}\langle(2L+1)^{-1}\rangle). \quad (19b)$$

In the state $|z, \zeta\rangle$ both inequalities become equalities and, moreover,

$$(\Delta\sigma)^2 = (\Delta\tau)^2 = \frac{1}{2}\hbar \quad (20a)$$

and

$$(\Delta\alpha)^2 = (\Delta\beta)^2 = \hbar(1 + \frac{1}{2}\langle(2L+1)^{-1}\rangle). \quad (20b)$$

We show in Sec. 4 that, if the system is in the state $|z_0, \zeta_0\rangle$ at time $t = 0$, then at time t , in the Schrödinger picture, it is in the state $e^{-(3/2)i\omega t}|z(t), \zeta(t)\rangle$, where

$$z(t) = e^{-2i\omega t} z_0$$

and

$$\zeta(t) = e^{-i\omega t} \zeta_0. \quad (21)$$

We then deduce that the expectation values $\langle\sigma\rangle, \langle\tau\rangle, \langle\alpha\rangle$, and $\langle\beta\rangle$ reproduce the corresponding behavior in time of their classical counterparts $\hat{\sigma}, \hat{\tau}, \hat{\alpha}$, and $\hat{\beta}$, as discussed in BL. Moreover, the dispersions of the quantum-mechanical variables remain constant during the motion at their minimum values as in Eqs. (20), so that the coherent angular momentum states can properly be called quasiclassical states.

Corresponding to a given classical motion of the oscillator there are therefore (at least) two quasiclassical descriptions in quantum mechanics, which are distinct in general. One is provided by the usual coherent states, another by coherent angular momentum states. In the special case that the classical motion is circular rather than elliptical, these two quasiclassical descriptions are the same. We deduce this as a consequence of the identification of the coherent angular momentum state $|0, \zeta\rangle$ with the usual coherent state $|z = \zeta\rangle$, an identification which follows from the result (16) and the known form for the states $|z\rangle$ in the coordinate representation.²²

It is important to note, however, that we find that, if the quasiclassical description of a given classical motion is given by coherent angular momentum states, the expectation values $\langle\mathbf{x}\rangle$ and $\langle\mathbf{p}\rangle$ do not exactly reproduce the corresponding behavior of the classical variables $\hat{\mathbf{x}}$ and $\hat{\mathbf{p}}$, unless that classical motion is circular. Rather we find, for example, that $\langle\mathbf{x}\rangle$ follows an elliptical path different from, though in the same plane as, the path followed by $\hat{\mathbf{x}}$. Furthermore, the state $|z, \zeta\rangle$ is not, in general, a minimum-uncertainty state for \mathbf{x} and \mathbf{p} and, when the system evolves in coherent angular momentum states, the dispersions Δx and Δp oscillate, but are always bounded. Therefore, in the coordinate representation (or the momentum representation) the state vector $e^{-3i\omega t/2} \times |z(t), \zeta(t)\rangle$ would appear as a pulsating wavepacket which follows the classical motion approximately. This reflects the fact that the variables \mathbf{x} and \mathbf{p} bear a special relation to the usual coherent states, not the coherent angular mo-

mentum states. In other representations, the *usual* coherent states are also presumably represented by pulsating packets which follow the classical motion only approximately. Furthermore, in the quasiclassical description provided by the usual coherent states, the expectation values of σ, τ, α , and β will not, in general, reproduce exactly the behavior of their classical counterparts.

The outstanding feature of the description by coherent angular momentum states, in the general case of an elliptic orbit, is that the uncertainty in the angular momentum of the system, as best measured, according to Delbourgo,¹¹ by $[\langle L^2 \rangle - \langle L \rangle \cdot \langle L \rangle]^{1/2}$, is smaller than it is for the description by the usual coherent states.

After a brief discussion of the classical limit which is obtained with $|z| \rightarrow \infty, |\xi| \rightarrow \infty$, and $\hbar \rightarrow 0$ we go on to Sec. 5, where we give a completeness relation for the states $|z, \xi\rangle$. They are, in fact, overcomplete and, just as for the usual coherent states, there is associated with these states a Hilbert space of analytic functions with a reproducing kernel. We briefly discuss this space and the associated elegant representation of the dynamical variables $\nu, \nu^\dagger, \lambda, \lambda^\dagger, K, L$, and H .

We conclude with some remarks in Sec. 6 about possible further developments.

2. THE COHERENT ANGULAR MOMENTUM STATES

We look for vectors satisfying Eqs. (10) in the form

$$|z, \xi\rangle = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \sum_{m=-l}^l b_{klm}(z, \xi) |klm\rangle. \quad (22)$$

In BL, Eqs. (53) and (58), we showed that

$$|klm\rangle = c_{klm} (\nu^\dagger)^k (\lambda_\epsilon^\dagger)^{|m|} (\lambda_3^\dagger)^{l-|m|} |0\rangle, \quad (23)$$

where

$$c_{klm} = (-\epsilon)^m [(2l)!/k!2^l!(l-m)!(l+m)!]^{1/2}, \quad (24)$$

ϵ is the sign of m , $|0\rangle$ is a normalized vector on which ν and λ vanish and $\lambda_\pm^\dagger = \lambda_1^\dagger \pm i\lambda_2^\dagger$. Supposing that the vector $|z, \xi\rangle$ of Eq. (22) does satisfy Eqs. (10), then we must have

$$\begin{aligned} b_{klm} |z, \xi\rangle &= \langle klm | z, \xi \rangle \\ &= (c_{klm})^* \langle 0 | \nu^k (\lambda_{-\epsilon})^{|m|} (\lambda_3)^{l-|m|} | z, \xi \rangle \\ &= c_{klm} z^k (\xi_{-\epsilon})^{|m|} (\xi_3)^{l-|m|} \langle 0 | z, \xi \rangle, \end{aligned} \quad (25)$$

where $\lambda_\pm = \lambda_1 \pm i\lambda_2$ and $\xi_\pm = \xi_1 \pm i\xi_2$.

Conversely, taking the coefficients in Eq. (22) to have the form

$$b_{klm}(z, \xi) = n(z, \xi) c_{klm} z^k (\xi_{-\epsilon})^{|m|} (\xi_3)^{l-|m|}, \quad (26)$$

with $n(z, \xi)$ an arbitrary function of z and ξ , one can check that the vector $|z, \xi\rangle$ so defined does satisfy Eqs. (10). To verify this, one needs to use the equations

$$\begin{aligned} \nu |klm\rangle &= k^{1/2} |k-1lm\rangle \\ \lambda_3 |klm\rangle &= \left[\frac{(l-m)(l+m)}{(2l-1)} \right]^{1/2} |kl-1m\rangle \\ \lambda_\pm |klm\rangle &= \pm \left[\frac{(l \mp m)(l \mp m - 1)}{(2l-1)} \right]^{1/2} \\ &\quad \times |kl-1m \pm 1\rangle, \end{aligned} \quad (27)$$

as given in BL [Eqs. (59)], and also to use Eq. (11) in the form

$$\xi_+ \xi_- = -(\xi_3)^2. \quad (28)$$

It is easily seen that this vector $|z, \xi\rangle$ is normalizable for arbitrary complex z and for arbitrary complex ξ satisfying Eq. (11). Using the orthonormality of the vectors $|klm\rangle$, we have

$$\begin{aligned} \langle z', \xi' | z, \xi \rangle &= n(z', \xi')^* n(z, \xi) \sum_{klm} (z'^* z)^k (\xi_3'^* \xi_3)^{l-|m|} \\ &\quad \times [(\xi'_{-\epsilon})^* \xi_{-\epsilon}]^{|m|} (c_{klm})^2. \end{aligned} \quad (29)$$

Because $\xi^2 = 0 = \xi'^2$, we have, with the help of the binomial theorem,

$$\begin{aligned} \sum_{m=-l}^l (\xi_3'^* \xi_3)^{l-|m|} [(\xi'_{-\epsilon})^* \xi_{-\epsilon}]^{|m|} \frac{(2l)!}{2^l(l+m)!(l-m)!} \\ = (\xi'^* \xi)^l, \end{aligned} \quad (30)$$

so that Eq. (29) reduces to

$$\begin{aligned} \langle z', \xi' | z, \xi \rangle &= n(z', \xi')^* n(z, \xi) \sum_{k,l} \frac{(z'^* z)^k}{k!} \frac{(\xi'^* \xi)^l}{l!} \\ &= n(z', \xi')^* n(z, \xi) \exp(z'^* z + \xi'^* \xi). \end{aligned} \quad (31)$$

Thus $|z, \xi\rangle$ is normalized if we take

$$n(z, \xi) = \exp(-\frac{1}{2}|z|^2 - \frac{1}{2}|\xi|^2) \quad (32)$$

and we have, from Eqs. (22), (24), and (26), the final form (12) for the coherent angular momentum states.

With this normalization, we have

$$\begin{aligned} \langle z', \xi' | z, \xi \rangle &= \exp[-\frac{1}{2}(|z'|^2 + |z|^2 + |\xi'|^2 + |\xi|^2) \\ &\quad + z'^* z + \xi'^* \xi], \end{aligned} \quad (33)$$

so that

$$|\langle z', \xi' | z, \xi \rangle|^2 = \exp[-\frac{1}{2}|z' - z|^2 - \frac{1}{2}|\xi' - \xi|^2]. \quad (34)$$

These vectors are, therefore, not orthogonal for $(z', \xi') \neq (z, \xi)$, but they approximate orthogonality as $|z' - z| \rightarrow \infty$ and $|\xi' - \xi| \rightarrow \infty$. We shall see in Sec. 5 that they are overcomplete.

3. EXPECTATION VALUES OF PHYSICAL VARIABLES AND THE MINIMUM UNCERTAINTY PROPERTY

When the system is in the state $|z, \xi\rangle$ we have at once

$$\begin{aligned} \langle \nu \rangle &= z, \quad \langle \nu^\dagger \rangle = z^*, \\ \langle \lambda \rangle &= \xi, \quad \langle \lambda^\dagger \rangle = \xi^*, \end{aligned} \quad (35)$$

so that, using Eqs. (10) and (17),

$$\begin{aligned} \langle \sigma \rangle &= \sqrt{\hbar/2} (z + z^*), \quad \langle \tau \rangle = -i\sqrt{\hbar/2} (z - z^*), \\ \langle \alpha \rangle &= \sqrt{\hbar/2} (\xi + \xi^*), \quad \langle \beta \rangle = -i\sqrt{\hbar/2} (\xi - \xi^*). \end{aligned} \quad (36)$$

Since $K = \nu^\dagger \nu$ and $L = \lambda^\dagger \lambda$ we have

$$\begin{aligned} \langle K \rangle &= |z|^2, \quad \langle L \rangle = |\xi|^2 \\ \text{and} \\ \langle H \rangle &= \hbar\omega(2|z|^2 + |\xi|^2 + \frac{3}{2}), \end{aligned} \quad (37)$$

and from the last of Eqs. (6) we have

$$\langle L \rangle = -i\hbar \xi^* \times \xi. \quad (38)$$

The probability of obtaining the value k on measuring K in the state $|z, \xi\rangle$ is given by

$$\begin{aligned}
p(k) &= \sum_{l,m} |\langle klm|z, \xi\rangle|^2 \\
&= |n(z, \xi)|^2 \frac{|z|^{2k}}{k!} \sum_{l,m} |\xi_3|^{2l-2m} |\xi_-|^{2m} \\
&\quad \times \frac{(2l)!}{2^l l!(l+m)(l-m)!} \\
&= \frac{|z|^{2k}}{k!} e^{-|z|^2}, \tag{39}
\end{aligned}$$

which corresponds to a Poisson distribution with mean $|z|^2$. Similarly, the probability of obtaining the value l on measuring L is given by

$$\begin{aligned}
p(l) &= \sum_{k,m} |\langle klm|z, \xi\rangle|^2 \\
&= \frac{|\xi|^{2l}}{l!} e^{-|\xi|^2}, \tag{40}
\end{aligned}$$

corresponding to a Poisson distribution with mean $|\xi|^2$.

Because the k values and l values are distributed in probability according to (independent) Poisson distributions, it follows that

$$\langle K^n \rangle = e^{-\gamma} \left(\gamma \frac{\partial}{\partial \gamma} \right)^n e^\gamma, \quad \gamma = |z|^2$$

and

$$\langle L^n \rangle = e^{-\gamma} \left(\gamma \frac{\partial}{\partial \gamma} \right)^n e^\gamma, \quad \gamma = |\xi|^2. \tag{41}$$

In particular,

$$\begin{aligned}
\langle L^2 \rangle &= \hbar^2 \langle L(L+1) \rangle \\
&= \hbar^2 (|\xi|^4 + 2|\xi|^2). \tag{42}
\end{aligned}$$

According to Delbourgo,¹¹ the quantity $[\langle L^2 \rangle - \langle L \rangle \cdot \langle L \rangle]^{1/2}$ provides the best measure of the uncertainty in the angular momentum of the system in a given state. In view of Eq. (38), we have in the state $|z, \xi\rangle$

$$\langle L \rangle \cdot \langle L \rangle = \hbar^2 |\xi|^4 \tag{43}$$

and hence

$$\langle L^2 \rangle - \langle L \rangle \cdot \langle L \rangle = 2\hbar^2 |\xi|^2, \tag{44}$$

a result to which we shall refer in Sec. 4.

The conditional probability of obtaining the value $m\hbar$ for L_3 , given that l has been observed for L , is given by

$$\begin{aligned}
p(m;l) &= \frac{1}{p(l)} \sum_k |\langle klm|z, \xi\rangle|^2 \\
&= \frac{|\xi_3|^{2l-2m} |\xi_-|^{2m}}{|\xi|^{2l}} \frac{(2l)!}{2^l (l+m)(l-m)!}. \tag{45}
\end{aligned}$$

Now, because $\xi^2 = 0$, we have

$$\sqrt{2}|\xi| = |\xi_+| + |\xi_-|, \tag{46}$$

which, with Eq. (45), enables us to write

$$p(m;l) = \binom{2l}{l+m} \theta^{l+m} (1-\theta)^{l-m}, \tag{47}$$

with

$$\theta = \frac{|\xi_-|}{\sqrt{2}|\xi|} = 1 - \frac{|\xi_+|}{\sqrt{2}|\xi|}, \quad 0 \leq \theta \leq 1. \tag{48}$$

Thus the m values are distributed, for a given l , in accor-

dance with a binomial distribution with mean

$$\sum_{m=-l}^l m p(m;l) = l(2\theta - 1) = l \left[\frac{|\xi_-| - |\xi_+|}{|\xi_-| + |\xi_+|} \right]. \tag{49}$$

[These results are not valid if $\xi = 0$, when we get simply $p(0;0) = 1$ and $p(m;l) = 0$ otherwise.] The unconditional probability of obtaining the value $m\hbar$ on measuring L_3 in the state $|z, \xi\rangle$ is given by

$$\begin{aligned}
p(m) &= \sum_{l=|m|}^{\infty} p(l) p(m;l) \\
&= e^{-|\xi|^2} \sum_{l=|m|}^{\infty} |\xi_-|^{l+m} |\xi_+|^{l-m} \\
&\quad \times \frac{(2l)!}{2^l l!(l+m)(l-m)!} \\
&= \left| \frac{\xi_-}{\xi_+} \right|^m \left| \frac{\xi_3^2}{2} \right|^{|m|} e^{-|\xi|^2} \sum_{n=0}^{\infty} |\xi_3|^{2n} \\
&\quad \times \frac{(2n+2|m|)!}{2^n n!(n+|m|)!(n+2|m|)!} \\
&= \left| \frac{\xi_-}{\xi_+} \right|^m \left| \frac{\xi_3^2}{2} \right|^{|m|} \\
&\quad \times \frac{e^{-|\xi|^2}}{(|m|)!} M(|m| + \frac{1}{2}, 2|m| + 1, 2|\xi_3|^2) \\
&= \left| \frac{\xi_-}{\xi_+} \right|^m \exp(-\frac{1}{2}|\xi_+|^2 - \frac{1}{2}|\xi_-|^2) \\
&\quad \times I_m(|\xi_3|^2), \tag{50}
\end{aligned}$$

where M is the confluent hypergeometric function and I_m is the modified Bessel function of order m .²⁰ Note that

$$\sum_{m=-\infty}^{\infty} x^m I_m(2y) = \exp[y(x+x^{-1})], \tag{51}$$

which ensures that $\sum_{m=-\infty}^{\infty} p(m) = 1$, as required. Note also that the result (50) can be written in the form

$$\begin{aligned}
p(m) &= |\xi_- \epsilon|^{2|m|} |\xi_+ \xi_-|^{-|m|} I_{|m|}(|\xi_+ \xi_-|) \\
&\quad \times \exp(-\frac{1}{2}|\xi_+|^2 - \frac{1}{2}|\xi_-|^2), \tag{52}
\end{aligned}$$

where ϵ is the sign of m , by using Eq. (28) and the properties of the modified Bessel functions. In this form $p(m)$ is well defined even if $\xi_+ \xi_- = 0$, because $z^{-|m|} I_{|m|}(z)$ is well defined at $z = 0$.

Let us now consider the sense in which coherent angular momentum states are minimum-uncertainty states. As the operators ν and ν^\dagger are boson operators, we know that the inequality (19a) holds in general and, from our experience with the usual coherent states, we know that in the state $|z, \xi\rangle$, this inequality becomes an equality with $(\Delta\sigma)^2 = (\Delta\tau)^2 = \frac{1}{2}\hbar$, as in Eq. (20a). Thus the states $|z, \xi\rangle$ are minimum-uncertainty states in the usual sense for the conjugate variables σ and τ .

By a simple extension of a familiar argument²³ it is easily shown that, if $\Delta\alpha$ and $\Delta\beta$ are defined as in Eq. (18), then

$$(\Delta\alpha)^2 + c^2(\Delta\beta)^2 \geq -ic \langle [\alpha_i, \beta_i] \rangle = \hbar c \langle [\lambda_i, \lambda_i^\dagger] \rangle, \tag{53}$$

for arbitrary real c , with the equality holding if and only if

$$(\alpha + ic\beta)|\psi\rangle = (\langle\alpha\rangle + ic\langle\beta\rangle)|\psi\rangle, \tag{54}$$

where $|\psi\rangle$ is the appropriate state vector. We know from the fifth of Eqs. (6) that

$$[\lambda_i, \lambda_i^\dagger] = \frac{4L+3}{2L+1}, \quad (55)$$

which is positive definite. The strongest inequality of the type (53) is therefore obtained by taking that positive value of c which makes $(\Delta\alpha)^2/c + c(\Delta\beta)^2$ a minimum, viz.

$$c = \Delta\alpha/\Delta\beta, \quad (56)$$

and the inequality then has the form

$$\Delta\alpha\Delta\beta \geq \frac{1}{2}\hbar\langle[\lambda_i, \lambda_i^\dagger]\rangle = \hbar(1 + \frac{1}{2}\langle(2L+1)^{-1}\rangle). \quad (57)$$

From Eq. (54) we see that this becomes an equality if and only if

$$[(\Delta\beta)\alpha + i(\Delta\alpha)\beta]|\psi\rangle = [(\Delta\beta)\langle\alpha\rangle + i(\Delta\alpha)\langle\beta\rangle]|\psi\rangle. \quad (58)$$

Now if $|\psi\rangle = |z, \xi\rangle$, we have

$$\begin{aligned} \langle\alpha\cdot\alpha\rangle &= \frac{1}{2}\hbar\langle(\lambda + \lambda^\dagger)\cdot(\lambda + \lambda^\dagger)\rangle \\ &= \frac{1}{2}\hbar\langle(\lambda^\dagger\cdot\lambda + \lambda\cdot\lambda^\dagger)\rangle \\ &= \frac{1}{2}\hbar\langle 2L + [\lambda_i, \lambda_i^\dagger] \rangle \\ &= \frac{1}{2}\hbar\left\langle \frac{4L^2 + 6L + 3}{2L+1} \right\rangle. \end{aligned} \quad (59)$$

The same value is obtained for $\langle\beta\cdot\beta\rangle$. Furthermore, according to Eqs. (36) and (37),

$$\begin{aligned} \langle\alpha\rangle\cdot\langle\alpha\rangle &= \hbar|\xi|^2 \\ &= \hbar\langle L \rangle, \end{aligned} \quad (60)$$

and the same value is obtained for $\langle\beta\rangle\cdot\langle\beta\rangle$. Combining Eqs. (59) and (60), we get

$$\langle\alpha\cdot\alpha\rangle - \langle\alpha\rangle\cdot\langle\alpha\rangle = \frac{1}{2}\hbar\left\langle \frac{4L+3}{2L+1} \right\rangle, \quad (61)$$

and the same value for $\langle\beta\cdot\beta\rangle - \langle\beta\rangle\cdot\langle\beta\rangle$. Thus we have

$$(\Delta\alpha)^2 = (\Delta\beta)^2 = \hbar(1 + \frac{1}{2}\langle(2L+1)^{-1}\rangle), \quad (62)$$

and the inequality (57) becomes an equality. That Eq. (58) is satisfied when $|\psi\rangle = |z, \xi\rangle$ is also now evident. Since $\Delta\alpha = \Delta\beta$, it reduces to the equation

$$\lambda|z, \xi\rangle = \langle\lambda\rangle|z, \xi\rangle, \quad (63)$$

which is satisfied because $|z, \xi\rangle$ is an eigenvector of λ .

In this sense then, the states $|z, \xi\rangle$ are minimum-uncertainty states for α and β , as well as for σ and τ . However, this is a somewhat weaker notion of minimum-uncertainty than that applying to σ and τ , in two respects. First, the inequality (57) does not place restrictions on the uncertainty products for individual components of α and β such as the product $\Delta\alpha_i\Delta\beta_j$. While α and β can be regarded in a certain sense as conjugate variables, the components α_i and β_j cannot be regarded as three pairs of independent conjugate variables, because $[\alpha_i, \alpha_j]$, $[\alpha_i, \beta_j]$ and $[\beta_i, \beta_j]$ are nonzero for $i \neq j$. Second, the right-hand side of the inequality (57) is not constant. Since $(2L+1)^{-1}$ has eigenvalues $1, \frac{1}{3}, \frac{1}{5}, \dots$, one sees that the greatest lower bound of $\langle(2L+1)^{-1}\rangle$ is 0, but also that there are no states in which this bound is attained. Thus, in addition to Eq. (57), one can say that in general

$$\Delta\alpha\Delta\beta > \hbar \quad (64)$$

and that there are *no* states of the system in which $\Delta\alpha\Delta\beta$ is minimized in an absolute sense. In the state $|z, \xi\rangle$, it is not hard to show from the result (40) that

$$\langle(2L+1)^{-1}\rangle = |\xi|^{-1} e^{-|\xi|^2} \int_0^{|\xi|} e^{y^2} dy. \quad (65)$$

What one can properly say, then, is that of all states for which $\langle(2L+1)^{-1}\rangle$ has a particular value say, A , (note that it then follows that $0 < A \leq 1$), some of the states in which $\Delta\alpha\Delta\beta$ is minimized are the states $|z, \xi\rangle$, with

$$|\xi|^{-1} e^{-|\xi|^2} \int_0^{|\xi|} e^{y^2} dy = A. \quad (66)$$

In the introduction we remarked that the states $|z, \xi\rangle$ are not minimum-uncertainty states for \mathbf{x} and \mathbf{p} . This can be seen most simply by observing²⁴ that any minimum-uncertainty state for \mathbf{x} and \mathbf{p} is an eigenvector of

$$(1 + \mu)\mathbf{a} + (1 - \mu)\mathbf{a}^\dagger \quad (67)$$

for some real $\mu > 0$. (The usual coherent states have $\mu = 1$.) It is reasonably obvious from the expressions (9) that no operator of the form (67) is diagonalized on the coherent angular momentum state $|z, \xi\rangle$ in general. (In the special case that $z = 0$, $|z, \xi\rangle$ becomes equal to one of the usual coherent states, as we saw in the introduction. Thus $|0, \xi\rangle$ is an eigenvector of \mathbf{a} .)

Let us now consider the uncertainties in position and momentum of the oscillator in the state $|z, \xi\rangle$. We note from Eq. (9) that

$$a_i = u(K, L)\lambda_i + \lambda_i^\dagger w(K, L)v, \quad (68)$$

where

$$\begin{aligned} u(K, L) &= [(2K + 2L + 3)/(2L + 3)]^{1/2}, \\ w(K, L) &= [2/(2L + 3)]^{1/2}. \end{aligned} \quad (69)$$

With the help of Eqs. (6) and (7) we then deduce that

$$\begin{aligned} a_i a_j &= u(K, L)u(K, L+1)\lambda_i\lambda_j \\ &\quad + \lambda_i^\dagger w(K, L)u(K, L+1)\lambda_j v \\ &\quad + \lambda_i^\dagger \lambda_j^\dagger w(K, L+1)w(K, L+1)v^2 \\ &\quad + \lambda_j^\dagger u(K, L+1)w(K, L+1)\lambda_i v \\ &\quad + \delta_{ij}u(K, L)w(K, L)v \\ &\quad - 2\lambda_i^\dagger u(K, L+1)w(K, L+1)(2L+3)^{-1}\lambda_j v \end{aligned} \quad (70)$$

and that

$$\begin{aligned} a_i^\dagger a_j &= \lambda_i^\dagger u(K, L)u(K, L)\lambda_j + \lambda_i^\dagger \lambda_j^\dagger u(K, L+1)w(K, L)v \\ &\quad + v^\dagger w(K, L)u(K, L+1)\lambda_i\lambda_j \\ &\quad + v^\dagger \lambda_j^\dagger w(K, L+1)w(K, L+1)\lambda_i v \\ &\quad + \delta_{ij}v^\dagger w(K, L)w(K, L)v \\ &\quad - 2v^\dagger \lambda_i^\dagger w(K, L+1)w(K, L+1)(2L+3)^{-1}\lambda_j v. \end{aligned} \quad (71)$$

Then we have, in the state $|z, \xi\rangle$,

$$\begin{aligned} \langle a_i \rangle &= \langle u(K, L) \rangle \xi_i + \langle w(K, L) \rangle \xi_i^* z \\ &= u\xi_i + w\xi_i^* z, \end{aligned} \quad (72)$$

say, and

$$\begin{aligned} \langle a_i a_j \rangle &= c_1 \xi_i \xi_j + c_2 \xi_i^* \xi_j z + c_3 \xi_i^* \xi_j^* z^2 \\ &\quad + c_4 \xi_j^* \xi_i z + c_5 \delta_{ij} z, \end{aligned} \quad (73)$$

$$\langle a_i^\dagger a_j \rangle = d_1 \zeta_i^* \zeta_j + d_2 \zeta_i^* \zeta_j^* z + d_3 \zeta_i \zeta_j z^* + d_4 \zeta_j^* \zeta_i z^* z + d_5 \delta_{ij} z^* z + d_6 \zeta_i^* \zeta_j z^* z, \quad (74)$$

where, for example,

$$c_1 = \langle u(K, L) u(K, L + 1) \rangle. \quad (75)$$

Then

$$\begin{aligned} (2M\omega/\hbar)^{1/2} \langle x_i \rangle &= \langle a_i + a_i^\dagger \rangle \\ &= \langle a_i \rangle + \langle a_i^\dagger \rangle^* \\ &= u(\zeta_i + \zeta_i^*) + w(\zeta_i^* z + \zeta_i z^*) \end{aligned} \quad (76)$$

and

$$\begin{aligned} (2M\omega/\hbar) \langle x_i x_j \rangle &= \langle (a_i + a_i^\dagger)(a_j + a_j^\dagger) \rangle \\ &= \langle a_i a_j \rangle + \langle a_i^\dagger a_j \rangle \\ &\quad + \langle a_j^\dagger a_i \rangle + \langle a_i a_j \rangle^* + \delta_{ij}. \end{aligned} \quad (77)$$

We introduce Δx , an overall measure of uncertainty in positions, by

$$\begin{aligned} \Delta x &= (\langle \mathbf{x} \cdot \mathbf{x} \rangle - \langle \mathbf{x} \rangle \cdot \langle \mathbf{x} \rangle)^{1/2} \\ &= [(\Delta x_1)^2 + (\Delta x_2)^2 + (\Delta x_3)^2]^{1/2}, \end{aligned} \quad (78)$$

and using Eqs. (73), (74), (76) and (77) we deduce that

$$\begin{aligned} (2M\omega/\hbar)(\Delta x)^2 &= (c_2 + c_4 - 2uw)|\zeta|^2(z + z^*) \\ &\quad + 2(d_1 - u^2)|\zeta|^2 \\ &\quad + 2(d_4 + d_6 - w^2)|\zeta|^2|z|^2 \\ &\quad + 3c_5(z + z^*) + 6d_5|z|^2 + 3. \end{aligned} \quad (79)$$

In a similar way, we deduce that

$$(2/M\omega\hbar)^{1/2} \langle p_i \rangle = -iu(\zeta_i - \zeta_i^*) - iw(\zeta_i^* z - \zeta_i z^*), \quad (80)$$

$$\begin{aligned} (2/M\omega\hbar)(\Delta p)^2 &= (2uw - c_2 - c_4)|\zeta|^2(z + z^*) \\ &\quad + 2(d_1 - u^2)|\zeta|^2 \\ &\quad + 2(d_4 + d_6 - w^2)|\zeta|^2|z|^2 \\ &\quad - 3c_5(z + z^*) + 6d_5|z|^2 + 3. \end{aligned} \quad (81)$$

We shall make further reference to these results in Sec. 4. They are not particularly revealing as they stand, but they do make it obvious that $|z, \zeta\rangle$ is not in general one of the usual coherent states, for which one always has

$$\begin{aligned} \Delta x &= (3\hbar/2M\omega)^{1/2}, \\ \Delta p &= (3M\omega\hbar/2)^{1/2}. \end{aligned} \quad (82)$$

4. QUASICLASSICAL BEHAVIOR AND THE CLASSICAL LIMIT

Classically, one may define the state of the oscillator at any time by giving the values of the classical variables $\hat{\mathbf{x}}$ and $\hat{\mathbf{p}}$, or equivalently, by giving the value of the complex variable $\hat{\mathbf{a}}$, which is the classical counterpart of the usual lowering operator \mathbf{a}

$$\hat{\mathbf{a}} = (2M\omega)^{-1/2}(i\hat{\mathbf{p}} + M\omega\hat{\mathbf{x}}). \quad (83)$$

Alternatively, as shown in BL, one may give the values of the complex variables $\hat{\nu}$ and $\hat{\lambda}$ (with $\hat{\lambda} \cdot \hat{\lambda} = 0$), which are the classical counterparts of ν and λ . For if one knows the values of $\hat{\nu}$ and $\hat{\lambda}$, one can calculate that of $\hat{\mathbf{a}}$, the vice versa [cf Eqs. (9)]. One can think of $\hat{\nu}$ as the coordinate of a point in the complex plane \mathbb{C} , and of $\hat{\lambda}$ as the coordinates of a point on the complex cone \mathbb{K}_3 , whose equation is $\hat{\lambda} \cdot \hat{\lambda} = 0$. Then the space $\mathbb{C} \times \mathbb{K}_3$ can be regarded as a sort of complex phase space for

the oscillator, with $(\hat{\nu}, \hat{\lambda})$ being the coordinates of the representative point for the state of the system. As time t varies, this point moves in accordance with the classical equations of motion, with

$$\begin{aligned} \hat{\nu}(t) &= \hat{\nu}(0)e^{-2i\omega t}, \\ \hat{\lambda}(t) &= \hat{\lambda}(0)e^{-i\omega t}. \end{aligned} \quad (84)$$

In the quantum mechanics problem, we have

$$\begin{aligned} [H, \lambda] &= -\hbar\omega\lambda, \\ [H, \nu] &= -2\hbar\omega\nu. \end{aligned} \quad (85)$$

If $|z_0, \zeta_0\rangle$ is the state vector of the system at $t = 0$, then in the Schrödinger picture the state vector at time t is

$$|\psi(t)\rangle = e^{-iHt/\hbar}|z_0, \zeta_0\rangle, \quad (86)$$

and from Eqs. (85) we deduce that

$$\begin{aligned} \nu|\psi(t)\rangle &= z_0 e^{-2i\omega t} |\psi(t)\rangle, \\ \lambda|\psi(t)\rangle &= \zeta_0 e^{-i\omega t} |\psi(t)\rangle. \end{aligned} \quad (87)$$

From the fact that H has the value $\hbar\omega(2k + l + \frac{3}{2})$ on $|klm\rangle$, we readily deduce from Eqs. (86) and (12) that, in fact,

$$|\psi(t)\rangle = e^{-3i\omega t/2} |z(t), \zeta(t)\rangle, \quad (88)$$

with

$$z(t) = z_0 e^{-2i\omega t}, \quad \zeta(t) = \zeta_0 e^{-i\omega t}. \quad (89)$$

We see that if the system is in a coherent angular momentum state at one time, it is so at all times.

The expectation values of ν and λ as functions of time are now given, according to Eqs. (35) and (89), by

$$\begin{aligned} \langle \nu \rangle(t) &= z(t), \\ \langle \lambda \rangle(t) &= \zeta(t), \end{aligned} \quad (90)$$

and we see by comparing Eqs. (84) and (89) that these expectation values are solutions of the classical equation of motion. Classical and quantum-mechanical descriptions which correspond are obtained by taking

$$\begin{aligned} \hat{\nu}(t) &= \sqrt{\hbar} z(t), \\ \hat{\lambda}(t) &= \sqrt{\hbar} \zeta(t). \end{aligned} \quad (91)$$

The factors of $\sqrt{\hbar}$ appear here because of a difference of a factor of $\sqrt{\hbar}$ in the definitions of classical and quantum-mechanical variables like $\hat{\mathbf{a}}$ in Eq. (83) and \mathbf{a} in Eq. (3). (The quantum-mechanical variables are dimensionless; the classical ones are not.)

We see also from Eqs. (20) that the values of $\Delta\sigma$, $\Delta\tau$, $\Delta\alpha$, and $\Delta\beta$ remain constant during the motion, with the products $\Delta\sigma\Delta\tau$ and $\Delta\alpha\Delta\beta$ at their minimum values. (In the case of $\Delta\alpha\Delta\beta$, this minimum value is $\hbar(1 + \frac{1}{2}\langle(2L + 1)^{-1}\rangle)$, which remains constant because $(2L + 1)^{-1}$ is a constant of the motion.)

We may say that if the system is evolving through a succession of coherent states, as in Eq. (88), its state at time t may be defined approximately by specifying a point $(\sqrt{\hbar} z(t), \sqrt{\hbar} \zeta(t))$ in the classical phase space $\mathbb{C} \times \mathbb{K}_3$. However, there is a "volume of uncertainty" of size $\approx \Delta\sigma\Delta\tau = \frac{1}{2}\hbar$ associated with the position of $\sqrt{\hbar} z$ in \mathbb{C} , and a volume of uncertainty defined by $\Delta\alpha\Delta\beta = \hbar(1 + \frac{1}{2}\langle(2L + 1)^{-1}\rangle)$ associated with the posi-

tion of $\sqrt{\hbar} \xi$ in \mathbb{K}_3 . This representative point follows a classical trajectory, and these volumes of uncertainty do not change with time. In this sense the coherent angular momentum states are justifiably called quasiclassical states of the oscillator. Consider a typical classical trajectory, with

$$\begin{aligned}\hat{\mathbf{x}} &= (A \cos \omega t, B \sin \omega t, 0), \quad A \geq B \geq 0 \\ \hat{\mathbf{p}} &= M\omega(-A \sin \omega t, B \cos \omega t, 0), \\ \hat{\mathbf{a}} &= (\frac{1}{2} M\omega)^{1/2} e^{-i\omega t} (A, iB, 0),\end{aligned}\quad (92)$$

or equivalently, (see BL, Sec. 4)

$$\begin{aligned}\hat{\nu} &= \frac{1}{2} (M\omega)^{1/2} (A - B) e^{-2i\omega t}, \\ \hat{\lambda} &= (\frac{1}{2} M\omega AB)^{1/2} e^{-i\omega t} (1, i, 0).\end{aligned}\quad (93)$$

The description in terms of the coherent angular momentum states corresponding to this classical trajectory is provided by taking the state vector at time t to be as in Eqs. (88) and (89), with

$$\begin{aligned}z_0 &= \frac{1}{2} (M\omega/\hbar)^{1/2} (A - B), \\ \zeta_0 &= (M\omega AB / 2\hbar)^{1/2} (1, i, 0).\end{aligned}\quad (94)$$

According to Eq. (76), in this state the expectation value of \mathbf{x} in particular is given by

$$\begin{aligned}\langle \mathbf{x} \rangle(t) &= u \sqrt{AB} (\cos \omega t, \sin \omega t, 0) \\ &+ \frac{1}{2} \omega (A - B) (M\omega AB / \hbar)^{1/2} (\cos \omega t, -\sin \omega t, 0) \\ &= (A' \cos \omega t, B' \sin \omega t, 0).\end{aligned}\quad (95)$$

Therefore $\langle \mathbf{x} \rangle$ follows an elliptical path which is in the same plane, with the same center and the same orientation as the elliptical path followed by $\hat{\mathbf{x}}$, but which has different sized axes. The ratios A'/A and B'/B involve the expectation values u and w , which are constants of the motion, but which are not simply functions of A and B . One can show from Eqs. (69) and (37) that as $|z_0|$ and $|\zeta_0|$ are increased, u tends towards $\frac{1}{2}(A+B)/\sqrt{AB}$ and w tends towards $(\hbar/M\omega AB)^{1/2}$, so that A' and B' tend to A and B , respectively, as the classical limit is approached (see below).

From the expression (79) we see that, because c_2, c_4, u etc. are constants of the motion, as are $|\zeta|^2$ and $|z|^2$, and because

$$z + z^* = (z_0 + z_0^*) \cos 2\omega t - i(z_0 - z_0^*) \sin 2\omega t,$$

the value of $(\Delta x)^2$ makes bounded oscillations with angular frequency 2ω about a fixed mean value. A similar remark applies to $(\Delta p)^2$. In the coordinate representation (or the momentum representation), the wavefunction must therefore pulsate while it only approximately follows the classical motion, but it does not disperse.²⁵

The reader should not hasten to conclude that the description corresponding to the classical motion, as provided by the coherent angular momentum states, is in any sense "less quasi-classical" than the description provided by the usual coherent states. In the latter case one would take the state vector to be (up to a phase factor) $|z(t)\rangle$ [cf. Eq. (4)], where

$$z(t) = (M\omega/2\hbar)^{1/2} e^{-i\omega t} (A, iB, 0)\quad (96)$$

for the particular motion described above. Then, as is well-known, $\langle \mathbf{x} \rangle$ and $\langle \mathbf{p} \rangle$ reproduce the behavior of $\hat{\mathbf{x}}$ and $\hat{\mathbf{p}}$ and,

moreover, Δx_i and Δp_i are constants, with $\Delta x_1 \Delta p_1 = \frac{1}{2} \hbar$, etc. However, it is evident from the definition in Eqs. (9) that $\langle \nu \rangle$ and $\langle \lambda \rangle$ will now not exactly reproduce the behavior of $\hat{\nu}$ and $\hat{\lambda}$ and $\Delta \sigma, \Delta \tau, \Delta \alpha$, and $\Delta \beta$ will presumably now be oscillatory.

(In the special case of a circular orbit, the two descriptions become the same. We have $A = B$ above and thus $z_0 = 0 = z(t)$. Because $\nu|0, \xi\rangle = 0$ implies $K|0, \xi\rangle = 0$, we deduce that $u = 1$ in this situation and so Eq. (95) reduces to $\langle \mathbf{x} \rangle = \hat{\mathbf{x}}$. More generally, corresponding to any circular orbit, we have $z(t) = 0$, $\mathbf{z}(t) \cdot \mathbf{z}(t) = 0$, and $\mathbf{z}(t) = \xi(t)$; and as explained in the Introduction, the coherent angular momentum state $|0, \xi\rangle$ is equal to the usual coherent state $|z = \xi\rangle$.)

The most important distinction between the two corresponding quasi-classical descriptions is brought out by a consideration of the uncertainty in the angular momentum. If the system is in the state $|z\rangle$, we readily deduce that

$$\langle \mathbf{L} \rangle^2 - \langle \mathbf{L} \rangle \cdot \langle \mathbf{L} \rangle = 2\hbar^2 \mathbf{z}^* \cdot \mathbf{z},\quad (97)$$

after noting that

$$\begin{aligned}L_i &= -i\hbar \epsilon_{ijk} a_j^\dagger a_k, \\ \mathbf{L}^2 &= \hbar^2 [a_j^\dagger a_k^\dagger a_j a_k + 2\mathbf{a}^\dagger \cdot \mathbf{a} - (\mathbf{a}^\dagger \cdot \mathbf{a}^\dagger)(\mathbf{a} \cdot \mathbf{a})].\end{aligned}\quad (98)$$

On the other hand, if the system is in the $|z, \xi\rangle$ state, we have the result (44). Corresponding to the particular classical trajectory described above, at time t we have z and ξ as in Eqs. (89) and (94), and \mathbf{z} as in Eq. (96). For the description provided by the coherent angular momentum states, we then have

$$\langle \mathbf{L} \rangle^2 - \langle \mathbf{L} \rangle \cdot \langle \mathbf{L} \rangle = 2\hbar M\omega AB\quad (99)$$

at all times, while for the description using the usual coherent states we obtain

$$\langle \mathbf{L} \rangle^2 - \langle \mathbf{L} \rangle \cdot \langle \mathbf{L} \rangle = \hbar M\omega (A^2 + B^2).\quad (100)$$

The latter is greater, by an amount $\hbar M\omega (A - B)^2$.

In the case of a circular orbit ($A = B$), the results agree, as they must in view of our earlier remarks, but in the general case of an elliptical orbit we see that the uncertainty in the angular momentum is greater in the usual quasi-classical description.

We conclude this section with some brief comments on the classical limit. In the usual treatment this is reached by considering the system in a succession of states $|z\rangle$, with $|z| \rightarrow \infty$, $\hbar \rightarrow 0$, and $(\sqrt{\hbar})z$ finite [and equal to $\sqrt{(M\omega/2)} \times e^{-i\omega t} (A, iB, 0)$ for the particular orbit described above]. In a similar way, we can approach the limit very simply by considering a succession of states $|z, \xi\rangle$, with $|z| \rightarrow \infty$, $|\xi| \rightarrow \infty$, $\hbar \rightarrow 0$, and $(\sqrt{\hbar})z$ and $(\sqrt{\hbar})\xi$ finite and equal to $\frac{1}{2} \sqrt{(M\omega)(A - B)} e^{-2i\omega t}$ and $\sqrt{(M\omega AB/2)} e^{-i\omega t} (1, i, 0)$ in the particular orbit]. Note that the case of a circular orbit is special and corresponds always to $z = 0$. It is evident from Eqs. (20) and (36) that as the limit is approached,

$$\begin{aligned}\frac{\Delta \sigma}{\langle \sigma \rangle} &\rightarrow 0, \quad \frac{\Delta \tau}{\langle \tau \rangle} \rightarrow 0, \\ \frac{\Delta \alpha}{(\langle \alpha \rangle \cdot \langle \alpha \rangle)^{1/2}} &\rightarrow 0, \quad \frac{\Delta \beta}{(\langle \beta \rangle \cdot \langle \beta \rangle)^{1/2}} \rightarrow 0,\end{aligned}\quad (101)$$

and also, from Eqs. (43) and (44), that

$$\frac{\langle \mathbf{L}^2 \rangle - \langle \mathbf{L} \rangle \cdot \langle \mathbf{L} \rangle}{\langle \mathbf{L} \rangle \cdot \langle \mathbf{L} \rangle} \rightarrow 0. \quad (102)$$

Thus the relative widths of the probability distributions go to zero for all these variables, and one can easily see from Eqs. (37) that the same is true for K , L , and H .

5. COMPLETENESS AND A HILBERT SPACE OF ANALYTIC FUNCTIONS

In this section we find it more convenient to work with the unnormalized vectors

$$\begin{aligned} |z, \xi\rangle &= \exp\left\{\frac{1}{2}|z|^2 + \frac{1}{2}|\xi|^2\right\} |z^*, \xi^*\rangle \\ &= \sum_{klm} a_{klm}(z^*, \xi^*) |klm\rangle, \end{aligned} \quad (103)$$

rather than with the $|z, \xi\rangle$ themselves. According to Eq. (33) we then have

$$\langle z', \xi' | z, \xi \rangle = \exp(z'z^* + \xi'\xi^*). \quad (104)$$

The coefficients a_{klm} appearing in Eq. (103) were defined in Eq. (13).

We first note that

$$\int d\mu(z, \xi) a_{klm}(z^*, \xi^*)^* a_{k'l'm'}(z^*, \xi^*) = \delta_{kk'} \delta_{ll'} \delta_{mm'}. \quad (105)$$

(A derivation appears in Appendix B.) In this equation

$$d\mu(z, \xi) = \frac{2}{\pi^3} d^2z d^2\xi \delta(\xi \cdot \xi) (2|\xi|^2 - 1) \exp(-|z|^2 - |\xi|^2), \quad (106)$$

and the integration is over all possible complex z and ξ . Of course, the coefficients a_{klm} and vectors $|z, \xi\rangle$ have only been defined for (z, ξ) on $\mathbb{C} \times \mathbb{K}_3$, but that is all that is needed in integrals like that in Eq. (105) and those below, because of the delta function in $d\mu$. The meaning of the notation is as in Ref. 26: If $z = x + iy$, $\xi = u + iv$, where x, y, u and v are real, then

$$d^2z d^2\xi \delta(\xi \cdot \xi) = dx dy d^3u d^3v \delta(u^2 - v^2) \delta(2\mathbf{u} \cdot \mathbf{v}). \quad (107)$$

It follows that

$$d\mu(z, \xi) = d\mu(z^*, \xi^*). \quad (108)$$

Now consider

$$\begin{aligned} &\int d\mu(z, \xi) |z, \xi\rangle \langle z, \xi| \\ &= \sum_{klm} \sum_{k'l'm'} \int d\mu a_{klm}(z^*, \xi^*)^* a_{k'l'm'}(z^*, \xi^*) |k'l'm'\rangle \langle klm| \\ &= \sum_{klm} |klm\rangle \langle klm|, \end{aligned} \quad (109)$$

using Eq. (105). As the vectors $|klm\rangle$ are complete and orthonormal, we have the result

$$\int d\mu(z, \xi) |z, \xi\rangle \langle z, \xi| = I, \quad (110)$$

expressing the completeness of the vectors $|z, \xi\rangle$ (and hence of the vectors $|z, \xi\rangle$).

They are, in fact, overcomplete. For example, we can deduce from Eq. (110) an expression of linear dependence:

$$|z', \xi'\rangle = \int d\mu(z, \xi) |z, \xi\rangle \langle z', \xi' | z, \xi \rangle, \quad (111)$$

with $\langle z', \xi' | z, \xi \rangle$ as in Eq. (104).

Now consider an arbitrary vector $|\phi\rangle$ and write

$$|\phi\rangle = \sum_{klm} \phi_{klm} |klm\rangle, \quad (112)$$

with

$$\phi_{klm} = \langle klm | \phi \rangle. \quad (113)$$

Using Eq. (103), we see that

$$\begin{aligned} (z, \xi | \phi) &= \sum_{klm} \phi_{klm} a_{klm}(z^*, \xi^*)^* \\ &= \sum_{klm} \phi_{klm} \left[\frac{(2l)!}{k! 2^l l! (l-m)! (l+m)!} \right]^{1/2} \\ &\quad \times z^k (\xi_3)^{l-|m|} (-\epsilon \xi_\epsilon)^{|m|} \\ &= \phi(z, \xi), \end{aligned} \quad (114)$$

say. Because $(z, \xi | \phi)$ is finite for all (z, ξ) on $\mathbb{C} \times \mathbb{K}_3$, it follows that the series in Eq. (114) converges for all (z, ξ) on $\mathbb{C} \times \mathbb{K}_3$ and that $\phi(z, \xi)$ is analytic there. Noting from Eq. (104) that

$$\langle z, \xi | z, \xi \rangle = \exp(|z|^2 + |\xi|^2),$$

we have from Eq. (114) and Schwartz's inequality, that

$$|\phi(z, \xi)| \leq A \exp\left\{\frac{1}{2}|z|^2 + \frac{1}{2}|\xi|^2\right\}, \quad (115)$$

with $A = \langle \phi | \phi \rangle^{1/2}$. Thus the growth of ϕ with $|z|$ and $|\xi|$ is limited. We note also from Eq. (110) that

$$\begin{aligned} \langle \phi | \phi \rangle &= \int d\mu(z, \xi) \langle \phi | z, \xi \rangle \langle z, \xi | \phi \rangle \\ &= \int d\mu(z, \xi) |\phi(z, \xi)|^2, \end{aligned} \quad (116)$$

so that, in addition to the inequality (115), ϕ satisfies

$$\int d\mu(z, \xi) |\phi(z, \xi)|^2 < \infty. \quad (117)$$

We see in this way that any vector $|\phi\rangle$ defines a function $\phi(z, \xi)$, analytic on $\mathbb{C} \times \mathbb{K}_3$, and satisfying there the conditions (115) and (117). From Eqs. (110) and (114) we have

$$|\phi\rangle = \int d\mu(z, \xi) \phi(z, \xi) |z, \xi\rangle. \quad (118)$$

Conversely, suppose that a function $\phi(z, \xi)$ of this type is given. Then we can define a vector $|\phi\rangle$ by Eq. (118) and check that it is normalizable and that $\phi(z, \xi) = \langle z, \xi | \phi \rangle$. To do this we first note, from Eqs. (103) and (110), that

$$\langle z', \xi' | klm \rangle = \int d\mu(z, \xi) \langle z', \xi' | z, \xi \rangle \langle z, \xi | klm \rangle,$$

i.e.,

$$a_{klm}(z'^*, \xi'^*)^* = \int d\mu(z, \xi) \exp(z'z^* + \xi'\xi^*) a_{klm}(z^*, \xi^*)^*. \quad (119)$$

Since ϕ is by assumption analytic, it can be expanded as a convergent series on $\mathbb{C} \times \mathbb{K}_3$:

$$\phi(z, \xi) = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} d_{k, l, i_1, \dots, i_l} z^k \xi_{i_1} \xi_{i_2} \dots \xi_{i_l}, \quad (120)$$

and one can use Eq. (28) to bring this expansion to the form

$$\phi(z, \zeta) = \sum_{klm} \phi_{klm} a_{klm}(z^*, \zeta^*)^* \quad (121)$$

for suitable coefficients ϕ_{klm} . Since ϕ satisfies conditions (115) and (117), one can employ term by term integration to deduce from Eqs. (119) and (121) that

$$\int d\mu(z, \zeta) \exp(z'z^* + \zeta'\zeta^*) \phi(z, \zeta) = \phi(z', \zeta'). \quad (122)$$

Now we have from the definition (118) of $|\phi\rangle$ that, using Eqs. (104) and (122),

$$\begin{aligned} \langle \phi | \phi \rangle &= \int d\mu(z, \zeta) \int d\mu(z', \zeta') \phi(z, \zeta) \phi(z', \zeta')^* \langle z', \zeta' | z, \zeta \rangle \\ &= \int d\mu(z', \zeta') |\phi(z', \zeta')|^2. \end{aligned} \quad (123)$$

Therefore $|\phi\rangle$ is normalizable. Furthermore, we can now deduce from Eq. (118), with the help of Eq. (122), that

$$\begin{aligned} \langle z', \zeta' | \phi \rangle &= \int d\mu(z, \zeta) \phi(z, \zeta) \langle z', \zeta' | z, \zeta \rangle \\ &= \phi(z', \zeta'), \end{aligned} \quad (124)$$

as claimed above.

We see that there is a 1-1 correspondence between vectors in our abstract Hilbert space and functions $\phi(z, \zeta)$ of the type described, and that this correspondence is defined by Eqs. (118) and (124). Supposing $|\phi\rangle$ and $|\psi\rangle$ are any two vectors corresponding to functions $\phi(z, \zeta)$ and $\psi(z, \zeta)$ in this way; then, using Eq. (122), we have from Eq. (118) and the corresponding equation for $|\psi\rangle$ that,

$$\begin{aligned} \langle \psi | \phi \rangle &= \int d\mu(z, \zeta) \int d\mu(z', \zeta') \psi(z', \zeta')^* \phi(z, \zeta) \langle z', \zeta' | z, \zeta \rangle \\ &= \int d\mu(z', \zeta') \psi(z', \zeta')^* \phi(z', \zeta'). \end{aligned} \quad (125)$$

We are now in a position to establish a realization of the abstract Hilbert space and algebra of operators (essentially, a $(\nu^\dagger, \lambda^\dagger)$ -representation) by taking $\phi(z, \zeta)$ as the representative of $|\phi\rangle$ in a Hilbert space \mathcal{H} of such functions with scalar product

$$\langle \phi, \psi \rangle = \int d\mu(z, \zeta) \phi(z, \zeta) \psi(z, \zeta)^*, \quad (126)$$

so that

$$\langle \phi, \psi \rangle = \langle \phi, \psi \rangle. \quad (127)$$

We see from Eq. (110) that (for $|\phi\rangle$ in the domain of ν^\dagger)

$$\begin{aligned} \nu^\dagger |\phi\rangle &= \int d\mu(z, \zeta) (z, \zeta | \nu^\dagger |\phi\rangle | z, \zeta) \\ &= \int d\mu(z, \zeta) z \phi(z, \zeta) | z, \zeta \rangle, \end{aligned} \quad (128)$$

so that in \mathcal{H} , ν^\dagger is represented by the operator which sends $\phi(z, \zeta)$ into $z\phi(z, \zeta)$. Similarly, λ^\dagger is represented by the operator which sends $\phi(z, \zeta)$ into $\zeta\phi(z, \zeta)$. From our experience with the usual coherent states we know that the representative in \mathcal{H} of ν is $\partial/\partial z$, but the representative of λ is more difficult to determine. We note first that, according to Eqs. (103) and (124), the representative in \mathcal{H} of $|klm\rangle$ is the function $a_{klm}(z^*, \zeta^*)^*$, which is homogeneous of degree k in z and homogeneous of degree l in the variables ζ_j . Equation

(105) expresses the orthonormality of these functions with respect to the scalar product (126) and of course, as the representatives of the $|klm\rangle$, they must be complete in \mathcal{H} . In particular, we note that the vacuum vector $|0\rangle$ ($= |k=0, l=0, m=0\rangle$) is represented by the function 1 and, accordingly, the vector

$$(\nu^\dagger)^k \lambda^\dagger_\alpha \lambda^\dagger_\beta \dots \lambda^\dagger_\tau |0\rangle, \quad (129)$$

where there are l subscripts $\alpha, \beta, \dots, \tau$, is represented by

$$T_{\alpha\beta\dots\tau}^{kl}(z, \zeta) = z^k \zeta_\alpha \zeta_\beta \dots \zeta_\tau. \quad (130)$$

These functions also form a complete set in \mathcal{H} , when k and l run over the nonnegative integers independently and $\alpha, \beta, \dots, \tau$ run over 1, 2, 3 independently, and they also represent eigenvectors of K and L corresponding to eigenvalues k and l . Note that $T_{\alpha\beta\dots\tau}^{kl}$ is completely symmetric and traceless in the subscripts. Since it is evident that

$$\begin{aligned} z \frac{\partial}{\partial z} T_{\alpha\beta\dots\tau}^{kl} &= k T_{\alpha\beta\dots\tau}^{kl}, \\ \zeta_\alpha \frac{\partial}{\partial \zeta_\alpha} T_{\alpha\beta\dots\tau}^{kl} &= l T_{\alpha\beta\dots\tau}^{kl}, \end{aligned} \quad (131)$$

we may conclude that, because of the completeness of these functions, K is represented by $z\partial/\partial z$ (a result already evident since we know that $K = \nu^\dagger \nu$) and that L is represented by $\zeta_\alpha \partial/\partial \zeta_\alpha$. Furthermore, we know from Eq. (56) of BL that the representative of λ_i must send $T_{\alpha\beta\gamma\dots\sigma}^{kl}$ into

$$\begin{aligned} &(\delta_{i\alpha} T_{\beta\gamma\dots\sigma}^{kl-1} + \delta_{i\beta} T_{\alpha\gamma\dots\sigma}^{kl-1} + \dots + \delta_{i\tau} T_{\alpha\beta\gamma\dots\sigma}^{kl-1}) \\ &- \frac{2}{(2l-1)} (\delta_{\alpha\beta} T_{i\gamma\dots\sigma}^{kl-1} + \delta_{\alpha\gamma} T_{i\beta\dots\sigma}^{kl-1} + \dots + \delta_{\alpha\tau} T_{i\beta\gamma\dots\sigma}^{kl-1} \\ &\quad + \delta_{\beta\gamma} T_{i\alpha\dots\sigma}^{kl-1} + \dots + \delta_{\beta\tau} T_{i\alpha\gamma\dots\sigma}^{kl-1} \\ &\quad + \dots \\ &\quad + \delta_{\sigma\tau} T_{i\alpha\beta\gamma\dots}^{kl-1}) \end{aligned} \quad (132)$$

which enables us to deduce that $(2L+1)\lambda_i$ is represented by the operator

$$\left(2\zeta_\alpha \frac{\partial}{\partial \zeta_\alpha} + 1 \right) \frac{\partial}{\partial \zeta_i} - \zeta_i \frac{\partial^2}{\partial \zeta_\alpha \partial \zeta_\alpha}. \quad (133)$$

One can check directly that this operator is hermitian conjugate to $\zeta_i(2\zeta_\alpha \partial/\partial \zeta_\alpha + 1)$ [the representative of $\lambda_i^\dagger(2L+1)$] with respect to the scalar product (126), but it is not a straightforward matter because of the delta function in $d\mu$. One needs to use some results on differentiation of the delta function of a complex variable, as described in Ref. (26). Formally, the operator conjugate to ζ_i is now seen to be $\partial/\partial \zeta_i - (2\zeta_\alpha \partial/\partial \zeta_\alpha + 1)^{-1} \zeta_i \partial^2/\partial \zeta_\alpha \partial \zeta_\alpha$, which is not a differential operator on \mathcal{H} . However, $L_i (= i\hbar \epsilon_{ijk} \lambda_j^\dagger \lambda_k)$ is seen to be represented by $-i\hbar \epsilon_{ijk} \zeta_j \partial/\partial \zeta_k$.

Summarizing, the realization of our abstract algebra is provided in a Hilbert space \mathcal{H} of analytic functions $\phi(z, \zeta)$ on $\mathbb{C} \times \mathbb{K}_3$, satisfying

$$|\phi(z, \zeta)| \leq A \left\{ \exp\left(\frac{1}{2}|z|^2 + \frac{1}{2}|\zeta|^2\right) \right\}$$

and

$$\int d\mu(z, \xi) |\phi(z, \xi)|^2 < \infty. \quad (134)$$

The scalar product in \mathcal{H} is

$$(\phi, \psi) = \int d\mu(z, \xi) \phi(z, \xi)^* \psi(z, \xi), \quad (135)$$

and the relevant operators are represented as

$$\begin{aligned} v &= \frac{\partial}{\partial z}, \quad v^\dagger = z, \quad K = z \frac{\partial}{\partial z}, \\ (2L + 1)\lambda &= \left(2\xi \frac{\partial}{\partial \xi} + 1 \right) \frac{\partial}{\partial \xi} - \xi \frac{\partial^2}{\partial \xi \partial \xi}, \\ \lambda^\dagger &= \xi, \quad L = \xi \frac{\partial}{\partial \xi}, \\ L_i &= -i\hbar \epsilon_{ijk} \xi_j \frac{\partial}{\partial \xi_k}. \end{aligned} \quad (136)$$

The common eigenvalues of K , L , and L_3 are represented as

$$|klm\rangle = \left[\frac{(2l)!}{k! 2^l l! (l+m)! (l-m)!} \right]^{1/2} \times z^k (\xi_3)^{l-|m|} (-\epsilon \xi_\epsilon)^{|m|}. \quad (137)$$

This Hilbert space has a reproducing kernel given by

$$K(z', \xi'; z, \xi) = \exp(z'^* z + \xi'^* \xi), \quad (138)$$

since Eqs. (122) and (126) together give

$$(K(z', \xi'; \cdot), \phi) = \phi(z', \xi'). \quad (139)$$

The function $K(z', \xi'; z, \xi)$ can be seen to be the representative in \mathcal{H} of $|z', \xi'\rangle$. The space \mathcal{H} is especially attractive as a carrier space for $SO(3)$. If we restrict our attention to functions $f(\xi)$ analytic on the cone \mathbb{K}_3 (in effect, we consider those ϕ in \mathcal{H} satisfying $K\phi = 0$), then we have a Hilbert subspace \mathcal{H}_0 with scalar product

$$(f_1, f_2) = \frac{2}{\pi^2} \int d^6 \xi \delta(\xi \cdot \xi) (2|\xi|^2 - 1) \exp(-|\xi|^2) \times f_1(\xi)^* f_2(\xi), \quad (140)$$

carrying a reducible unitary representation of $SO(3)$, with hermitian generators

$$\hbar^{-1} L_i = -i \epsilon_{ijk} \xi_j \frac{\partial}{\partial \xi_k}. \quad (141)$$

If we label as (l) the $(2l + 1)$ -dimensional irreducible representation of $SO(3)$, we see that

$$\mathcal{H}_0 = \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \dots, \quad (142)$$

where \mathcal{H}_{0l} is $(2l + 1)$ -dimensional and carries the representation (l) . A basis for \mathcal{H}_{0l} is provided by the orthonormal functions

$$|0lm\rangle = \left[\frac{(2l)!}{2^l l! (l+m)! (l-m)!} \right]^{1/2} (\xi_3)^{l-|m|} (-\epsilon \xi_\epsilon)^{|m|} \quad (143)$$

of Eq. (137); or alternatively, by the $(2l + 1)$ linearly independent elements of the traceless, symmetric rank- l tensor

$$T_{\alpha\beta\dots\tau}^{0l}(\xi) = \xi_\alpha \xi_\beta \dots \xi_\tau. \quad (144)$$

The decomposition (142) merely symbolizes the expansion of any of the analytic functions f in \mathcal{H}_0 in series form:

$$f(\xi) = \sum_{l=0}^{\infty} d^{l\alpha\beta\dots\tau} \xi_\alpha \xi_\beta \dots \xi_\tau. \quad (145)$$

This space \mathcal{H}_0 , which can be compared with the Bargmann space²⁷ for $SU(2)$, can be said to provide a realization of the "modified boson" structure set up by Lohe and Hurst.²⁸ They introduced operators satisfying the same algebraic relations as our λ and λ^\dagger , but nothing corresponding to our v and v^\dagger . Recently Bargmann and Todorov²⁹ have described a very similar carrier space for $SO(3)$. They also consider analytic functions $f(\xi)$ on the cone \mathbb{K}_3 , but choose a more complicated scalar product, and consequently find a more complicated reproducing kernel. In their space, however, the operator conjugate to ξ_i is simply $(\xi \cdot \partial / \partial \xi + \frac{1}{2}) \partial / \partial \xi_i - \frac{1}{2} \xi_i \partial^2 / \partial \xi \partial \xi$, and these two vector operators, together with our L and $\hbar^{-1} L$, generate a unitary representation of $SO(3,2)$. Their space was not derived from a consideration of coherent states of any kind, but we have mentioned in BL that there is an $so(3,2)$ algebra associated with the oscillator, spanned by our operators L , $\hbar^{-1} L$, Λ , and Λ^\dagger , where

$$\Lambda = (2L + 1)^{1/2} \lambda, \quad \Lambda^\dagger = \lambda^\dagger (2L + 1)^{1/2}. \quad (146)$$

Had we chosen to define coherent angular momentum states by diagonalizing the operators Λ rather than λ , we would have arrived at the Hilbert space of Bargmann and Todorov. The coherent states so obtained would evidently be generalized coherent states for $so(3,2)$ (in particular) in the sense of Barut and Girardello¹⁷ and Perelomov.¹⁸ We have already given in BL some reasons for our preference for the operators λ and λ^\dagger . The main point is that with our definitions, the expectation values of H , K , L , and L in the coherent angular momentum states have simple properties—more simple than if we were to follow the alternative path. In particular, the nice property that in a coherent angular momentum state the l values are distributed in probability according to a Poisson distribution would be lost if we were to diagonalize the Λ_i .

6. CONCLUDING REMARKS

The existence of a second set of coherent or quasiclassical states for the oscillator places the usual ones in a new perspective. The two sets share several interesting properties, as we have seen, and we therefore hope to find interesting applications of our new states. In particular, we hope to be able to construct a new diagonal representation of the density matrix for the oscillator.

The usual coherent states are also quasiclassical states for the Hamiltonian

$$H' = H + \alpha L_3$$

corresponding to a charged oscillator in a uniform magnetic field. The same is true for coherent angular momentum states, since the diagonalized operators v , λ_3 , and λ_\pm , like a_3 and $a_\pm (= a_1 \pm ia_2)$, are all shift operators for H' . The new states are also quasi-classical states for a Hamiltonian of the form

$$H' = H + \alpha L_3 + \beta L,$$

which is not an exact Hamiltonian for any physical system

but may be of interest in the approximate description of some molecular spectra. In that connection, we may also expect the coherent angular momentum states to be of value in the analysis of Hamiltonians of the form

$$H' = H + \alpha L_3 + \beta L + \gamma L^2,$$

although they will not then define states which are quasiclassical in the same sense as above.

The analysis we have developed above and in BL can no doubt be extended to $n > 3$ dimensions. The Hilbert space for the isotropic oscillator in n dimensions carries only symmetric representations of $so(n)$, labelled by a single nonnegative integer l . Accordingly, the first step in the generalization would be the introduction of a scalar operator L which has nonnegative integer eigenvalues l . One would then proceed to resolve n -vector boson operators \mathbf{a} , \mathbf{a}^\dagger into shift operators for L . There appear to be various interesting alternative paths to follow from that point, corresponding to various chains of orthogonal subgroups of $so(n)$.

In the case $n = 2$, the boson operators a_\pm are shift operators for the $so(2)$ scalar L_3 , and accordingly, the usual coherent states and the coherent angular momentum states may be identified.

We recall that the eigenvalue problem for the three-dimensional isotropic oscillator can also be solved by separation of variables in a cylindrical-polar coordinate system.

Are there then "cylindrical" coherent states, as well as "Cartesian" and "spherical" ones? The answer is "yes," but they can be taken to be the usual (Cartesian) ones. (Diagonalize a_3 and a_\pm , which are shift operators for N_3 and L_3 .)

With regard to the definition of coherent angular momentum states for other Hamiltonians of the form

$$H = \mathbf{p}^2/2M + V(|\mathbf{x}|),$$

it is clear that one cannot, in general, hope to obtain quasiclassical states in the sense described above. The existence of this property for the oscillator depends on the special feature that the total angular momentum quantum number l appears linearly in the formula for the energy eigenvalue, when expressed in terms of l and the radial quantum number k :

$$E_{kl} = \hbar\omega(2k + l + \frac{1}{2}).$$

However, in the more general case one can still diagonalize vector lowering operators for L in order to define overcomplete sets of states for the "angular part" of the Hilbert space, and one may be able to use these to construct representations in which the angular part of the density matrix is diagonal.

ACKNOWLEDGMENT

We thank Professor L. Bass for several helpful discussions.

APPENDIX A

The combination of Eqs. (12) and (14) with the expanded form of $Y_{lm}(\theta, \phi)$ and property (28) gives

$$|z, \xi\rangle = \exp(-\frac{1}{2}|z|^2 - \frac{1}{2}|\xi|^2) \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \sum_{m=-l}^l z^k (\xi_3)^{l-m} (-\xi_-)^m \left[\frac{(2l)!}{k!2^l l!(l+m)!(l-m)!} \right]^{1/2} (-1)^k \times \left[\frac{2a^3 k!}{\Gamma(k+l+\frac{3}{2})} \right]^{1/2} \xi^l e^{-\frac{1}{2}\xi^2} L_k^{(l+1)}(\xi^2) (-1)^m \left[\frac{(2l+1)(l-m)!}{4\pi(l+m)!} \right]^{1/2} P_l^m(\cos\theta) e^{im\phi} \quad (\text{A1})$$

$$= \exp(-\frac{1}{2}|z|^2 - \frac{1}{2}|\xi|^2 - \frac{1}{2}\xi^2) \times \left[\frac{a^3}{\pi^{3/2}} \right]^{1/2} \times \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \left[\frac{\Gamma(l+\frac{3}{2})}{\Gamma(k+l+\frac{3}{2})} \right]^{1/2} (-z)^k L_k^{(l+1)}(\xi^2) \times \sum_{m=-l}^l \frac{1}{(l+m)!} (\sqrt{2}\xi)^l (\xi_3)^{l-m} (\xi_-)^m P_l^m(\cos\theta) e^{im\phi}, \quad (\text{A2})$$

using the result that

$$\Gamma(l+\frac{3}{2}) = \frac{(2l+1)! \sqrt{\pi}}{2^{2l+1} l!}. \quad (\text{A3})$$

We are able to evaluate the sum over m by noting that for any vector \mathbf{v} , with spherical polar coordinates v, θ, ϕ ,

$$\mathbf{v} \cdot \xi = v \xi_3 \left\{ \frac{1}{2} \left[\frac{\xi_-}{\xi_3} e^{i\phi} + \frac{\xi_+}{\xi_3} e^{-i\phi} \right] \sin\theta + \cos\theta \right\} = v \xi_3 \left\{ \frac{1}{2} [A - 1/A] \sin\theta + \cos\theta \right\}, \quad (\text{A4})$$

where $A = \xi_- e^{i\phi} / \xi_3$. We can further show, by induction, that

$$\left\{ \frac{1}{2} [A - 1/A] \sin\theta + \cos\theta \right\}^l = l! \sum_{m=-l}^l A^m P_l^m(\cos\theta) / (l+m)!, \quad (\text{A5})$$

from which it follows that

$$\frac{(\mathbf{v} \cdot \xi)^l}{l!} = \sum_{m=-l}^l \frac{1}{(l+m)!} v^l (\xi_3)^{l-m} (\xi_-)^m P_l^m(\cos\theta) e^{im\phi}. \quad (\text{A6})$$

With the use of this result in the case $\mathbf{v} = (\sqrt{2})\mathbf{ax}$, Eq. (A2) becomes

$$|z, \xi\rangle = \exp(-\frac{1}{2}|z|^2 - \frac{1}{2}|\xi|^2 - \frac{1}{2}\xi^2) \times \left[\frac{a}{\sqrt{\pi}} \right]^{3/2} \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \left[\frac{\Gamma(l+\frac{3}{2})}{\Gamma(k+l+\frac{3}{2})} \right]^{1/2} (-z)^k L_k^{(l+1/2)}(\xi^2) \frac{(\sqrt{2}\mathbf{ax} \cdot \xi)^l}{l!}, \quad (\text{A7})$$

as in Eq. (15).

When $z = 0$ only the $k = 0$ term contributes, so that

$$|0, \xi\rangle = \exp\left(-\frac{1}{2}|\xi|^2 - \frac{1}{2}a^2|\mathbf{x}|^2\right) \left[\frac{a}{\sqrt{\pi}}\right]^{3/2} \sum_{l=0}^{\infty} \frac{(\sqrt{2ax \cdot \xi})^l}{l!}, \quad (\text{A8})$$

which is equivalent to Eq. (16).

APPENDIX B

Using the definitions (13) and (106) of a_{klm} and $d\mu$, respectively, and property (28), we can write

$$\begin{aligned} & \int d\mu(z, \xi) a_{klm}(z^*, \xi^*)^* a_{KLM}(z^*, \xi^*) \\ &= \frac{2}{\pi^3} \left[\frac{(2l)!(2L)!}{k!K!l!L!2^{l+L}(l+m)!(L+M)!(l-m)!(L-M)!} \right]^{1/2} \int d^2z \exp(-|z|^2) z^k (z^*)^K \\ & \quad \times \int d^6\xi (2|\xi|^2 - 1) \exp(-|\xi|^2) \delta(\xi \cdot \xi) (\xi_3)^{l-m} (\xi_3^*)^{L-M} (-\xi_+)^m (-\xi_-^*)^M. \end{aligned} \quad (\text{B1})$$

It is straightforward to show that

$$\int d^2z \exp(-|z|^2) z^k (z^*)^K = \pi k! \delta_{k,K}. \quad (\text{B2})$$

However, the next integration is not so easy. Letting $\xi = \mathbf{x} + i\mathbf{y}$, where \mathbf{x} and \mathbf{y} are the real vectors, the delta function becomes [see Ref. (26)]

$$\delta(\xi \cdot \xi) = \delta(|\mathbf{x}|^2 - |\mathbf{y}|^2) \delta(2\mathbf{x} \cdot \mathbf{y}), \quad (\text{B3})$$

and the integral becomes

$$\begin{aligned} & \int d^6\xi (2|\xi|^2 - 1) \exp(-|\xi|^2) \delta(\xi \cdot \xi) (\xi_3)^{l-m} (\xi_3^*)^{L-M} (-\xi_+)^m (-\xi_-^*)^M \\ &= \int d^3x d^3y \exp(-|\mathbf{x}|^2 - |\mathbf{y}|^2) \delta(|\mathbf{x}|^2 - |\mathbf{y}|^2) \delta(2\mathbf{x} \cdot \mathbf{y}) (2|\mathbf{x}|^2 + 2|\mathbf{y}|^2 - 1) \\ & \quad \times (-1)^{m+M} (x_3 + iy_3)^{l-m} (x_3 - iy_3)^{L-M} [x_1 - y_2 + i(x_2 + y_1)]^m [x_1 - y_2 - i(x_2 + y_1)]^M. \end{aligned} \quad (\text{B4})$$

We can introduce the new variables r, θ, ϕ, R, ψ , and u by the following:

$$[x_1, x_2, x_3] = [r \sin\theta \cos\phi, r \sin\theta \sin\phi, r \cos\theta]$$

and

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \cos\theta \cos\phi & -\sin\phi & \sin\theta \cos\phi \\ \cos\theta \sin\phi & \cos\phi & \sin\theta \sin\phi \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} R \cos\psi \\ R \sin\psi \\ u \end{bmatrix}. \quad (\text{B5})$$

The matrix is orthogonal, with determinant equal to 1. It is easily inverted to give

$$u = (\mathbf{x} \cdot \mathbf{y}) / r. \quad (\text{B6})$$

Now

$$\delta(2\mathbf{x} \cdot \mathbf{y}) = \delta(2ru) = \frac{1}{2r} \delta(u), \quad (\text{B7})$$

and the integral (B4) becomes, after the u integration

$$\begin{aligned} & \frac{1}{2} \int r dr \sin\theta d\theta d\phi R dR d\psi \exp(-r^2 - R^2) \delta(r^2 - R^2) (2r^2 + 2R^2 - 1) (-1)^{m+M} (r \cos\theta - iR \sin\theta \cos\psi)^{l-m} \\ & \quad \times (r \cos\theta + iR \sin\theta \cos\psi)^{L-M} [r \sin\theta \cos\phi - R \cos\theta \sin\phi \cos\psi - R \cos\phi \sin\psi \\ & \quad + i(r \sin\theta \sin\phi + R \cos\theta \cos\phi \cos\psi - R \sin\phi \sin\psi)]^m [r \sin\theta \cos\phi - R \cos\theta \sin\phi \cos\psi - R \cos\phi \sin\psi \\ & \quad - i(r \sin\theta \sin\phi + R \cos\theta \cos\phi \cos\psi - R \sin\phi \sin\psi)]^M. \end{aligned} \quad (\text{B8})$$

Since

$$\delta(r^2 - R^2) = \frac{1}{r+R} \delta(r-R) \quad (\text{B9})$$

here, we can perform the R integration and obtain

$$\frac{1}{4} \int r dr \sin\theta d\theta d\phi d\psi \exp(-2r^2)r^{l+L}(4r^2-1)(-1)^{m+M}(\cos\theta - i \sin\theta \cos\psi)^{l-m}(\cos\theta + i \sin\theta \cos\psi)^{L-M} \\ \times [\sin\theta - \sin\psi + i \cos\theta \cos\psi]^m [\sin\theta - \sin\psi - i \cos\theta \cos\psi]^M e^{i(m-M)\phi} \quad (\text{B10})$$

$$= \frac{1}{2} \pi \delta_{m,M} \int r dr \sin\theta d\theta d\psi \exp(-2r^2)r^{l+L}(4r^2-1)(\cos\theta - i \sin\theta \cos\psi)^l (\cos\theta + i \sin\theta \cos\psi)^L \\ \times (1 + \sin\theta \sin\psi)^{-m} (1 - \sin\theta \sin\psi)^m. \quad (\text{B11})$$

Here we make yet another change of variables from θ, ψ to α, β , where

$$[\sin\alpha \cos\beta, \sin\alpha \sin\beta, \cos\alpha] = [\cos\theta, \sin\theta \cos\psi, \sin\theta \sin\psi] \quad (\text{B12})$$

and the integral becomes

$$\frac{1}{2} \pi \delta_{m,M} \int r dr \sin\alpha d\alpha d\beta \exp(-2r^2)r^{l+L}(4r^2-1)(1 + \cos\alpha)^{-m}(1 - \cos\alpha)^m (\sin\alpha)^{l+L} e^{i(L-l)\beta} \quad (\text{B13})$$

$$= \pi^2 \delta_{m,M} \delta_{l,L} \int r dr \exp(-2r^2)r^{2l}(4r^2-1) \int \sin\alpha d\alpha (1 + \cos\alpha)^{l-m}(1 - \cos\alpha)^{l+m} \quad (\text{B14})$$

$$= \frac{\pi^2}{2} \frac{l! 2^l (l-m)! (l+m)!}{(2l)!} \delta_{l,L} \delta_{m,M}, \quad (\text{B15})$$

using the properties of the Gamma and Beta functions. Substitution of (B15) and (B2) into (B1) gives the required result, Eq. (105).

¹E. Schrödinger, *Naturwiss.* **14**, 664 (1926).

²J.R. Klauder, *Ann. Phys. (N.Y.)* **11**, 123 (1960).

³R.J. Glauber, *Phys. Rev.* **131**, 2766 (1963).

⁴J.R. Klauder and E.C.G. Sudarshan, *Fundamentals of Quantum Optics* (Benjamin, New York, 1968), Chap. 7.

⁵H.M. Nussenzveig, *Introduction to Quantum Optics* (Gordon and Breach, New York, 1973), Chap. 3.

⁶P. Bonifacio, D.M. Kim, and M.O. Scully, *Phys. Rev.* **187**, 441 (1969).

⁷P.W. Atkins and J.C. Dobson, *Proc. R. Soc. London, Ser. A* **321**, 321 (1971).

⁸D. Bhaumik, T. Nag, and B. Dutta-Roy, *J. Phys. A: Math. Gen.* **8**, 1868 (1975).

⁹T.M. Makhviladze and L.A. Shelepin, in *Proceedings (Trudy) of the P.N. Lebedev Physics Institute*, Vol. 70, edited by D.V. Skobel'tsyn. English translation (Consultants Bureau, New York, 1975).

¹⁰J. Mostowski, *Phys. Lett. A* **56**, 369 (1976).

¹¹R. Delbourgo, *J. Phys. A: Math. Nucl. Gen.* **10**, 1837 (1977).

¹²T.S. Santhanam preprint, Australian National University, 1978.

¹³J.M. Radcliffe, *J. Phys. A: Gen. Phys.* **4**, 313 (1971).

¹⁴J. Schwinger, in *Quantum Theory of Angular Momentum*, edited by L.C. Biedenharn and H. van Dam (Academic, New York, 1965).

¹⁵V.V. Mikhailov, *Izv. Akad. Nauk SSSR, Ser. Fiz.* **37**, 2230 (1973). English

Translation, *Bull. Acad. Sci. USSR Phys. Ser.* **37**, No. 10, 187 (1974).

¹⁶A.J. Bracken and H.I. Leemon, *J. Math. Phys.* (to appear).

¹⁷A.O. Barut and L. Girardello, *Commun. Math. Phys.* **21**, 41 (1971).

¹⁸A.M. Perelomov, *Commun. Math. Phys.* **26**, 2221 (1976).

¹⁹J.L. Powell and B. Crasemann, *Quantum Mechanics* (Addison-Wesley, Reading, 1961), Secs. 7.4, 7.5.

²⁰M. Abramowitz and I.A. Stegun (eds.), *Handbook of Mathematical Functions* (Dover, New York, 1965).

²¹A.R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton U.P., Princeton, N.J., 1960).

²²P. Stehle, *Quantum Mechanics* (Holden-Day, San Francisco, 1966), p. 51.

²³Ref. 22, p. 21.

²⁴Ref. 22, p. 55.

²⁵The fact that the wavefunction pulsates is not in itself surprising; every wavefunction for the oscillator is periodic in time.

²⁶I.M. Gel'fand and G.E. Shilov, *Generalized Functions* (Academic, New York, 1964), Vol. 1, Sec. B2.1.

²⁷V. Bargmann, *Commun. Pure Appl. Math.* **14**, 187 (1961); *Rev. Mod. Phys.* **34**, 829 (1962).

²⁸M.A. Lohe and C.A. Hurst, *J. Math. Phys.* **12**, 1882 (1971).

²⁹V. Bargmann and I.T. Todorov, *J. Math. Phys.* **18**, 1141 (1977).

Iteration of single- and two-channel Schrödinger equations^{a)}

R. Müller and H. J. W. Müller-Kirsten

Department of Physics, University of Kaiserslautern, 6750 Kaiserslautern, West Germany

(Received 4 November 1980; accepted for publication 14 November 1980)

A general perturbation technique is developed for the iteration of one- and two-channel Schrödinger equations for potentials, which can be expanded around a minimum. The channel coupling is assumed to be weak. In order to facilitate the numerical or algebraic computer calculation of terms of higher order, a recurrence relation is derived for some particularly important coefficients. The eigenvalues and oscillatorlike solutions are then derived explicitly up to and including the third-order iteration. Furthermore, we demonstrate that certain parts of every n th order iteration can be lumped together in a manner which is independent of the specific form of the potential. Finally, these methods are applied to the calculation of the eigenvalues of the one-channel equation for linear and logarithmic potentials with or without a weak Coulomb contribution.

PACS numbers: 03.65.Ge, 02.30.Hq

1. INTRODUCTION

Recently Dashen *et al.*¹ have given a detailed investigation of multichannel potential scattering with at least one permanently confined channel.

The success of nonrelativistic models in reproducing the observed mass spectrum of heavy quark-antiquark states led also to a revival of interest²⁻⁵ in the multichannel formalism.^{6,7} The problem can be formulated in such a way that its extension of single-channel theory is particularly transparent.¹ Nevertheless, the single-channel potential theory normally leads to broad widths; however, narrow widths can be generated by the weak coupling to a second channel.⁷ It is therefore plausible to investigate the two-channel problem defined by the transitions such as $c\bar{c} - D\bar{D}^* - D\bar{D}^*$, where c is the charmed quark and D, D^* are $c\bar{u}$ or $c\bar{d}$ bound states. For arbitrary potentials we have to use perturbation methods, which require in higher order the evaluation of a large number of expressions. The present investigation was also motivated by the desire to give the iteration in third-order explicitly, for a certain class of potentials.

In the following we describe an iteration procedure, which has already been applied to a large number of other problems (see Ref. 8-13). With the help of REDUCE¹⁴ it is now possible to do these algebraic computations on the computer, so that the necessary expressions are obtained within a reasonable time.

In Sec. 2 we recapitulate the iteration procedure, to define names and parameters. In subsequent sections we elaborate on details relevant to our application. In Sec. 3 we give a new recurrence relation for the coefficient functions, which is also useful for numerical computations, as well as for algebraic summations. Further we demonstrate in Sec. 4 that some parts of the n th iteration can be written as terms of a geometric progression and thus can be combined.

With the iteration procedure we obtain the wavefunctions and their eigenvalues in the range of a minimum of the potential. If we represent the wavefunction in some other

way, as in Sec. 5, we can show that the representation of the wavefunction is confined to the complete set of eigenfunctions of the harmonic oscillator, although the iteration procedure seems to use a larger set. This implies that the iteration procedure is restricted to the same complete set of eigenfunctions as the usual perturbation ansatz.

In Appendices A-F we give the general result for the iteration of the eigenvalues up to third-order for an arbitrary potential, which can be expanded in a Taylor series around a minimum. This general iteration procedure is then applied to the linear and logarithmic potentials and their combination with a Coulomb potential. We calculate their s states and Regge trajectories.

2. THE ITERATION PROCEDURE

2.1. The one-dimensional Schrödinger equation

In the following we recapitulate briefly the iteration procedure developed elsewhere,⁸⁻¹³ in order to introduce names and parameters, which are then used.

We start with the s -wave equation

$$\left(-\frac{1}{2\mu} \frac{d^2}{dx^2} + V(x)\right)\psi(x) = E\psi(x), \quad (1)$$

where

$$V(x) = \sum_{i=0}^{\infty} a_i x^i. \quad (2)$$

First we expand the potential around one of its minima at $x = x_0$. Setting

$$z = x - x_0, \quad (3)$$

we have

$$V(z) = \sum_{i=0}^{\infty} b_i z^i. \quad (4)$$

Next we rewrite the equation in the form

$$\left(\frac{1}{2\mu} \frac{d^2}{dz^2} + E - b_0 - b_2 z^2\right)\psi = \left(\sum_{i=3}^{\infty} b_i z^i\right)\psi, \quad (5)$$

and make the following substitutions:

^{a)}Supported in part by the Deutsche Forschungsgemeinschaft.

$$\begin{aligned}
E - b_0 &= \bar{E}, \\
b_2 &= g^2, \\
M^2 &= 2\mu, \\
y^2 &= 2gMz^2,
\end{aligned}
\tag{6}$$

and

$$\bar{E}M/2g = \frac{1}{2}p + \Delta/2g \tag{7}$$

or

$$E = pg/M + M\Delta + b_0,$$

where pg/M is the contribution to the eigenvalue coming from the left hand side of Eq. (5), p is exactly or approximately an odd integer (depending on the boundary conditions), and Δ represents the contributions of higher order terms of the potential to the eigenvalues.

Hence

$$(-2)\left(\frac{d^2}{dy^2} + \frac{p}{2} - \frac{y^2}{4}\right)\psi(y) = \left(\sum_{i=0}^{\infty} h_i y^i\right)\psi(y), \tag{8}$$

where

$$\begin{aligned}
h_0 &= \Delta/g, \\
h_1 &= h_2 = 0, \\
h_i &= -b_i(M/g)(1/2gM)^{i/2}, \quad i > 2.
\end{aligned}
\tag{9}$$

Our differential equation now has the form

$$\mathcal{D}_p \psi = (\sum h_i y^i) \psi, \tag{10}$$

where

$$\mathcal{D}_p = (-2)\left(\frac{d^2}{dy^2} + \frac{p}{2} - \frac{y^2}{4}\right). \tag{11}$$

Given any arbitrary potential, we rewrite the Schrödinger equation in the form (10) in order to fix the coefficients h_i . Our results will be expressed in terms of h_i and p . The overall solution of Eq. (10) can be written

$$\psi = \psi^{(0)} + \psi^{(1)} + \psi^{(2)} + \dots, \tag{12}$$

where $\psi^{(i)}$ indicates the contribution of the i th iteration.

The iteration procedure starts with

$$\psi^{(0)} = \psi_p, \tag{13}$$

where ψ_p is the solution of

$$\mathcal{D}_p \psi_p = 0. \tag{14}$$

This solution is the well-known parabolic cylinder function i.e.,

$$\psi_p \equiv D_{(p-1)/2}(y) \equiv D_{-a-1/2}(y), \tag{15}$$

where $a = -p/2$. The square integrability over $0 < y < \infty$ implies

$$a = -(2n+1)/2, \quad \text{or } p = 2n+1, \quad n \in \text{IN}. \tag{16}$$

With the recurrence relation

$$y\psi_p = \psi_{p+2} + (p-1)/2\psi_{p-2} \tag{17}$$

we can express the right side of (10) as a sum over various ψ_{p+2j} . Our first step is to calculate the coefficients S_i of the relation

$$y^i \psi_p = \sum_{j=-i}^{+i} S_i(p, 2j) \psi_{p+2j}. \tag{18}$$

For $i = 1, \dots, 8$ the coefficients $S_i(p, 2j)$ are given in Appendix A.

For our subsequent iteration it is useful to rearrange the sums derived from the right-hand side of Eq. (8) in the following form:

$$\begin{aligned}
\left(\sum_i h_i y^i\right)\psi_p &= \sum_i h_i \left(\sum_{j=-i}^{+i} S_i(p, 2j)\psi_{p+2j}\right) \\
&= \sum_{j=-\infty}^{+\infty} C(p, p+2j)\psi_{p+2j}.
\end{aligned}
\tag{19}$$

The coefficients $C(p, p+2j)$ are the sums

$$C(p, p+2j) = \sum_{i=0}^{\infty} h_{2i+|j|} S_{2i+|j|}(p, 2j). \tag{20}$$

In Appendix B we give the explicit form of these coefficients for values of j ranging from 0 to ± 7 . With these definitions, we get the following rest term $R^{(0)} = \psi_p$ on the right hand side of (10),

$$R^{(0)} = \left(\sum_i h_i y^i\right)\psi_p = \sum_j C(p, p+2j)\psi_{p+2j}. \tag{21}$$

Since

$$\mathcal{D}_{p+i} = \mathcal{D}_p - i, \tag{22}$$

we get

$$\mathcal{D}_p \psi_{p+i}/i = \psi_{p+i}. \tag{23}$$

This expression implies, that we can compensate any term $\mu\psi_{p+i}$ ($i \neq 0$) on the right-hand side of (10) by adding to $\psi^{(0)}$ appropriate higher order contributions

$$\mu\psi_{p+i}/i.$$

The contribution of first order to ψ is $\psi^{(1)}$ i.e.,

$$\psi^{(1)} = \sum_{j \neq 0} C(p, p+2j)(1/2j)\psi_{p+2j}. \tag{24}$$

The part proportional to ψ_p must be zero to this order of the iteration. In this way we can determine Δ . To first order we have

$$\psi_p C(p, p) = 0. \tag{25}$$

Iterating we get from $\psi^{(1)}$ the rest term

$$\begin{aligned}
R^{(1)} &= \left(\sum_i h_i y^i\right)\psi^{(1)} \\
&= \sum_{j \neq 0} C(p, p+2j) \frac{1}{2j} \sum_i C(p+2j, p+2i+2j)\psi_{p+2i+2j},
\end{aligned}
\tag{26}$$

and then from $\psi^{(2)}$ a term $R^{(2)}$ and so on.

The quantity Δ is determined by the equation

$$0 = C(p, p) + \sum_{i \neq 0} C(p, p+2i) \frac{1}{2i} C(p+2i, p) + \dots, \tag{27}$$

and the weight w_{2j} of a certain ψ_{p+2j} in $\psi = \psi^{(0)} + \psi^{(1)} + \dots$ is given by

$$\begin{aligned}
w_{2j}(p) &= \frac{1}{2j} \left(C(p, p+2j) + \sum_{i \neq 0} C(p, p+2i) \frac{1}{2i} C(p+2i, p+2j) + \dots \right).
\end{aligned}
\tag{28}$$

The w_{2j} 's are functions of Δ and p . The iterated wavefunction is therefore

$$\psi = \psi^{(0)} + \psi^{(1)} + \psi^{(2)} + \dots = \sum_j w_{2j} \psi_{p+2j}, \quad (29)$$

where $w_0 = 1$. Further statements regarding w_{2j} are given in Sec. 5.

2.2. The two-channel equation

If we write the two-channel equation¹² in the following form, we can reduce the iteration to a procedure similar to that for the one-dimensional case. We have a system of equations with potentials which we choose as in Ref. 12, i.e., as harmonic terms ($g^2 + a_{ii}^{(2)}z^2$) plus anharmonic contributions. Then

$$\begin{bmatrix} \frac{1}{2\mu} \frac{d^2}{dz^2} + E - g^2 z^2 - V_{11}(z) & -V_{12}(z) \\ -V_{21}(z) & \frac{1}{2\mu} \frac{d^2}{dz^2} + E - g^2 z^2 - V_{22}(z) \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} = 0, \quad (30)$$

where

$$V_{ii} = \sum_{k=2}^{\infty} a_{ii}^{(k)} z^k, \quad (31)$$

and for $i \neq j$

$$V_{ij} = \sum_{k=0}^{\infty} a_{ij}^{(k)} z^k.$$

Next we make the substitutions

$$M^2 = 2\mu, \quad y^2 = 2gMz^2 \\ h_{ij}^{(k)} = -a_{ij}^{(k)}(M/g)(2gM)^{-k/2}, \quad k > 0, \quad (32)$$

and we set again

$$EM/2g = \frac{1}{2}p + \Delta/2g. \quad (33)$$

Below we distinguish between Δ_1 and Δ_2 which are the values of Δ associated with the two channels; of course, the total E is the same for both channels. Defining the matrices

$$H_k = \begin{bmatrix} h_{11}^{(k)} & h_{12}^{(k)} \\ h_{21}^{(k)} & h_{22}^{(k)} \end{bmatrix} \quad \text{for } k > 0,$$

and

$$H_0 = \begin{bmatrix} \Delta_1/g & h_{12}^{(0)} \\ h_{21}^{(0)} & \Delta_2/g \end{bmatrix}, \quad (34)$$

$$\mathcal{D}_{pp} = \begin{bmatrix} \mathcal{D}_p & \cdot \\ \cdot & \mathcal{D}_p \end{bmatrix}, \quad (35)$$

where

$$\mathcal{D}_p = (-2) \left(\frac{d^2}{dy^2} + \frac{p}{2} - \frac{y^2}{4} \right),$$

we can write our two-channel equation in the form

$$\mathcal{D}_{pp} \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} = \sum_k H_k y^k \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix}, \quad (36)$$

and we can now iterate the equation in the same way as we treated the one-dimension case.

We start with

$$\psi^{(0)} \equiv \psi_{pp} \begin{bmatrix} \psi_p \\ \psi_p \end{bmatrix} \quad (37)$$

and

$$\psi_p = D_{(p-1)/2}.$$

The iteration procedure now yields matrices and the coefficients $C(p, p+2j)$ are determined by

$$C(p, p+2j) = \sum_{i=0}^{\infty} H_{2i+|j|} S_{2i+|j|}(p, 2j). \quad (38)$$

This is the generalization of Eq. (20). It should be noted, that the S_i 's are the same in the one-dimensional case.

For the calculation of the eigenvalues, we have to solve the matrix equation

$$0 = \left[C(p, p) + \sum_{i \neq 0} C(p, p+2i) \frac{1}{2i} C(p+2i, p) + \dots \right] \begin{bmatrix} \psi_p \\ \psi_p \end{bmatrix}. \quad (39)$$

We thus obtain two equations for the eigenvalues. We have to distinguish the corrections of the two channels and we get two results.

The weight w_{2j} of the functions ψ_{p+2j} is now defined by a vector given by

$$w_{2j} \psi_{p+2j} = \left[C(p+2j) \frac{1}{2j} + \dots \right] \begin{bmatrix} \psi_{p+2j} \\ \psi_{p+2j} \end{bmatrix}. \quad (40)$$

The complete solution is then (apart from an overall factor)

$$\Psi \equiv \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} = \sum_{j=-\infty}^{\infty} w_{2j} \psi_{p+2j}. \quad (41)$$

3. GENERAL EXPRESSIONS FOR THE COEFFICIENTS $C(p, p+2j)$

We now discuss the derivation of the coefficients $C(p, p+2j)$ for either of the single- or the two-channel problems formulated above. The $S_i(p, j)$ can be calculated from the recurrence relation

$$S_m(p, j) = S_{m-1}(p, j+2)(p+j+1)/2 \\ + S_{m-1}(p, j-2), \quad (42)$$

where

$$S_0(p, 0) = 1$$

and

$$S_m(p, j) = 0 \quad \text{if } |j| > 2m.$$

But this relation is not the most suitable for computer calculations. If we wish to calculate a coefficient $C(p, p+2j)$ for a specified value of j , we need to know a large number of such coefficients. For this reason, we generalize the procedure.

It is found to be convenient to factorize S_i in the form (for $j \geq 0$)

$$S_{2n+j}(p, \pm 2j) = W(p, \pm j) \frac{1}{2^n} \binom{2n+j}{n} f_j^n(p, \pm j), \quad (43)$$

where $W(p, \pm j)$ is defined as:

$$W(p, \pm j) = \begin{cases} 1 & \text{if } +j \geq 0 \\ \frac{(p-1)(p-3)\dots(p+1-2j)}{2} & \text{if } -j < 0 \end{cases} \quad (44)$$

and the function $f_j^n(p)$ is a polynomial of order n in p . In particular we have

$$S_j(p, \pm 2j) = W(p, \pm j)$$

and

$$S_{j-2}(p, \pm 2j) = 0. \quad (45)$$

The coefficients C then become

$$C(p, p \pm 2j) = W(p, \pm j) \sum_{n=0}^{\infty} H_{2n+j} \frac{1}{2^n} \binom{2n+j}{n} f_j^n(p \pm j), \quad (46)$$

where H_{2n+j} are the coefficients defined in Sec. 2.2 (or h_{2n+j} in Sec. 2.1), derived from a knowledge of the given potentials. The calculation of $S_{2n+j}(p, \pm 2j)$ then implies the calculation of $f_j^n(p \pm j)$. We found the general expression for $f_j^n(p)$ by trial and error, i.e.,

$$f_j^n(p) = p^n + \left(\sum_{k=1}^{n-1} k^2 + j \sum_{k=1}^{n-1} k \right) p^{n-2} + \dots \quad (47)$$

or

$$f_j^n(p) = p^n + \frac{1}{6}n(n-1)(2n-1+3j)p^{n-2} + \dots$$

From the first few terms of $f_j^n(p)$ we can develop a recurrence relation in much the same way as for systems of orthogonal polynomials. But we cannot find an interval $[a, b]$ and a weight function $w(p)$, on the real axis, so that

$$\int_a^b w(p) f_j^n(p) f_j^m(p) dp \sim \delta_{nm}. \quad (48)$$

The reason is the following. In such a system, the zeros of $f_j^n(p)$ have to be within $[a, b]$. But any zero which we have calculated lies on the imaginary axis. This means that if we want to construct an orthogonal system on the real axis, we have to rotate $f_j^n(p)$ in the complex plane. Here we are not interested in these functions, although our calculations proceed parallel to the problem of handling orthogonal systems.

Orthogonal polynomials P_n possess a recurrence relation of the form

$$P_{n+1}(x) = (a_n + xb_n)P_n - c_n P_{n-1}, \quad (49)$$

where x is the variable and a_n , b_n , and c_n depend only on n . In our case the relation is

$$f_j^{n+1}(p) = p f_j^n(p) + n(n+j) f_j^{n-1}(p). \quad (50)$$

We have checked this relation for more than 70 f_j^n 's with the help of REDUCE. With the help of this recurrence relation, we can derive another recurrence relation for $S_m(p, \pm 2j)$.

Replacing the $f_j^n(p)$'s in (50) by S_m 's, with the help of (43), we obtain

$$\begin{aligned} & S_{m+2}(p, \pm 2j) \\ &= (m+1)(m+2)/(m+2-j)(m+2+j) \\ & \times [2(p \pm j)S_m(p, \pm 2j) + m(m-1)S_{m-2}(p, \pm 2j)], \end{aligned} \quad (51)$$

with the constraints

$$S_j(p, \pm 2j) = W(p, \pm j)$$

and

$$S_{j-2}(p, \pm 2j) = 0. \quad (52)$$

For particular values of the coefficients H_{2n+j} of the potential, we can regard $C(p, p \pm 2j)$ as a generating function and construct a differential equation from the recurrence re-

lation for S_m . If this differential equation is sufficiently simple, depending on the coefficients H_{2n+j} , we can integrate it and obtain for $C(p, p \pm 2j)$ an exact expression, i.e., we find from

$$F_j = \sum_n f_j^n(p) \frac{1}{x^n} \quad (x \neq 0) \quad (53)$$

and

$$f_j^{n+1} = p f_j^n + n(n+j) f_j^{n-1} \quad (54)$$

the differential equation

$$x^2 F_j'' - (1+j)x F_j' + (1+j+px-x^2)F_j = 0, \quad (55)$$

with the solution

$$F_j = x^{1+j} e^x {}_1F_1\left[\frac{1}{2}(p+j+1), 1+j, -2x\right]. \quad (56)$$

4. ALGEBRAIC ITERATION OF THE EIGENVALUES

4.1. Summation of parts of the iteration

The first description of the iteration procedure, given in Sec. 2 is useful for a numerical iteration of the wavefunction in inverse powers of the coupling constant. In algebraic calculations it is not so interesting to calculate the complete wavefunction. Our primary interest here concerns the calculation of the eigenvalues. We give the lowest order contributions of Eqs. (27), and (39) explicitly.

In calculating these contributions we find that some terms can be summed by hand. We now demonstrate how this can be done. First we classify the different contributions in certain combinations of H_i . Each combination with only one H_i is an element of the first iteration I T 1.

$$\text{I T 1} = C(p, p) = \sum_i H_i S_i(p, 0). \quad (57)$$

Any combination of a pair $H_i H_j$ belongs to the second iteration

$$\text{I T 2} = \sum_{i \neq 0} C(p, p+2i)(1/2i)C(p+2i, p), \quad (58)$$

and so on. We can write each iteration in the form

$$\begin{aligned} \text{I T 1} &= \sum_{i=0}^{\infty} S_i H_i, \\ \text{I T 2} &= \sum_i \sum_j T_{ij} H_i H_j, \\ \text{I T 3} &= \sum_i \sum_j \sum_k U_{ijk} H_i H_j H_k, \\ \text{I T 4} &= \sum_i \sum_j \sum_k \sum_l V_{ijkl} H_i H_j H_k H_l. \end{aligned} \quad (59)$$

The coefficients S_i , T_{ij} , ... are defined by comparison with the n th order iteration. They are given in Appendix C. The coefficients S_i , T_{ij} , ... are scalar functions of the quantum number p and thus commute with the matrices H_i .

From the first iteration we get

$$S_i \equiv S_i(p, 0). \quad (60)$$

The T_{ij} 's of the second iteration can be obtained as follows. The first correction to the wavefunction is

$$\psi^{(1)} = \sum_j H_j \sum_{k \neq 0} S_j(p, 2k) \frac{1}{2k} \psi_{p+2k}. \quad (61)$$

Thus the part proportional to H_j is constructed in the following way:

- (i) multiply ψ_p by y^j and use Eq. (18);
- (ii) replace ψ_{p+2k} by $(1/2k) \psi_{p+2k}$ if $k \neq 0$;
- (iii) ignore ψ_p .

The right-hand sided of Eqs. (10) and (36) now operate on $\psi^{(1)}$ from the left. This right hand side contains only one term, which is proportional to H_i . Hence we get the combination

$$H_i y^j \left(H_j \sum_{k \neq 0} S_j(p, 2k) \frac{1}{2k} \psi_{p+2k} \right) = H_i H_j \left(\sum_{k \neq 0} S_j(p, 2k) \frac{1}{2k} \sum_l S_i(p+2k, 2l) \psi_{p+2k+2l} \right). \quad (63)$$

The contribution to the second iteration of the eigenvalue is the term proportional to ψ_p ; we call this term T_{ij} , i.e.,

$$T_{ij} = \sum_{k=-j, k \neq 0}^{+j} S_j(p, 2k) \frac{1}{2k} S_i(p+2k, -2k). \quad (64)$$

In the same way we obtain corresponding expressions for U_{ijk} , V_{ijkl} , ... i.e.,

$$U_{ijk} = \sum_{n_0=-k, n_0 \neq 0}^{+k} S_k(p, 2n_0) \times \sum_{n_1=-j, n_0+n_1 \neq 0}^{+j} S_j(p+2n_0, 2n_1) \frac{1}{4n_0(n_0+n_1)} S_i[p+2(n_0+n_1), -2(n_0+n_1)]. \quad (65)$$

The coefficient function of a combination

$$H_i (H_0)^n H_k \quad (66)$$

of one of the next iterations can be related to Eq. (64). We demonstrate this for $n=1$ with the help of Eq. (65).

The combination $H_i H_0 H_k$ belongs to the coefficient function U_{i0k} . In (65) we set $j=0$. With $S_0(p+2n_0, 0) = 1$, we see that the sum over n_1 reduces to a factor $1/2n_0$. We obtain therefore,

$$U_{i0k} = \sum_{n_0=-k, n_0 \neq 0}^{+k} S_k(p, 2n_0) \left(\frac{1}{2n_0} \right)^2 S_i(p+2n_0, -2n_0). \quad (67)$$

By the same reduction we obtain a factor $1/2n_0$ for each H_0 in the coefficient function of

$$H_i (H_0)^n H_k. \quad (68)$$

The complete contribution of (68) to the iteration of the eigenvalues is therefore

$$H_i (H_0)^n H_k \cdot \sum_{n_0=-k, n_0 \neq 0}^{+k} S_k(p, 2n_0) \left(\frac{1}{2n_0} \right)^{n+1} S_i(p+2n_0, -2n_0). \quad (69)$$

Now we insert $(H_0)^n$ into the sum. The number n varies from zero to infinity and $(H_0/2n_0)^n$ represents a term of a geometric progression. Summing this part of our iteration gives

$$\left(\frac{1}{2n_0} \right) \frac{1}{1 - H_0/2n_0} = \frac{1}{2n_0 - H_0} \quad (70)$$

within (69). In Sec. 6 we calculate H_0 explicitly and we will see that H_0 is proportional to $1/g$. For a sufficiently large coupling constant g , we have $|H_0| < 1$. We now define

$$\hat{T}_{ij} = \sum_{n_0=-j, n_0 \neq 0}^{+j} S_j(p, 2n_0) \frac{1}{2n_0 - H_0} S_i(p+2n_0, -2n_0), \quad (71)$$

and \hat{U}_{ijk} , \hat{V}_{ijkl} , ... correspondingly, with the condition, that each of these quantities is zero, if one of the indices is zero.

Our iteration of the eigenvalues is now obtained from:

$$0 = \sum_i S_i H_i + \sum_i \sum_j H_i \hat{T}_{ij} H_j + \dots \quad (72)$$

\hat{T}_{ij} is a matrix, so we have to be careful about its position in the above sums. For a triple combination $H_i H_j H_k$, two of the substitutions (70) are required and we have to distinguish between factors $1/(2k - H_0)$ arising from the first and second substitutions.

Summing over all combinations $H_i H_j$ means that for a specific combination $H_i H_j$, the complete contribution is

$$H_i T_{ij} H_j + H_j T_{ji} H_i. \quad (73)$$

In the one-dimensional case we get

$$h_i T_{ij} h_j + h_j T_{ji} h_i = 2h_i h_j T_{ij}, \quad (74)$$

where $T_{ij} = T_{ji}$. The coefficients \hat{T}_{ij} are given in detail in Appendix D.

4.2. The choice of the order of $1/g$

In our iteration of the eigenvalues we ignore all contributions which are of order higher than $n+1$ in $1/g$, where g is the coupling constant. According to our definitions [Eqs. (9), (32)], the coefficients H_i are of the following order in $1/g$:

$$H_i \sim (1/g)^{1+i/2}. \quad (75)$$

The number of possible combinations of H_i appropriate to a particular power $(1/g)^n$ grows rapidly with n . Many of these combinations contain H_1 or H_2 . Because of their low order in $1/g$, they lead to many combinations of H_1 , H_2 , and H_i which contribute to the iteration of the eigenvalue. In the one-dimensional case, it can be arranged that h_1 and h_2 are zero. In the two-channel problem H_1 and H_2 are zero only in special cases, and then we have fewer terms which determine H_0 (i.e., Δ). In Table I we give the number of terms which contribute to the n th order iteration of $H_0 = \Delta/g$, neglecting terms of order $(1/g)^{n+1}$. In column Ia we give the number of coefficients S_i , T_{ij} , ... if $H_1 \neq 0$ and $H_2 \neq 0$. Column Ib gives the number of these coefficients if $H_1 = H_2 = 0$. Columns IIa, and IIb give the number of the corresponding coefficients \hat{T}_{ij} , ...

From Table I we conclude that it is reasonable in the case of coefficients in column Ia to iterate up to $O(1/g^6)$; this has also the advantage, that our equation for H_0 remains linear. E.g., to $O(1/g^6)$ we have the equation: (in the one-dimensional case with $h_1 = h_2 = 0$)

$$0 = h_0(1 + h_3^2 U_{303}) + h_4 S_4 + h_6 S_6 + h_8 S_8 + h_{10} S_{10} + T_{33} h_3^2 + T_{44} h_4^2 + 2T_{35} h_3 h_5 + O(1/g^6), \quad (76)$$

or with Eq. (9)

TABLE I. This table gives the number of terms which contribute to Eq. (27) if terms of $O(1/g^{n+1})$ are neglected. Column Ia gives the number of terms if we use the coefficient functions T_{ij}, U_{ijk}, \dots . Column Ib gives the same as Ia but with $h_1 = h_2 = 0$. Column IIa gives the number of terms if we use $\hat{T}_{ij}, \hat{U}_{ijk}, \dots$ and IIb the same for $h_1 = h_2 = 0$.

n	Ia	Ib	IIa	IIb
2	2	1	2	1
3	4	2	4	2
4	9	3	8	3
5	22	5	17	5
6	67	16	37	9
7	139	19	73	17
8	296	39	141	26

$$\begin{aligned} \Delta \left(1 + \frac{1}{g^5} \frac{b_3^2 U_{303}}{8M} \right) &= \frac{1}{g^2} b_4 S_4 + \frac{1}{g^3} S_6 + \frac{1}{g^4} \left(\frac{b_8 S_8}{2^4 M^3} - \frac{b_3^2 T_{33}}{8M} \right) \\ &+ \frac{1}{g^5} \left(\frac{b_{10} S_{10}}{2^5 M^4} - \frac{b_4^2 T_{44}}{2^4 M^2} - \frac{b_3 b_5 T_{35}}{2^3 M^2} \right) \\ &+ O\left(\frac{1}{g^6}\right). \end{aligned} \quad (77)$$

5. REPRESENTATION OF ψ_{p+j} IN TERMS OF ψ_p AND ψ'_p

We have also calculated the first few terms of the wavefunction algebraically. Here we give an alternative representation of the wavefunction and then discuss the results for the coefficients w_{2j} of

$$\psi = \sum_{j=-\infty}^{\infty} w_{2j} \psi_{p+2j}. \quad (78)$$

We can rewrite each ψ_{p+2j} in terms of ψ_p and $\psi'_p = d\psi/dy$. For this reason we define two functions $C_1(y,p)$ and $C_2(y,p)$ with the property

$$\sum_{j=-\infty}^{\infty} w_{2j}(p) \psi_{p+2j}(y) = C_1(y,p) \psi_p + C_2(y,p) \psi'_p, \quad (79)$$

where $w_{2j}(p)$ are the coefficients determined earlier [Eq. (28)] and contained in (78).

We obtain our second representation from the recurrence relations

$$\psi_{p+2} = y\psi_p - \frac{1}{2}(p-1)\psi_{p-2} \quad (80)$$

and

$$\psi_{p+2} = \frac{1}{2}y\psi_p - \psi'_p. \quad (81)$$

We now set

$$\psi_{p+2j} = \frac{1}{W(p,j)} [F_j(y,p)\psi_p + G_j(y,p)\psi'_p], \quad (82)$$

where $W(p,j)$ is again the function defined by (44), and F_j, G_j are defined by (82). An explicit list of F_j, G_j is given in Appendix E. From the recurrence relations for F_j and G_j , to be derived below, we see that F_j and G_j are polynomials in y and p . With a different choice of the function $W(p,j)$ in the denominator of (82), we would get complicated recurrence relations with rational expressions. From (82) and (80) we can deduce recurrence relations for F_j and G_j . We find

$$\frac{F_{j+1}}{W(p,j+1)} = \frac{yF_j}{W(p,j)} - \frac{p-1+2j}{2W(p,j-1)} F_{j-1}$$

and

$$\frac{G_{j+1}}{W(p,j+1)} = \frac{yG_j}{W(p,j)} - \frac{p-1+2j}{2W(p,j-1)} G_{j-1}. \quad (83)$$

We subdivide these recurrence relations into two sets

- (i) if $j > 0$ and $W(p,j) = 1$,
- (ii) if $j < 0$ and $W(p,j) \neq 1$.

If $j > 0$, we get

$$F_{j+1} = yF_j - \frac{1}{2}(p-1+2j)F_{j-1},$$

with

$$F_0 = 1, \quad F_1 = y/2 \quad (85)$$

and

$$G_{j+1} = yG_j - \frac{1}{2}(p-1+2j)G_{j-1},$$

with

$$G_0 = 0, \quad G_1 = -1. \quad (86)$$

For negative j it is more practical to redefine the index j so that we get the index $-|j+1|$ on the left-hand side of the recurrence relations.

With these substitutions, we have

$$F_{-(j+1)} = yF_{-j} - \frac{1}{2}(p+1-2j)F_{-(j-1)} \quad (87)$$

and

$$G_{-(j+1)} = yG_{-j} - \frac{1}{2}(p+1-2j)G_{-(j-1)}, \quad (88)$$

where $j > 0$ and the boundary values are

$$\begin{aligned} F_0 &= 1, \quad F_{-1} = y/2, \\ G_0 &= 0, \quad G_{-1} = 1. \end{aligned} \quad (89)$$

The functions $C_1(y,p)$ and $C_2(y,p)$ of (79) are now defined by

$$\begin{aligned} C_1(y,p) &= \sum_{j=-\infty}^{+\infty} w_{2j} \frac{F_j(y,p)}{W(p,j)}, \\ C_2(y,p) &= \sum_{j=-\infty}^{+\infty} w_{2j} \frac{G_j(y,p)}{W(p,j)}. \end{aligned} \quad (90)$$

These expressions have an important consequence for the coefficients w_{2j} . C_1 and C_2 are well-defined functions of y . For negative values of j the function $W(p,j)$ has $|j|$ zeros at $p = 1, 3, \dots, -2j - 1$. These zeros must be compensated by $w_{2j} F_j(y,p)$ and $w_{2j} G_j(y,p)$. However, the functions F_j and G_j do not provide this cancellation for the following reasons:

(i) For negative j , F_j and G_j are polynomials in p of degree lower than $|j|$ so they cannot compensate $|j|$ zeros of $W(p,j)$.

(ii) The zeros of F_j and G_j depend also on y and we can find a value of y for which neither y nor G_j is zero at $p = 1, 3, \dots$. Consequently, w_{2j} must compensate the function $W(p,j)$, i.e.,

$$w_{2j} \sim W(p,j). \quad (91)$$

In the case of the first contribution to w_{2j} we see immediately, that it is proportional to $W(p,j)$.

$$w_{2j} = 1/2j [C(p,p+2j) + \dots]$$

and

$$C(p, p+2j) \sim W(p, j). \quad (92)$$

(see Appendix A).

The proportionality has also been verified for several other contributions derived from the next iterations. The overall factor $W(p, j)$ has no consequence for $j > 0$; hence the common factor is one. However, for negative j we have

$$\begin{aligned} w_{-2} &= \left(-\frac{1}{2}\right) \frac{p-1}{2} (H_1 + \dots), \\ w_{-4} &= \left(-\frac{1}{4}\right) \frac{p-1}{2} \frac{p-3}{2} (H_2 + \dots), \\ w_{-6} &= \left(-\frac{1}{6}\right) \frac{p-1}{2} \frac{p-3}{2} \frac{p-5}{2} (H_3 + \dots), \end{aligned} \quad (93)$$

and so on.

Here we have a nontrivial factor which implies that for a fixed number $p = 1, 3, \dots$ the weights $w_{2j}(p)$ are zero for $(-2j) > p$. The function ψ_{p+2j} with the lowest index, for which $w_{2j} \neq 0$, is always ψ_1 . This is reasonable if we compare (78) with ansatz of the usual perturbation method.¹⁶ There one starts with the ansatz

$$\phi = \sum_n c_n \phi_n, \quad (94)$$

where $\{\phi_n\}$ is the complete set of eigenfunctions of the unperturbed Hamiltonian.

For the harmonic oscillator this complete set is given by the parabolic cylinder functions ψ_p , $p = 1, 3, \dots$. The functions ψ_q with negative q do not belong to this set and the iteration procedure reduces the coefficients w_{2j} of these functions automatically to zero, leaving

$$\begin{aligned} \psi &= \sum_{j=-\infty}^{+\infty} w_{2j} \psi_{p+2j} = \sum_{j=(1-p)/2}^{\infty} w_{2j} \psi_{p+2j} \\ &= \sum_{k=1,3,5}^{\infty} w_{k-p} \psi_k. \end{aligned} \quad (95)$$

Although the denominator $W(p, j)$ in (90) is cancelled out by factors of w_{2j} , we do not get an additional contribution to the wavefunction. From (82) we get

$$W(p, y) \psi_{p+2j} = F_j \psi_p + G_j \psi'_p = 0 \quad \text{if } p < -2j. \quad (96)$$

ψ_{p+2j} is a known function for each p and j , and from $\psi_{p+2j} < \infty$ it follows that the product $W(p, j) \psi_{p+2j}$ is zero if $p < 2j$, and thus the sum $F_j \psi_p + G_j \psi'_p = 0$ if $p < -2j$.

Finally we make some remarks concerning $C_1(y, p)$ and $C_2(y, p)$. If we represent the wavefunction in terms of w_{2j} , the number and the relative magnitude of the terms grows very rapidly, so that in this way, we obtain only a crude approximation of the wavefunction.

The construction of $C_i(y, p)$ is therefore more economical if it is possible to sum up a large number of terms in the definition (90), especially for the ground state $p = 1$ and at the origin $y = 0$. The weight w_{2j} is given by the expansion ($j \geq 0$)

$$\begin{aligned} w_{2j} &= (1/2j) \{ H_j + \frac{1}{2}(2+j)(p+j)H_{j+2} \\ &\quad + \frac{1}{8}(4+j)(3+j)[(p+j)^2 + 1+j]H_{j+4} + \dots \}. \end{aligned} \quad (97)$$

If we set $p = 1$ we get for F_j and G_j the same recurrence

relation as for Hermite polynomials¹⁵; from (85) and (86),

$$F_{j+1} = yF_j - jF_{j-1}$$

and

$$G_{j+1} = yG_j - jG_{j-1}. \quad (98)$$

The generating function depends on the boundary conditions, but we can expect that the contribution of

$$\sum_{j=0}^{\infty} w_{2j} F_j(y, 1) \simeq 1 + \sum_{j=1}^{\infty} \frac{1}{2^j} H_j F_j(y, 1) \quad (99)$$

to $C_1(y, p)$ is of the same kind as the contribution of the generating function of the Hermite polynomials. Therefore, for the class of physically interesting potentials, the coefficients H_j are of the form

$$H_j \sim \alpha^j / j! \quad (100)$$

where α is a parameter, which represents the perturbation of the harmonic potential.

The generating function of the Hermite polynomials is

$$\exp(2xz - z^2) = \sum_n \frac{z^n}{n!} H_n(x). \quad (101)$$

Similarly we have

$$-\exp\left(y\alpha - \frac{\alpha^2}{2}\right) = \sum_n \frac{\alpha^n}{n!} G_{n+1}(y, p=1). \quad (102a)$$

To obtain $C_2(y, p)$, we have to compensate an additional (-1) in (102), so we get

$$C_2(y, p) = 1 - \exp(y\alpha - \alpha^2/2). \quad (102b)$$

6. QUARK-POTENTIAL MODELS

We first of all recapitulate some lowest order iterations for potentials which have been discussed in the literature.^{10,11} We then consider additional contributions and give here a more extensive numerical analysis of the results. It suffices to compare equation (10) with the corresponding equations in previous papers^{10,11} in order to fix the parameters h_i . With these h_i we then get the approximation determined by the n th iteration.

Before we start, we comment on the expansion in powers of the coupling parameter. It is true that we always combine all contributions of the same order $(1/g)^n$. But the combinations of the coefficient functions S_i, T_{ij}, \dots depend on combination of the coupling parameter with the coefficients h_i of the potential. Sometimes, i.e., in some problem, it is advantageous to substitute the Taylor coefficients a_j of the potential via $a_i = g \cdot b_i$, in order to obtain simple expressions. Thus we switch from a combination of h_j 's to powers $(1/g)^n$. In Sec. 2 we have seen that the coupling constant g and the coefficients h_i are related in the following way:

$$h_i \sim 1/g^{1+i/2}, \quad i = 0, 1, \dots \quad (103)$$

In the case of the linear potential¹¹ it was found to be convenient to divide out a factor of g^2 . The correspondence there is

$$\begin{aligned} h_0 &\sim 1/g, \\ h_i &\sim \frac{1}{g^{1+i/2}}, \quad i > 0. \end{aligned} \quad (104)$$

In the first case $h_8 S_8$ and $h_3^2 T_{33}$ are the same order in $1/g$; in the second case $h_4 S_4$ and $h_3^2 T_{33}$. If we now truncate our iteration at a certain order $(1/g)^n$, the contribution neglected in the two cases are not the same. There is no general criterion for deciding which ordering is the more suitable; this has to be decided from case to case and depends on the specific form of the potential under consideration.

We find that for our potentials the effect is not sufficiently significant in order to yield deviations which change the results considerably. However, we can take a more statistical viewpoint for our choice of relevant contributions. There is a correspondence between the power of y and the degree of the polynomials of the quantum number p arising in the coefficient functions. A high degree of p requires high powers of y . So it is perhaps more practical to sum all contributions of the same order in p . If we combine all contributions of the same order we have to use \hat{T}_{ij} , \hat{U}_{ijk} , ... and the order of p is half of the indices. The contributions of order p^n are

$$\begin{aligned} p^0 &\rightarrow S_0, \\ p^1 &\rightarrow S_2, \hat{T}_{11}, \\ p^2 &\rightarrow S_4, \hat{T}_{13}, \hat{T}_{31}, \hat{T}_{22}, \\ &\hat{U}_{112}, \hat{U}_{121}, \hat{U}_{211}, \hat{V}_{1111}. \end{aligned} \quad (105)$$

The relation ratios of the coefficient functions \hat{T}_{ij} , \hat{U}_{ijk} , ... grows very rapidly, so it is more convenient to consider all polynomials up to the degree p^n than to take all contributions of an order $(1/g)^n$.

6.1. The power potential $V(r) = r^\lambda$, $\lambda > 1$

We do not repeat each step of the calculations of Ref. 11. All we need here are the coefficients h_i and nothing else. After the transformation $r = e^z$ has been applied to the appropriate radial wave equation, we get an effective potential of the form

$$v(z) = \alpha e^{2z} - \beta e^{(2+\lambda)z} + \delta e^z, \quad (106)$$

where

$$\begin{aligned} \alpha &= 2\mu(E - V_0)/\hbar, \quad \beta = 2\mu g_1/\hbar^2, \\ \delta &= 2\mu g_0/\hbar^2, \quad \gamma = L^2 - \frac{1}{4} = l(l+1), \end{aligned} \quad (107)$$

μ is the reduced mass, g_1 is the coupling constant of r^λ , and g_0 the coupling constant of a Coulomb potential.

We write this potential in the form

$$v(z) = v(z_0) + \sum_{i=2}^{\infty} \frac{(z-z_0)^i}{i!} v^{(i)}(z_0), \quad (108)$$

where $v^{(i)}(z_0)$ is the Taylor coefficient, i.e.,

$$v^{(i)}(z_0) = \alpha 2^i e^{2z_0} - \beta (2+\lambda)^i e^{(2+\lambda)z_0} + \delta e^{z_0}, \quad (109)$$

and z_0 is determined by

$$v^{(1)}(z_0) = 0. \quad (110)$$

The coupling parameter h is

$$h = [-2v^{(2)}(z_0)]^{1/2}. \quad (111)$$

The variable is named ω and given by $\omega = h(z - z_0)$. With the differential operator (11) we can rewrite the appropriate equation in the form

$$\mathcal{D}_q \psi = \left[\frac{2\Delta}{h} - \sum_{i=3}^{\infty} \left(\frac{v^{(i)}(z_0)}{v^{(2)}(z_0)} \right) \frac{\omega^i}{i! h^{i-2}} \right] \psi. \quad (112)$$

We now have to compare (112) with (10). The relations between the relevant parameters are

$$\begin{aligned} h_0 &\equiv 2\Delta/h, \\ h_1 &= h_2 = 0, \\ h_i &\equiv \frac{v^{(i)}(z_0)}{v^{(2)}(z_0)} \frac{1}{i!} \frac{1}{h^{i-2}}, \quad i \geq 3. \end{aligned} \quad (113)$$

Setting

$$h_i = -c_i/h^{i-2}, \quad i \geq 3, \quad (114)$$

we obtain the following set of iterations for Δ :

$$\begin{aligned} \text{IT1} &= \frac{2\Delta}{h} - c_4 \frac{1}{h^2} S_4 - c_6 \frac{1}{h^4} S_6 - O\left(\frac{1}{h^6}\right), \\ \text{IT2} &= \frac{1}{h^2} c_3^2 T_{33} + \frac{1}{h^4} (c_4^2 T_{44} + 2c_3 c_5 T_{35}) + O\left(\frac{1}{h^6}\right), \\ \text{IT3} &= \frac{2\Delta}{h} \left(\frac{1}{h^2} c_3^2 U_{303} \right) - c_3^2 c_4 \frac{1}{h^4} (U_{334} + U_{343} + U_{433}) \\ &\quad + O\left(\frac{1}{h^5}\right), \\ \text{IT4} &= \left(\frac{2\Delta}{h} \right)^2 \frac{c_3^2}{h^2} V_{3003} + \frac{c_3^4}{h^4} V_{3333} + O\left(\frac{1}{h^5}\right). \end{aligned} \quad (115)$$

To this order of our iteration, we obtain an equation, which is quadratic in Δ .

In the case of the linear potential ($\lambda = 1$, $\delta = 0$) we have

$$c_i = (1/i!)(3^{i-1} - 2^{i-1}). \quad (116)$$

This gives

$$\begin{aligned} \Delta h [1 + (q/h^2)(0.395q^2 + 1.28)] + (\Delta h)^2 (1/h^4) \\ \times (0.30q^2 + 0.10) \\ = -(1/72)(51q^2 + 1) \\ - (1/h^2)(q/256)(71.71q^4 + 253.9q^2 + 4.86) \\ + O(1/h^3). \end{aligned} \quad (117)$$

We can iterate the quadratic equation and express Δh in powers in q provided

$$q(0.395q^2 + 1.28) < h^2. \quad (118)$$

Then we get

$$\begin{aligned} -\Delta h &= \frac{51q^2 + 1}{72} + \frac{q}{h^2} \frac{67q^2 + 1}{2^5 \times 3^3} \\ &\quad + \frac{1}{h^4} \left(\frac{4925}{62 \ 208} q^4 - \frac{551}{31 \ 104} q^2 - \frac{3965}{186 \ 624} \right) \\ &\quad + O\left(\frac{1}{h^6}\right). \end{aligned} \quad (119)$$

Numerical test show, that for Δh the difference between the approximation (119) and the quadratic equation (117) is less than 1% if (118) is satisfied.

Now the relation between h and the energy E is

$$h^3 \sim E^{3/2}. \quad (120)$$

This means that in calculating Δh for small values of E we have to use the quadratic equation (117). Calculating s states for the linear potential presents a further problem. The contributions of the next order $O(1/h^4)$ are already so large for

relatively small h , that the asymptotic approximation for the s states has to be truncated after terms of $O(1/h^2)$.

6.2 Iteration with higher contributions

In Sec. 4 we showed that we can sum a part of the iteration. We give here the result with the coefficient functions \hat{T}_{ij}, \dots and later we compare this with the result of Δh in Sec. 6.1. Of course, we get contributions, as explained in 6.1, so we can expect the numerical result to differ, particularly in the range of low energies. But we can also expect these results to be closer to the exact result, and for this reason we include higher contributions.

The following numerical results for Δh differ from those of Sec. 6.1 markedly only in the domain of low energies. This can be seen in the following way. In the coefficient functions \hat{T}_{ij}, \dots , we have denominators of the form $n h^2 \pm \Delta h$, where n is an integer. If we set $\Delta h \equiv 0$ we get the coefficient functions T_{ij} which were used in Sec. 6.1. So we get the same results where we can neglect the term Δh in comparison with h^2 . From Eq. (119) we know that for sufficiently large h the quantity Δh becomes constant. Thus, irrespective of our choice of the potential we can conclude that for sufficiently large values of h we have $h^2 \gg |\Delta h|$. In the range of large h , it is therefore no problem to replace the denominator by the first few terms of a geometric progression in powers of $\Delta h/nh^2$.

For small values of $h \sim 1$, we cannot make a general statement for arbitrary potentials. For the linear and loga-

arithmic potentials Δh and h^2 are of the same magnitude. We are there approaching the limit of validity for the expansion of the denominator as a geometric progression.

The difference in the numerical results arises from the difference of the denominator and its representation by a few terms of the geometric progression.

Here we quote all terms up to and including those of polynomials of degree four. Doing the calculation with the help of the coefficient functions \hat{T}_{ij}, \dots instead of ordering in powers of $1/h$ implies this criterion. The reason is very simple. The value of the coefficient functions grows very rapidly and the complete expression would not be manageable. In this case there are only contributions from the first and second iterations:

$$\begin{aligned} \text{IT1} &= \frac{2\Delta}{h} - c_4 \frac{1}{h^2} S_4 - c_6 \frac{1}{h^4} S_6 - c_8 S_8 \frac{1}{h^6} + O(q^5), \\ \text{IT2} &= \frac{1}{h^2} c_3^2 \hat{T}_{33} + c_3 c_5 \frac{1}{h^4} (\hat{T}_{35} + \hat{T}_{53}) \\ &\quad + c_4^2 \frac{1}{h^4} \hat{T}_{44} + O(q^5). \end{aligned} \quad (121)$$

Substituting the coefficient functions gives

$$\begin{aligned} \text{IT1} &= \frac{2\Delta}{h} - \frac{c_4}{h^2} \frac{3}{2}(q^2 + 1) - \frac{c_6}{h^4} \frac{3}{2} q(q^2 + 5) \\ &\quad - \frac{c_8}{h^6} \frac{36}{8} (q^4 + 14q^2 + 9) \\ &\quad + O(q^5), \end{aligned} \quad (122)$$

and

$$\begin{aligned} \text{IT2} &= \frac{c_3^2}{16} \left(\frac{q^3 + 9q^2 + 23q + 15}{3h^2 - \Delta h} - \frac{q^3 - 9q^2 + 23q - 15}{3h^2 + \Delta h} + 9 \frac{q^3 + 3q^2 + 3q + 1}{h^2 - \Delta h} - 9 \frac{q^3 - 3q^2 + 3q - 1}{h^2 + \Delta h} \right) \\ &\quad + \frac{5}{8h^2} c_3 c_5 \left(\frac{q^4 + 12q^3 + 50q^2 + 84q + 45}{3h^2 - \Delta h} - \frac{q^4 - 12q^3 + 50q^2 - 84q + 45}{3h^2 + \Delta h} \right. \\ &\quad \left. + 3 \frac{q^4 + 4q^3 + 8q^2 + 8q + 3}{h^2 - \Delta h} - 3 \frac{q^4 - 4q^3 + 8q^2 - 8q + 3}{h^2 + \Delta h} \right) \\ &\quad + \frac{c_4^2}{2h^2} \left(\frac{q^4 + 16q^3 + 86q^2 + 176q + 105}{16(4h^2 - \Delta h)} - \frac{q^4 - 16q^3 + 86q^2 - 176q + 105}{16(4h^2 + \Delta h)} \right. \\ &\quad \left. + \frac{q^4 + 8q^3 + 23q^2 + 28q^2 + 12}{2h^2 - \Delta h} - \frac{q^4 - 8q^3 + 23q^2 - 28q + 12}{2h^2 + \Delta h} \right) \\ &\quad + O(q^5). \end{aligned} \quad (123)$$

The complete expression is

$$O = \text{IT1} + \text{IT2} + O(q^5). \quad (124)$$

We have also calculated the contributions of order (q^5) , which are given in their general form in Appendix F.

6.3. Regge trajectories $\alpha_q(E)$ and particle spectroscopy

In Ref. 11 it was shown that for the linear potential the angular momentum l is given by

$$(l + \frac{1}{2})^2 = h^4/12 - \frac{1}{2} q h^2 - \Delta h \quad (125)$$

Thus in calculating the Regge trajectories $\alpha_q \equiv l$ we need Δh as a function of $\alpha = 2\mu(E - V_0)/h$. In this section we give the results for some iterations. In particular we are interested in the behavior of these trajectories for s wave states.

First of all we consider the numerical results of Eq. (117) and their approximations given by Eq. (119). Numerical evaluation shows, that the approximation is good in a far wider range, i.e., even if the condition

$$c_3^2(q/72)(41q^2 + 133) < h^2 \quad (126)$$

is not strictly fulfilled. Figure 1 gives these results for $q = 1$ and $q = 3$; the corresponding Regge trajectories are shown in Figs. 3 and 4.

The solutions Δh of the quadratic equation (117) are not necessarily real. For sufficiently small values of h the solutions can be complex. However, it can be shown that for $h \gg 1$ the solutions are always real. Difficulties arise only near $h \sim 1$, i.e., in the domain of the s states. Our curves in Figs. 1-4 start at this point, where Eq. (117) develops real solutions.

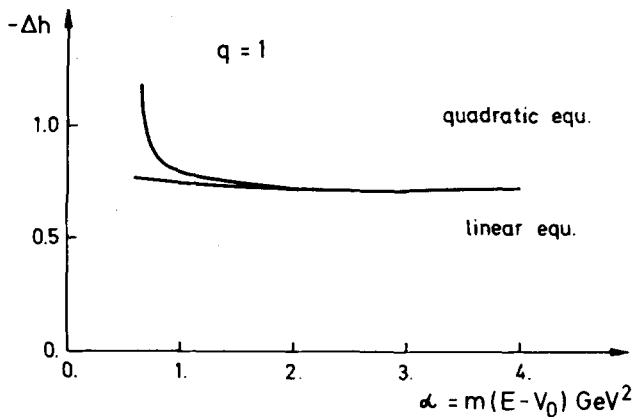


FIG. 1. Δh for $q = 1$, as obtained from Eq. (117) or the linear approximation Eq. (119).

For $q = 1$ the difference between the results obtained with these methods of approximation is small. The case of $q = 3$ presents a new problem.

The Regge trajectories do not become negative, so that we cannot determine an s state. Here we can use only the approximation (119). This seems to be an acceptable and smooth extrapolation from high energies to low energies.

We require additional contributions to our iterations, but we have to remember, of course, that after a certain number of iterations the asymptotic expansion begins to diverge. In the case of the linear potential the expansion behaves so badly, that the next iteration cannot be used. The best way to include higher order contributions seems to be to use the coefficient functions \hat{T}_i, \dots and to take into account all contributions up to and including polynomials of degree q^n . This kind of procedure has further advantages. In this approximation, we always obtain a real result for Δh . The reason is the following. We always add to the iteration of Δh symmetric contributions of the form:

$$\frac{P(q)}{nh^2 - \Delta h} \pm \frac{P(-q)}{nh^2 + \Delta h}, \quad (127)$$

where n is an integer and $P(q)$ a polynomial. If we multiply our equation for Δh by the common denominator, we always get a polynomial of odd degree, so that an additional term of the form (127) raises the degree of the polynomial by 2. In this way we have removed the problem of complex results for Δh . But the other problem remains, i.e., the early divergence

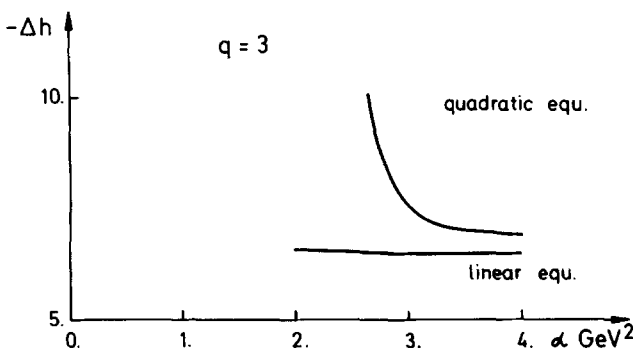


FIG. 2. The same as Fig. 1 with $q = 3$.

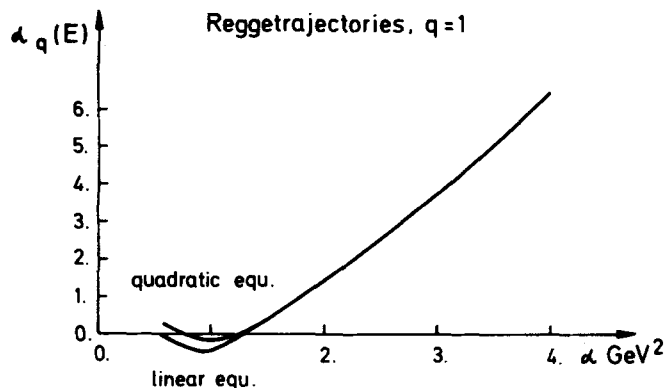


FIG. 3. Regge-trajectories for the corresponding results of Δh from Eq. (117) and Eq. (119) for $q = 1$.

of the asymptotic approximation for the linear potential.

In Figs. 5 and 6 we give the correction Δh and the Regge trajectories for the same values of parameters as were used for Figs. 1–4. Figures 5 and 6 show the results for $q = 1$ for calculations up to polynomials of different degree.

Detailed analysis shows that for $h \approx 1$ the contributions of polynomials of higher degree dominate over those of lower degree. This is a consequence of the asymptotic behavior of the iteration. However, from Fig. 6 we see also that it is an alternating effect, so it is suggestive to use a summation procedure which cancels the effect of alternating contributions to this order. A simple and effective method is the Hölder–Cesaro summation procedure.¹⁷

If we use an arithmetic procedure, we obtain the mean of the three curves of Fig. 6, so the iteration up to $O(q^3)$ is sufficient. Without more complicated techniques of summation this would be the limit of our asymptotic iteration for the linear-power potential. It should be noted that we have adjusted the parameter β in such a way, that the two quantities underlined in Table II correspond to well-known experimental values.

The difference between the present and earlier results stems from a false sign in Ref. 11. Further we see that the unknown parameters (β, δ) can be varied over a large range. Finally Figs. 7 and 8 give the Regge trajectories for different values of β and $q = 1, 3, 5, 7$ using the approximation up to polynomials of degree 3.

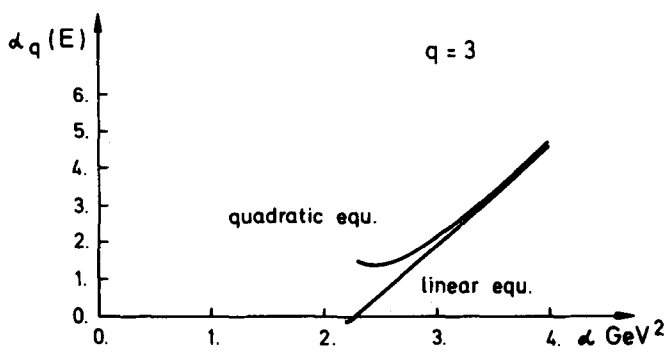


FIG. 4. The Regge-trajectories for $q = 3$ obtained from Eqs. (117) and (119).

TABLE II. The linear potential .

	$m = 1.65$ GeV;		$(m = \text{Quark mass})$		
	1	2	3	4	
$\beta =$	0.3934	0.3726	0.2390	0.2335	GeV ³
$\delta =$	0	$\frac{1}{2}\beta^{1/3}$	0	$\frac{1}{2}\beta^{1/3}$	GeV
$q = 1$	3.096	3.096	3.096	3.096	Input
$q = 3$	3.684	3.684	3.684	3.684	
$q = 5$	4.178	4.165	4.229	4.225	
$q = 7$	4.619	4.592	4.717	4.706	

1 and 2: masses of s states of Eq. (119).
3 and 4: masses of s states of Eq. (124) up to $O(q^3)$.

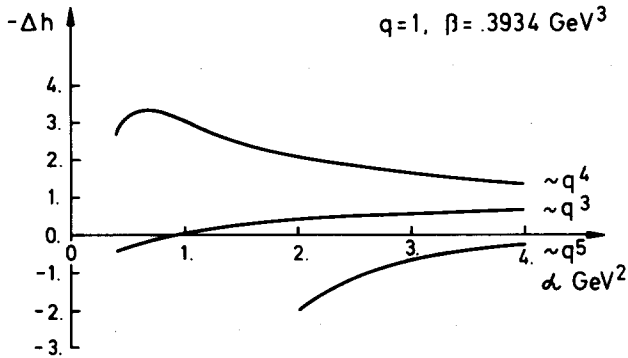


FIG. 5. Comparison of the results for Δh obtained from Eq. (124) if all contributions of order q^3, q^4, q^5 are taken into account.

6.4. The logarithmic potential combined with a Coulomb potential

Another potential which has attracted considerable interest in connection with the spectroscopy of the newly discovered heavy quark-antiquark states is the logarithmic potential.¹⁰ We consider this potential in the form

$$V(r) = -g_c/r + g \ln(r/r_0), \quad (128)$$

where $g, g_c \geq 0$.

After separation of the motion of the center of mass, the Schrödinger equation for the radial wavefunction

$$\psi_{(r)} = (1/r^{1/2})\phi$$

is

$$\frac{d^2\phi(r)}{dr^2} + \left[\alpha - \frac{\gamma}{r^2} - \beta \left(\ln r - \frac{\delta}{r} \right) \right] \phi(r) = 0, \quad (129)$$

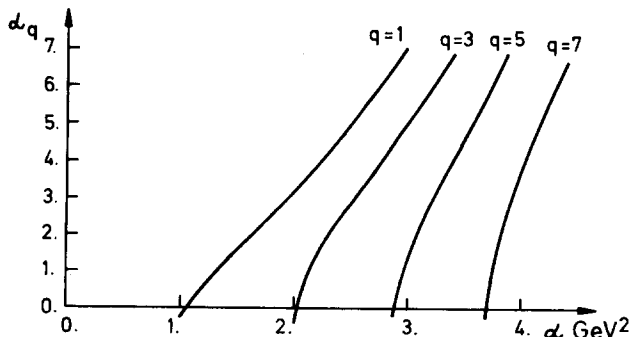


FIG. 7. Regge-trajectories for the linear potential obtained from Eq. (124) and for $\delta = 0, \beta = 0.2390 \text{ GeV}^3$ and terms up to q^3 .

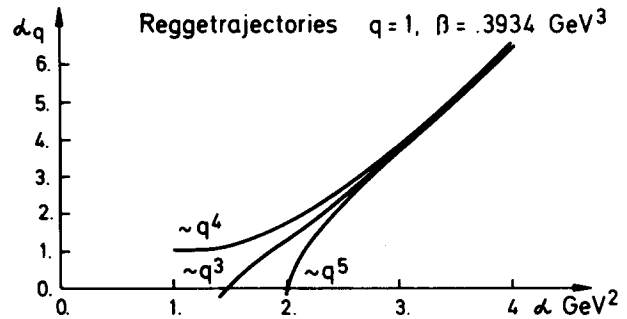


FIG. 6. Regge-trajectories for the Δh -values of Fig. 5.

where

$$\alpha = (2\mu/\hbar^2)(E + g \ln r_0),$$

$$\beta = \frac{2\mu g}{\hbar^2}, \quad \gamma = l(l+1),$$

$$\delta = g_c/g. \quad (130)$$

Setting

$$r = e^{z-c} \quad (131)$$

and choosing

$$c = -\alpha/\beta, \quad (132)$$

we obtain our basic equation

$$\frac{d^2\psi}{dz^2} + [-L^2 + U(z)]\psi(z) = 0, \quad (133)$$

where

$$L^2 \equiv \gamma + \frac{1}{4} = (l + \frac{1}{2})^2 \quad (134)$$

and

$$U(z) = \beta e^{2(z-c)} (\delta e^{-z+c} - z). \quad (135)$$

The maximum of $U(z) - L^2$ at z_0 is given by

$$\delta e^{-z_0+c} - 2z_0 - 1 = 0. \quad (136)$$

Setting

$$x = \delta \frac{1}{2} e^{c+\frac{1}{2}}, \quad (137)$$

we obtain

$$z_0 \approx -\frac{1}{2} + \frac{1}{4}(3 + 1/(1+x)^2) \ln(1+x). \quad (138)$$

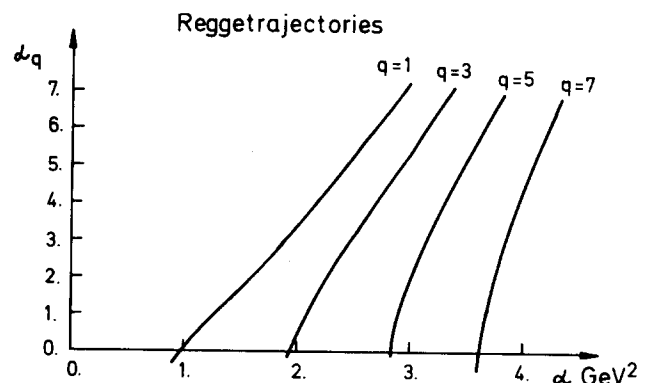


FIG. 8. Regge-trajectories for the linear potential combined with a Coulomb potential obtained from Eq. (124) for $\delta = \frac{1}{2}, \beta^{1/3}; \beta = 0.2335 \text{ GeV}^3$.

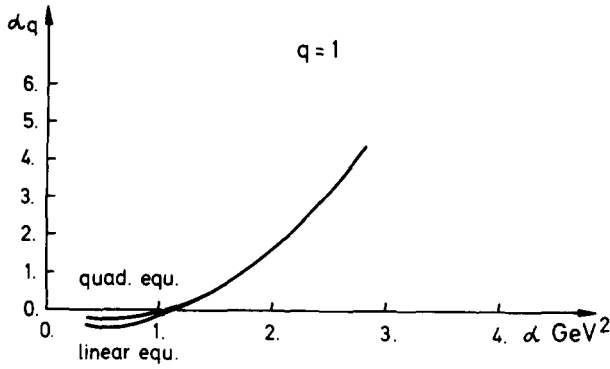


FIG. 9. Comparison of the results for Δh obtained from the quadratic equation and the linear approximation.

Again we expand $U(z)$ around z_0 . Then

$$U(z) = U(z_0) + \sum_{i=2}^{\infty} \frac{(z-z_0)^i}{i!} U^{(i)}(z_0), \quad (139)$$

where

$$U^{(i)}(z_0) = \beta e^{2(z_0-c)} (\delta e^{c-z_0} - 2^i z_0 - i 2^{i-1}). \quad (140)$$

Next we set

$$h^2 = [-2U^{(2)}(z_0)]^{1/2} \quad (141)$$

and change the independent variable to

$$\omega = h(z-z_0). \quad (142)$$

Then

$$\frac{d^2\psi}{d\omega^2} + \frac{1}{h^2} \left(-L^2 + U(\omega=0) + \sum_{i=2}^{\infty} U^{(i)} \frac{1}{i!} \frac{\omega^i}{h^i} \right) \psi(\omega) = 0. \quad (143)$$

This equation can now be written in our standard form (10). Thus

$$\mathcal{D}_q \psi(\omega) = \left(\frac{2\Delta}{h} + \sum_{i=3}^{\infty} \frac{2U^{(i)}}{i! h^{i+2}} \omega^i \right) \psi, \quad (144)$$

with

$$\begin{aligned} h_0 &= 2\Delta/h, \\ h_1 &= h_2 = 0, \\ h_i &= \frac{2U^{(i)}}{i! h^{i+2}} = -\frac{U^{(i)}}{i! U^{(2)}} \frac{1}{h^{i-2}}. \end{aligned} \quad (145)$$

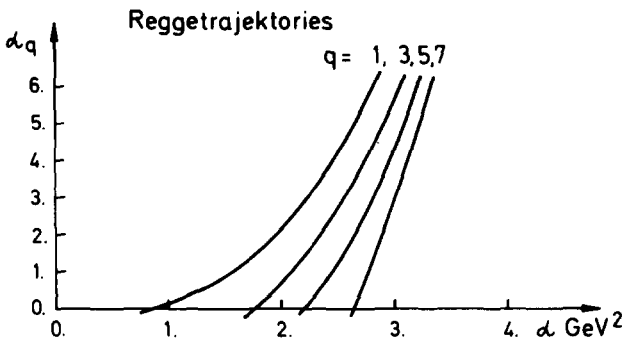


FIG. 11. The linear approximation (147) with the parameters $\beta = 1.004 \text{ GeV}^3$ and $\delta = \beta/4$.

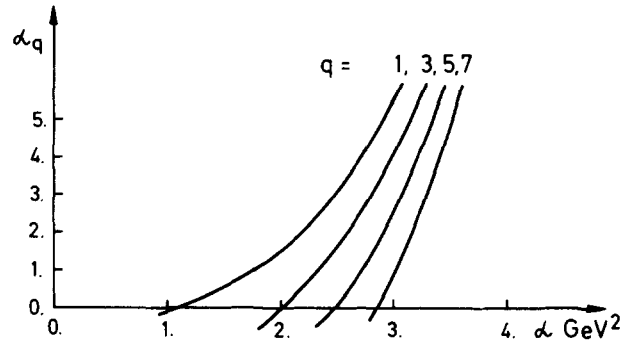


FIG. 10. Regge-trajectories using the linear approximation (147) for the logarithmic potential with $\delta = 0$ and $\beta = 1.113 \text{ GeV}^3$.

We now determine the correction Δ in terms of these coefficients. For $g_c = 0$ we obtain particularly simple expressions of z_0 and hence for the coefficients h_i . Thus

$$z_0 = -\frac{1}{2}$$

and

$$h_i = -\frac{(1-i) 2^{i-2}}{i!} \frac{1}{h^{i-2}}. \quad (146)$$

A crude approximation, similar to (119), of the linear potential is

$$\begin{aligned} -\Delta h &= \frac{1}{2^3 \times 3^2} (33q^2 + 1) \\ &+ \frac{q}{2^5 \times 3^3} (43q^2 + 1) \frac{1}{h^2} \\ &+ \frac{1}{h^4} \frac{1}{199 \times 2^6 \times 3^2 \times 5^2} \\ &\times (264\,445q^4 - 23\,790q^2 - 43\,199) \\ &+ O\left(\frac{1}{h^6}\right). \end{aligned} \quad (147)$$

We now perform calculations similar to those for the linear potential. The Regge-trajectories are obtained from

$$L^2 = (l + \frac{1}{2})^2 = U(\omega=0) - \frac{1}{2} q h^2 - \Delta h. \quad (148)$$

Generally speaking we find that the logarithmic potential is

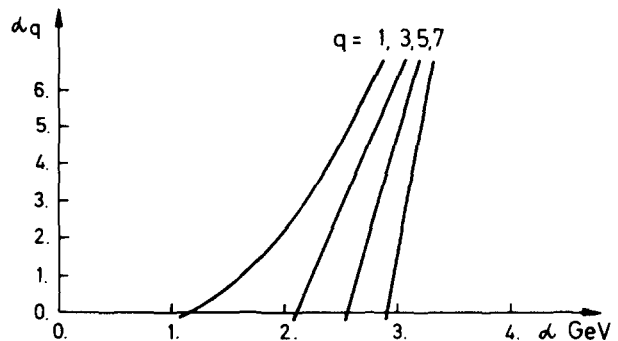


FIG. 12. Regge-trajectories calculated by an iteration similar to Eq. (124) including terms of $O(q^3)$, with $\delta = 0$ and $\beta = 0.951 \text{ GeV}^3$.

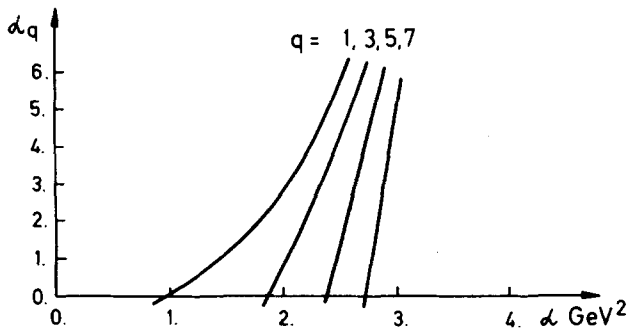


FIG. 13. The same as Fig. 12, but with $\delta = \beta/4$ and $\beta = 0.884 \text{ GeV}^3$.

easier to handle in an iterative treatment than the linear potential. In particular the coefficients h_i decrease more rapidly. However, this is not very important in the calculation of the s states. Here we present the results for the s states using the approximation (147) and an equation which is similar to (121).

Figure 9 shows the Regge-trajectory for $q = 1$, using (147) and the solution of the original quadratic equation. Figures 10–13 show the same Regge-trajectory for different values of β and δ . Finally, Table III summarizes the values of the meson masses of the s states using again typical values of the quark masses.

TABLE III. The logarithmic potential.

	s states with						
	$m = 1.56 \text{ GeV}$						
	1	2	3	4	5		
β	0.888	1.113	1.004	0.951	0.884	GeV ²	
δ	0	0	$\beta/4$	0	$\beta/4$		
$q = 1$	3.095	3.096	3.096	3.096	3.096	} Input	
$q = 3$	3.685	3.684	3.684	3.684	3.684		
$q = 5$	4.008	4.014	4.002	4.020	4.008		GeV
$q = 7$	4.233	4.242	4.218	4.242	4.220		GeV
$Q = 9$	4.405	4.416	4.381	4.405	4.374		GeV

1: from numerical integration (Ref. 18).

2 and 3: approximation with Eq. (147).

4 and 5: approximation with Eq. (124).

APPENDIX A: THE COEFFICIENTS $S_j(p, 2j)$

Here we give the explicit form of the coefficient functions S_j defined by Eq. (18). They are obtained from the recurrence relation of the parabolic cylinder function ψ_p , i.e.,

$$z\psi_p = \psi_{p+2} + \frac{1}{2}(p-1)\psi_{p-2}.$$

The function $W(p, j)$ referred to in the text [cf. Eq. (44)] is defined as

$$W(p, j) = \begin{cases} 1 & \text{if } j \geq 0, \\ \frac{p-1}{2} \frac{p-3}{2} \dots \frac{p+1+2j}{2} & \text{if } j < 0. \end{cases}$$

Results of $z^n \psi_p(z)$ are:

$$\begin{aligned} z\psi_p &= W(p, 1)\psi_{p+2} + W(p, -1)\psi_{p-2}, \\ z^2\psi_p &= \psi_p = W(p, 2)\psi_{p+4} + W(p, -2)\psi_{p-4} + p\psi_p, \\ z^3\psi_p &= W(p, 3)\psi_{p+6} + W(p, -3)\psi_{p-6} + \frac{3}{2}(p+1)W(p, 1)\psi_{p+2} + \frac{3}{2}(p-1)W(p, -1)\psi_{p-2}, \\ z^4\psi_p &= W(p, 4)\psi_{p+8} + W(p, -4)\psi_{p-8} + 2(p+2)W(p, 2)\psi_{p+4} + 2(p-2)W(p, -2)\psi_{p-4} + \frac{3}{2}(p^2+1)\psi_p, \\ z^5\psi_p &= W(p, 5)\psi_{p+10} + W(p, -5)\psi_{p-10} + \frac{5}{2}(p+3)W(p, 3)\psi_{p+6} + \frac{5}{2}(p-3)W(p, -3)\psi_{p-6} \\ &\quad + \frac{5}{2}(p^2+2p+3)W(p, 1)\psi_{p+2} + \frac{5}{2}(p^2-2p+3)W(p, -1)\psi_{p-2}, \\ z^6\psi_p &= \\ &W(p, 6)\psi_{p+12} + W(p, -6)\psi_{p-12} + 3(p+4)W(p, 4)\psi_{p+8} + 3(p-4)W(p, -4)\psi_{p-8} + \frac{15}{4}(p^2+4p+7)W(p, 2)\psi_{p+4} \\ &\quad + \frac{15}{4}(p^2-4p+7)W(p, -2)\psi_{p-4} + \frac{5}{2}p(p^2+5)\psi_p, \\ z^7\psi_p &= W(p, 7)\psi_{p+14} + W(p, -7)\psi_{p-14} + \frac{7}{2}(p+5)W(p, 5)\psi_{p+10} + \frac{7}{2}(p-5)W(p, -5)\psi_{p-10} \\ &\quad + \frac{21}{4}(p^2+6p+13)W(p, 3)\psi_{p+6} + \frac{21}{4}(p^2-6p+13)W(p, -3)\psi_{p-6} \\ &\quad + \frac{35}{8}(p^3+3p^2+11p+9)W(p, 1)\psi_{p+2} + \frac{35}{8}(p^3-3p^2+11p-9)W(p, -1)\psi_{p-2}, \\ z^8\psi_p &= \\ &W(p, 8)\psi_{p+16} + W(p, -8)\psi_{p-16} + 4(p+6)W(p, 6)\psi_{p+12} + 4(p-6)W(p, -6)\psi_{p-12} + 7(p^2+8p+21)W(p, 4)\psi_{p+8} \\ &\quad + 7(p^2-8p+21)W(p, -4)\psi_{p-8} + 7(p^3+6p^2+23p+30)W(p, 2)\psi_{p+4} \\ &\quad + 7(p^3-6p^2+23p-30)W(p, -2)\psi_{p-4} + \frac{35}{8}(p^4+14p^2+9)\psi_p. \end{aligned}$$

APPENDIX B: LIST OF THE COEFFICIENTS $C(p+2j)$

The following are the expressions for the coefficients C calculated from (20). We neglect all contributions of $O(1/g^9)$:

$$\begin{aligned} C(p, p) &= H_0 + pH_2 + \frac{3}{2}(p^2+1)H_4 + \frac{5}{2}p(p^2+5)H_6 + \frac{35}{8}(p^4+14p^2+89)H_8 + \frac{63}{8}p(p^4+30p^2+89)H_{10} \\ &\quad + \frac{231}{16}(p^6+55p^4+439p^2+225)H_{12} + \frac{429}{16}p(p^6+91p^4+1519p^2+3429)H_{14} + O(1/g^9), \end{aligned}$$

$$\begin{aligned} C(p, p+2) &= H_1 + \frac{3}{2}(p+1)H_3 + \frac{5}{2}(p^2+2p+3)H_5 + \frac{35}{8}(p^3+3p^2+11p+9)H_7 + \frac{63}{8}(p^4+4p^3+26p^2+44p+45)H_9 \\ &\quad + \frac{231}{16}(p^5+5p^4+50p^3+130p^2+309p+225)H_{11} + \frac{429}{16}p(p^6+6p^5+85p^4+300p^3+1219p^2+1854p+1575)H_{13}, \end{aligned}$$

$$\begin{aligned}
C(p, p-2) &= W(p, -1) [H_1 + \frac{3}{2}(p-1)H_3 + \frac{5}{2}(p^2-2p+3)H_5 + \frac{35}{8}(p^3-3p^2+11p-9)H_7 \\
&\quad + \frac{63}{8}(p^4-4p^3+26p^2-44p+45)H_9 + \frac{231}{16}(p^5-5p^4+50p^3-130p^2+309p-225)H_{11} \\
&\quad + \frac{429}{16}(p^6-6p^5+85p^4-300p^3+1219p^2-1854p+1575)H_{13}], \\
C(p, p+4) &= H_2 + 2(p+2)H_4 + \frac{1}{4}(p^2+4p+7)H_6 + 7(p^3+6p^2+23p+30)H_8 \\
&\quad + \frac{195}{8}(p^4+8p^3+50p^2+136p+165)H_{10} + \frac{29}{4}(p^5+10p^4+90p^3+380p^2+989p+1050)H_{12}, \\
C(p, p-4) &= W(p, -2) [H_2 + 2(p-2)H_4 + \frac{1}{4}(p^2-4p+7)H_6 + 7(p^3-6p^2+23p-30)H_8 \\
&\quad + \frac{195}{8}(p^4-8p^3+50p^2-136p+165)H_{10} + \frac{29}{4}(p^5-10p^4+90p^3-380p^2+989p-1050)H_{12}], \\
C(p, p+6) &= H_3 + \frac{5}{2}(p+3)H_5 + \frac{7}{4}(p^2+6p+13)H_7 + \frac{7}{2}(p^3+9p^2+41p+69)H_9 \\
&\quad + \frac{165}{8}(p^4+12p^3+86p^2+300p+441)H_{11}, \\
C(p, p-6) &= W(p, -3) [H_3 + \frac{5}{2}(p-3)H_5 + \frac{7}{4}(p^2-6p+13)H_7 \\
&\quad + \frac{7}{2}(p^3-9p^2+41p-69)H_9 + \frac{165}{8}(p^4-12p^3+86p^2-300p+441)H_{11}], \\
C(p, p+8) &= H_4 + 3(p+4)H_6 + 7(p^2+8p+21)H_8 + 15(p^3+12p^2+65+132)H_{10}, \\
C(p, p-8) &= W(p, -4) [H_4 + 3(p-4)H_6 + 7(p^2-8p+21)H_8 + 15(p^3-12p^2+65p-132)H_{10}], \\
C(p, p+10) &= H_5 + \frac{7}{2}(p+5)H_7 + 9(p^2+10p+31)H_9, \\
C(p, p-10) &= W(p, -5) [H_5 + \frac{7}{2}(p-5)H_7 + 9(p^2-10p+31)H_9], \\
C(p, p+12) &= H_6 + 4(p+6)H_8 \\
C(p, p-12) &= W(p, -6) [H_6 + 4(p-6)H_8], \\
C(p, p+14) &= H_7, \\
C(p, p-14) &= W(p, -7)H_7.
\end{aligned}$$

APPENDIX C: THE COEFFICIENTS $S_i, T_{ij}, U_{ijk}, \dots$

Here we give the coefficient functions S_i, T_{ij}, \dots which are discussed in Sec. 4.

We have the following general results:

- (i) The coefficient functions are zero if the first or last index is zero.
- (ii) A coefficient function is zero if the sum of the indices is an odd integer.

The terms $S_i(p)$ are:

$$\begin{aligned}
S_0 &= 1, \\
S_2 &= p, \\
S_4 &= \frac{3}{2}(p^2+1), \\
S_6 &= \frac{5}{2}p(p^2+5), \\
S_8 &= \frac{35}{8}(p^4+14p^2+9), \\
S_{10} &= \frac{63}{8}p(p^4+30p^2+89), \\
S_{12} &= \frac{231}{16}(p^6+55p^4+439p^2+225), \\
S_{14} &= \frac{429}{16}p(p^6+91p^4+1519p^2+3429).
\end{aligned}$$

Terms which contain $i = 1$ are:

$$\begin{aligned}
T_{11} &= \frac{1}{2}, \\
T_{13} &= T_{31} = \frac{3}{2}p, \\
T_{15} &= T_{51} = \frac{15}{4}(p^2+1), \\
T_{17} &= T_{71} = \frac{35}{4}p(p^2+5); \\
U_{101} &= \frac{1}{4}p, \\
U_{103} &= U_{301} = \frac{3}{8}(p^2+1), \\
U_{105} &= U_{501} = \frac{5}{8}(p^2+5)p, \\
U_{107} &= U_{701} = \frac{35}{32}(p^4+14p^2+9),
\end{aligned}$$

$$\left. \begin{aligned}
U_{211} &= \frac{1}{16}(p^2+3) \\
U_{121} &= \frac{1}{8}(p^2+5) \\
U_{112} &= U_{211}
\end{aligned} \right\} \Sigma = \frac{1}{4}(p^2+4),$$

$$\left. \begin{aligned}
U_{411} &= \frac{1}{8}p(p^2+11) \\
U_{141} &= \frac{1}{8}p(p^2+28) \\
U_{114} &= U_{411}
\end{aligned} \right\} \Sigma = \frac{p}{8}(3p^2+50),$$

$$\left. \begin{aligned}
U_{123} &= \frac{1}{24}(5p^2+61) \\
U_{231} &= \frac{1}{16}p(p^2+23) \\
U_{312} &= \frac{1}{48}p(5p^2+43) \\
U_{132} &= U_{231} \\
U_{321} &= U_{123} \\
U_{213} &= U_{312}
\end{aligned} \right\} \Sigma = \frac{p}{12}(2p^2+107);$$

$$\begin{aligned}
V_{1001} &= \frac{1}{8} \\
V_{1003} &= V_{3001} = \frac{3}{8}p, \\
V_{1005} &= V_{5001} = \frac{15}{16}(p^2+1), \\
V_{2011} &= \frac{1}{16}p \\
V_{2101} &= \frac{1}{8}p \\
V_{1201} &= \frac{3}{8}p \\
V_{1012} &= \frac{1}{8}p \\
V_{1102} &= 1/16p \\
V_{1021} &= \frac{3}{8}p
\end{aligned} \right\} \Sigma = \frac{9}{8}p,$$

$$\left. \begin{aligned}
V_{4011} &= \frac{3}{16}(p^2+1) \\
V_{4101} &= \frac{3}{8}(p^2+1) \\
V_{1401} &= \frac{15}{16}(p^2+1) \\
V_{1041} &= \frac{15}{16}(p^2+1) \\
V_{1014} &= \frac{3}{8}(p^2+1) \\
V_{1104} &= \frac{3}{16}(p^2+1)
\end{aligned} \right\} \Sigma = \frac{27}{8}(p^2+1),$$

$$\begin{aligned}
V_{1111} &= \frac{1}{8}p, \\
V_{1113} &= \frac{1}{32}(9p^2 + 7), \\
V_{1131} &= \frac{3}{32}(3p^2 + 5), \\
V_{1311} &= \frac{3}{32}(3p^2 + 5), \\
V_{3111} &= \frac{1}{32}(9p^2 + 7); \\
W_{10001} &= \frac{1}{16}p, \\
W_{10003} &= W_{30001} = \frac{3}{32}(p^2 + 1), \\
W_{10005} &= W_{50001} = \frac{5}{32}p(p^2 + 5), \\
W_{10111} &= W_{11101} = \frac{1}{16}(p^2 + 3), \\
W_{11011} &= \frac{1}{128}(p^2 + 3); \\
X_{1\ 000\ 1} &= \frac{1}{32}, \\
X_{1\ 0000\ 3} &= X_{3\ 0000\ 1} = 3/32 p.
\end{aligned}$$

Terms containing $i = 2$ are:

$$\begin{aligned}
T_{22} &= \frac{1}{2}p, \\
T_{42} = T_{24} &= \frac{3}{2}(p^2 + 1), \\
T_{62} = T_{26} &= \frac{15}{4}p(p^2 + 5); \\
U_{202} &= \frac{1}{32}(p^2 + 3), \\
U_{402} = U_{204} &= \frac{1}{16}(p^2 + 11), \\
U_{602} = U_{206} &= \frac{15}{128}(p^4 + 26p^2 + 21), \\
U_{222} &= \frac{1}{32}p(p^2 + 19); \\
V_{2002} &= \frac{1}{32}p, \\
V_{2004} = V_{4002} &= \frac{3}{32}(p^2 + 1), \\
V_{2202} = V_{2022} &= \frac{1}{32}(2p^2 + 3), \\
W_{20002} &= \frac{1}{512}(p^2 + 3), \\
W_{20004} = W_{40002} &= \frac{1}{256}p(p^2 + 11).
\end{aligned}$$

Terms with index $i > 2$ are:

$$\begin{aligned}
T_{33} &= \frac{1}{4}(15p^2 + 7), \\
T_{35} = T_{53} &= \frac{5}{4}p(7p^2 + 19), \\
T_{37} = T_{73} &= \frac{105}{16}(3p^4 + 26p^2 + 11), \\
T_{39} = T_{93} &= \frac{7}{16}p(33p^4 + 670p^2 + 1337), \\
T_{44} &= \frac{1}{4}p(17p^2 + 67), \\
T_{46} = T_{64} &= \frac{15}{16}(11p^4 + 118p^2 + 63), \\
T_{48} = T_{84} &= \frac{63}{8}p(3p^4 + 70p^2 + 167); \\
U_{303} &= \frac{1}{72}p(41p^2 + 133), \\
U_{305} = U_{503} &= \frac{5}{288}(55p^4 + 482p^2 + 207), \\
U_{307} = U_{703} &= \frac{7}{96}p(23p^4 + 450p^2 + 847), \\
U_{309} = U_{903} &= \frac{7}{192}(83p^6 + 3125p^4 + 6277p^2 + 5715), \\
U_{334} &= \frac{3}{64}p(5p^4 + 334p^2 + 749), \\
U_{343} &= \frac{1}{96}(37p^4 + 2742p^2 + 4541), \\
U_{433} &= U_{334},
\end{aligned}$$

$$\begin{aligned}
U_{336} &= (527p^6 + 54\ 725p^4 + 356\ 273p^2 + 131\ 355)/1152, \\
U_{633} &= U_{336} \\
U_{363} &= (293p^6 + 49\ 775p^4 + 271\ 547p^2 + 86\ 625)/576, \\
U_{354} &= (239p^6 + 37\ 445p^4 + 248\ 921p^2 + 93\ 555)/768, \\
U_{345} &= 669p^6 + 71\ 795p^4 + 351\ 691p^2 + 125\ 685/960, \\
U_{435} &= (181p^6 + 17\ 655p^4 + 104\ 259p^2 + 45\ 585)3/1280, \\
U_{345} &= U_{543}, \\
U_{453} &= U_{354}, \\
U_{534} &= U_{435}, \\
U_{404} &= \frac{1}{512}(65p^4 + 1558p^2 + 873), \\
U_{444} &= \frac{3}{1024}(65p^6 + 9623p^4 + 75\ 263p^2 + 33\ 705), \\
U_{406} = U_{604} &= \frac{3}{512}p(41p^4 1830p^2 + 4849); \\
V_{3003} &= \frac{1}{144}(123p^2 + 43), \\
V_{3333} &= \frac{1}{288}p(615p^4 + 14\ 972p^2 + 21\ 721), \\
V_{3705} = V_{5003} &= \frac{25}{144}p(11p^2 + 23), \\
V_{4004} &= \frac{5}{256}p(13p^2 + 47), \\
V_{3403} &= (693p^4 + 3782p^2 + 1085)/192, \\
V_{4033} &= \frac{3}{1024}(263p^4 + 2682p^2 + 959), \\
V_{4303} &= \frac{1}{48}(86p^4 + 601p^2 + 225), \\
V_{3043} &= V_{3403}, \\
V_{3034} &= V_{4303}, \\
V_{3304} &= V_{4033}; \\
W_{30003} &= \frac{5}{2592}p(73p^2 + 221), \\
W_{33303} &= (7507p^6 + 593\ 695p^4 + 2\ 521\ 681p^2 \\
&\quad + 643\ 005)/82\ 944, \\
W_{33033} &= (4294p^6 + 57\ 7468p^4 + 3\ 174\ 784p^2 \\
&\quad + 806\ 463)/165\ 888, \\
W_{30333} &= W_{33303}, \\
W_{40033} &= \frac{3}{4096}p(19p^4 + 1210p^2 + 2451), \\
W_{40303} &= \frac{1}{6144}p(301p^4 + 9534p^2 + 16\ 853), \\
W_{43003} &= W_{30034}, \\
W_{34003} &= W_{33004}, \\
W_{30403} &= \frac{1}{3456}p(1261p^4 + 16\ 422p^2 + 21\ 317), \\
W_{33004} &= \frac{1}{3456}p(289p^4 + 19\ 950p^2 + 24\ 521), \\
W_{30034} &= \frac{1}{768}p(39p^4 + 2362p^2 + 4063), \\
W_{30043} &= W_{33004}, \\
W_{30304} &= W_{40303}; \\
X_{300003} &= \frac{1}{5184}(1095p^2 + 367).
\end{aligned}$$

APPENDIX D: THE COEFFICIENTS $\hat{T}_{ij} \dots$

Here we give the coefficient functions \hat{T}_{ij} which we need if $h_1 = h_2 = 0$. We quote these up to $O(p^5)$:

$$\hat{T}_{33} = \frac{1}{8} \left(\frac{p^3 + 9p^2 + 23p + 15}{6 - H_0} - \frac{p^3 - 9p^2 + 23p - 15}{6 + H_0} \right. \\ \left. + 9 \frac{p^3 + 3p^2 + 3p + 1}{2 - H_0} - \frac{p^3 - 3p^2 + 3p - 1}{2 + H_0} \right),$$

$$\hat{T}_{35} = \hat{T}_{53} = \frac{1}{16} \left(\frac{p^4 + 12p^3 + 50p^2 + 84p + 45}{6 - H_0} \right. \\ \left. - \frac{p^4 - 12p^3 + 50p^2 - 84p + 45}{6 + H_0} \right) \\ + \frac{18}{8} \left(\frac{p^4 + 4p^3 + 8p^2 + 8p + 3}{2 - H_0} - \frac{p^4 - 4p^3 + 8p^2 - 8p + 3}{2 + H_0} \right),$$

$$\hat{T}_{37} = \hat{T}_{73} = \frac{195}{32}(p^5 + 5p^4 + 18p^3 + 34p^2 + 29p + 9)/(2 - H_0) \\ - \frac{195}{32}(p^5 - 5p^4 + 18p^3 - 34p^2 + 29p - 9)/(2 + H_0) \\ + \frac{21}{32}(p^5 + 15p^4 + 90p^3 + 270p^2 + 389p + 195)/(6 - H_0) \\ - \frac{21}{32}(p^5 - 15p^4 + 90p^3 - 270p^2 + 389p - 195)/(6 + H_0),$$

$$\hat{T}_{46} = \hat{T}_{64} = \frac{18}{8}(p^5 + 10p^4 + 42p^3 + 92p^2 + 101p + 42)/(4 - H_0) \\ - \frac{18}{8}(p^5 - 10p^4 + 42p^3 - 92p^2 + 101p - 42)/(4 + H_0) \\ + \frac{3}{16}(p^5 + 20p^4 + 150p^3 + 520p^2 + 809p + 420)/(8 - H_0) \\ - \frac{3}{16}(p^5 - 20p^4 + 150p^3 - 520p^2 + 809p - 420)/(8 + H_0),$$

$$\hat{T}_{55} = \frac{25}{8}(p^5 + 5p^4 + 14p^3 + 22p^2 + 21p + 9)/(2 - H_0) \\ - \frac{25}{8}(p^5 - 5p^4 + 14p^3 - 22p^2 + 21p - 9)/(2 + H_0) \\ + \frac{23}{32}(p^5 + 15p^4 + 86p^3 + 234p^2 + 297p + 135)/(6 - H_0) \\ - \frac{23}{32}(p^5 - 15p^4 + 86p^3 - 234p^2 + 297p - 135)/(6 + H_0) \\ + \frac{1}{32}(p^5 + 25p^4 + 230p^3 + 950p^2 + 1689p + 945)/(10 - H_0) \\ - \frac{1}{32}(p^5 - 25p^4 + 230p^3 - 950p^2 + 1689p - 945)/(10 + H_0)$$

$$\hat{U}_{343} = \frac{3}{16}(p^5 + 21p^4 + 168p^3 + 624p^2 + 1031p + 555)/(H_0^2 - 12H_0 + 36) \\ + \frac{3}{16}(p^5 - 21p^4 + 168p^3 - 624p^2 + 1031p - 555)/(H_0^2 + 12H_0 + 36) \\ + \frac{3}{4}(p^5 + 14p^4 + 72p^3 + 166p^2 + 167p + 60)/(H_0^2 - 8H_0 + 12) \\ + \frac{3}{4}(p^5 - 14p^4 + 72p^3 - 166p^2 + 167p - 60)/(H_0^2 + 8H_0 + 12) \\ + \frac{3}{32}(p^5 + 7p^4 + 6p^3 - 22p^2 - 7p + 15)/(H_0^2 - 4H_0 - 12) \\ + \frac{3}{32}(p^5 - 7p^4 + 6p^3 + 22p^2 - 7p - 15)/(H_0^2 + 4H_0 - 12) \\ + \frac{27}{16}(p^5 + 7p^4 + 20p^3 + 28p^2 + 19p + 5)/(H_0^2 - 4H_0 + 4) \\ + \frac{27}{16}(p^5 - 7p^4 + 20p^3 - 28p^2 + 19p - 5)/(H_0^2 + 4H_0 + 4) \\ + \frac{3}{32}(p^5 - 7p^4 + 6p^3 + 22p^2 - 7p - 15)/(H_0^2 + 4H_0 - 12) \\ + \frac{3}{32}(p^5 + 7p^4 + 6p^3 - 22p^2 - 7p + 15)/(H_0^2 - 4H_0 - 12) \\ + \frac{9}{4}p(p^4 - 2p^2 + 1)/(H_0^2 - 4),$$

$$\hat{U}_{334} = \hat{U}_{433} = \frac{3}{32}(p^5 + 23p^4 + 198p^3 + 778p^2 + 1337p + 735)/(H_0^2 - 14H_0 + 48) \\ + \frac{3}{32}(p^5 - 23p^4 + 198p^3 - 778p^2 + 1337p - 735)/(H_0^2 + 14H_0 + 48) \\ + \frac{3}{32}(p^5 + 17p^4 + 102p^3 + 262p^2 + 281p + 102)/(H_0^2 - 10H_0 + 16) \\ + \frac{3}{32}(p^5 - 17p^4 + 102p^3 - 262p^2 + 281p - 102)/(H_0^2 + 10H_0 + 16) \\ + \frac{3}{8}(p^5 + 16p^4 + 96p^3 + 266p^2 + 335p + 150)/(H_0^2 - 10H_0 + 24) \\ + \frac{3}{8}(p^5 - 16p^4 + 96p^3 - 266p^2 + 335p - 150)/(H_0^2 + 10H_0 + 24) \\ + \frac{9}{8}(p^5 + 10p^4 + 38p^3 + 68p^2 + 57p + 18)/(H_0^2 - 6H_0 + 8) \\ + \frac{9}{8}(p^5 - 10p^4 + 38p^3 - 68p^2 + 57p - 18)/(H_0^2 + 6H_0 + 8) \\ + \frac{3}{8}(p^5 - 4p^4 + 10p^2 - p - 6)/(H_0^2 + 2H_0 - 8) \\ + \frac{3}{8}(p^5 + 4p^4 - 10p^2 - p + 6)/(H_0^2 - 2H_0 - 8).$$

APPENDIX E: EXPLICIT EXPRESSIONS OF $F_j(y,p)$ AND $G_j(y,p)$

j	F_j	G_j
+ 4	$(2y^4 - 10y^2 - 4py^2 + p^2 + 6p + 5)/4$	$-y(y^2 - p - 4)$
+ 3	$3y(2y^2 - 3p - 5)/4$	$-(2y^2 - p - 3)/2$
+ 2	$(y^2 - p - 1)/2$	$-y$
+ 1	$y/2$	-1
0	1	0
- 1	$y/2$	+ 1
- 2	$(y^2 - p + 1)/2$	y
- 3	$y(2y^2 - 3p + 5)/4$	$(2y^2 + 3 - p)/2$
- 4	$(2y^4 + y^2(10 - 4p) + p^2 - 6p + 5)/4$	$y(y^2 - p + 4)$

APPENDIX F: THIRD-ORDER ITERATION OF THE EIGENVALUE OF ONE-DIMENSIONAL SCHRÖDINGER EQUATION

If we set for the coefficients h_i

$$h_0 \equiv 2\Delta / h,$$

$$h_1 = h_2 = 0,$$

$$h_i = -c_i / h^{i-2},$$

$$i \geq 3,$$

we get a polynomial of fourth-order in Δ . If we iterate this polynomial we get the following solution:

$$\begin{aligned}
 & -\Delta h \\
 = & \frac{1}{8} [7c_3^2 - 6c_4 + 3p^2(5c_3^2 - 2c_4)] + (p/32h^2) [p^2(280c_5c_3 + 68c_4^2 - 900c_4c_3^2 + 705c_3^4 - 40c_6) \\
 & + 760c_5c_3 + 268c_4^2 - 1836c_4c_3^2 + 1155c_3^4 - 200c_6] + (1/256h^4) [5p^4(504c_3^2 - 7728c_5c_4c_3 + 15624c_5c_3^3 \\
 & - 600c_4^3 + 19956c_4^2c_3^2 - 46530c_4c_3^4 + 528c_4c_6 + 23151c_3^6 - 4344c_3^2c_6 + 1008c_3c_7 - 112c_8) \\
 & + 2p^2(8680c_5^2 - 117360c_5c_4c_3 + 190920c_5c_3^3 + 13656c_4^3 + 248052c_4^2c_3^2 - 479970c_4c_3^4 \\
 & + 14160c_4c_6 + 209055c_3^6 - 68280c_3^2c_6 + 21840c_3c_7 - 3920c_8) + 8856c_5^2 - 90672c_5c_4c_3 + 118216c_5c_3^3 \\
 & - 12312c_4^3 + 161044c_4^2c_3^2 - 263634c_4c_3^4 + 15120c_4c_6 + 101479c_3^6 - 48440c_3^2c_6 \\
 & + 18480c_3c_7 - 5040c_8] + O(1/h^6).
 \end{aligned}$$

¹R. F. Dashen, J. B. Healy, and I. J. Muzinich, Phys. Rev. D **14**, 2773 (1976); Ann. Phys. (N.Y.) **102**, 1 (1976).

²G. V. Gehlen and V. Rittenberg, Phys. Lett. B **71**, 373 (1977).

³B. R. Karlsson and B. Kerbikov, Nucl. Phys. B **141**, 241 (1978); C. Dullemond and E. van Beveren, Ann. Phys. (N.Y.) **105**, 318 (1977).

⁴S. Nussinov and D. P. Sidhu, Fermilab Report No. 76/70-THY, (1976).

⁵D. Horn and D. E. Novoseller, Phys. Rev. D **17**, 1763 (1978).

⁶The standard reference is R. G. Newton, *Scattering Theory of Waves and Particles* (McGraw-Hill, New York, 1966).

⁷H. Feshbach, Ann. Phys. (N.Y.) **5**, 357 (1958); K. Gottfried, in *Elementary Particle Physics and Scattering Theory*, 1967 Brandeis Summer Institute in Theoretical Physics, edited by M. Chretien and S. S. Schweber (Gordon and Breach, New York, 1970), Vol. II, p. 148.

⁸R. B. Dingle and H. J. W. Müller, J. Reine Angew. Math. **211**, 11 (1962); *ibid.* **216**, 123 (1964); H. J. W. Müller, *ibid.* **211**, 33 (1962); H. J. W. Müller,

Math. Nachr. **32**, 157 (1966).

⁹The wave equation for the Gauss potential has been treated in H. J. W. Müller, J. Math. Phys. **11**, 355 (1970).

¹⁰H. J. W. Müller-Kirsten and S. K. Bose, J. Math. Phys. **20**, 2471 (1979).

¹¹H. J. W. Müller-Kirsten, G. E. Hite, and S. K. Bose, J. Math. Phys. **20**, 1878 (1979).

¹²H. J. W. Müller-Kirsten and R. Müller, Phys. Rev. D **20**, 2541 (1979).

¹³R. S. Kaushal and H. J. W. Müller-Kirsten, J. Math. Phys. **20**, 11 (1979).

¹⁴A. C. Hearn, REDUCE 2, Users Manual, University of Utah, Salt Lake City, March 1973.

¹⁵M. Abramowitz and A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1970).

¹⁶Schiff, *Quantum Mechanics* (McGraw Hill, New York, 1968).

¹⁷G. H. Hardy, *Divergent Series* (Oxford University Press, 1973).

¹⁸C. Quigg and J. L. Rosner, Phys. Lett. B **71**, 153 (1977); B **72**, 462 (1978).

Derivation of "Bethe's hypothesis" from the quantum inverse scattering transformation for the nonlinear Schrödinger equation

Andreas Wiesler

Fakultät für Physik der Universität Freiburg, Hermann-Herder-Str. 3, 7800 Freiburg, West Germany

(Received 28 November 1979; accepted for publication 14 March 1980)

Using the quantum version of the inverse scattering transformation for the nonlinear Schrödinger equation, eigenstates of the Hamiltonian can be constructed. We show that these eigenstates are of the Bethe form.

PACS numbers: 03.65.Ge, 03.70.+k

INTRODUCTION

It has been noticed recently^{1,2} that field theoretic models in 1 + 1 dimensions possessing an inverse scattering transformation may be solved exactly at the quantum level. In this note we treat the nonlinear Schrödinger theory with Hamiltonian

$$\mathcal{H} = \int dx (\psi_x^+ \psi_x + c \psi^+ \psi^+ \psi \psi). \quad (1)$$

We consider the case $c > 0$. The canonical commutation relation reads $[\psi(x), \psi^+(y)] = \delta(x - y)$ and in N -particle Fock space we have

$$\mathcal{H} = \sum_{j=1}^N \left(-\frac{d^2}{dx_j^2} \right) + c \sum_{j \neq l} \delta(x_j - x_l). \quad (2)$$

The main observation made in Refs. 1 and 2 is that the equations of the inverse scattering transformation may be taken over to the quantized theory. For our purposes only the operator version of the Zakharov-Shabat eigenvalue problem³ will be needed:

$$\phi_{1x} = -\frac{1}{2}iq\phi_1(x) + \sqrt{c}\phi_2(x)\psi(x), \quad (3)$$

$$\phi_{2x} = +\frac{1}{2}iq\phi_2(x) + \sqrt{c}\psi^+(x)\phi_1(x).$$

The special operator Jost solution (ϕ_1, ϕ_2) we are looking for is defined by the asymptotic behavior

$$x \rightarrow -\infty: \phi_1 \rightarrow e^{-iqx/2}, \quad \phi_2 \rightarrow 0. \quad (4)$$

Then the Jost operators $A(q)$ and $B^+(q)$ are defined as $x \rightarrow +\infty$ by

$$\phi_1 e^{iqx/2} \rightarrow A(q), \quad \phi_2 e^{-iqx/2} \rightarrow \sqrt{2\pi c} B^+(q). \quad (5)$$

The following commutation relations have been established¹:

$$[\mathcal{H}, B^+(q)] = q^2 B^+(q) \quad (6)$$

and

$$[\mathcal{H}, A(q)] = 0,$$

$$A(q)B^+(q') = \left(1 - \frac{ic}{q - q'}\right) B^+(q')A(q). \quad (7)$$

From (6) it follows that the state vectors

$$B^+(q_1) \dots B^+(q_N) |0\rangle \quad (8)$$

are eigenstates of \mathcal{H} with eigenvalue $\sum_i q_i^2$. $|0\rangle$ is the ground state of \mathcal{H} and is identical with the vacuum of the ψ fields.

From (7) and $A(q)|0\rangle = |0\rangle$ it follows that the states (8) are eigenstates of $A(q)$ too. The expansion of the eigenvalue in powers of q^{-1} may be compared with the corresponding expansion of $A(q) = A[q; \psi^+, \psi]$ in order to obtain the eigenvalues of the infinite sequence of operators commuting with \mathcal{H} .

On the other hand, the eigenfunctions of (2) have been calculated in Refs. 4 and 5. Their particular form is known in the literature as "Bethe's hypothesis" or "Bethe-ansatz."⁶

Following Lieb and Liniger,⁴ these eigenfunctions read $\Psi^{(+)}(q_1 \dots q_N | x_1 \dots x_N)$

$$= \frac{1}{(2\pi)^{N/2}} \frac{1}{N!} \sum_{P \in \gamma^N} A(P) E \left(\begin{matrix} x_1 \dots x_N \\ P_1 \dots P_N \end{matrix} \right); \quad (9)$$

here

$$E \left(\begin{matrix} x_1 \dots x_N \\ P_1 \dots P_N \end{matrix} \right) = \sum_{Q \in \gamma^N} \theta(x_{Q_1} - x_{Q_2}, \dots) \theta(x_{Q_2} - x_{Q_3}, \dots) \times \exp\left(-\sum_j i q_j x_{Q_j}\right) \quad (10)$$

denotes the "ordered exponential function" and γ^N is the set of all permutations of the numbers 1 to N .

The coefficients $A(P)$ are products of two-particle S matrices of the form

$$S_{ij} = \frac{i|q_i - q_j| + c}{i|q_i - q_j| - c}, \quad (11)$$

where q_i and q_j are the momenta of the corresponding particles. The precise form of $A(P)$ depends on the permutation P and can be determined from simple rules given in Ref. 4. All $A(P)$'s are uniquely determined up to a common multiplicative constant which we have chosen such that $A(I) = 1$, where I is the identity permutation. The wavefunction obtained in this way represents an outgoing wave.⁵

In this paper we want to prove that the states given by (8) are of the Bethe type and closely related to (9), namely,²

$$B^+(q_1) \dots B^+(q_N) |0\rangle = \prod_{I < J} [1 + Y(q_I, q_J)] |+, q_1 \dots q_N\rangle \quad (12)$$

where

$$Y(q_I, q_J) = ic / |q_I - q_J| \quad (13)$$

and

$$|+, q_1 \dots q_N\rangle = \int dx_1 \dots dx_N \Psi^{(+)}(q_1 \dots q_N | x_1 \dots x_N)$$

$$\times \psi^+(x_1) \cdots \psi^+(x_N) | 0 \rangle. \quad (14)$$

In the nonlinear Schrödinger theory, therefore, "Bethe's hypothesis" appears to be a consequence of the inverse scattering transformation.

THE PROOF

In this section we show that the following relation holds:

$$B^+(K) | +, q_1 \cdots q_N \rangle = \prod_{j=1}^N [1 + Y(K, q_j)] | +, q_1 \cdots q_N K \rangle. \quad (15)$$

By repeated application of (15), Eq. (12) immediately follows.

We start with the explicit representation of $B^+(K)$ in terms of free field operators

$$B^+(K) = \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{\infty} \int d^{j+1} \mathbf{x} d^j \mathbf{y} g_{2j}(\mathbf{x} | \mathbf{y} | K) \times \psi^+(x_1) \cdots \psi^+(x_{j+1}) \psi(y_1) \cdots \psi(y_j), \quad (16)$$

where

$$g_{2j}(\mathbf{x} | \mathbf{y} | K) = c^j \prod_{l=1}^j \theta(x_{l+1} - y_l) \theta(y_l - x_l) \times \exp[-iK(x_1 + x_2 + \cdots + x_{j+1} - y_1 - y_2 - \cdots - y_j)]. \quad (17)$$

Here $d^j \mathbf{x} \equiv dx_1 \cdots dx_j$, $\mathbf{x} = x_1 \cdots x_j$, θ denotes the step function, and in (17) a symmetrization in \mathbf{x} and \mathbf{y} is to be included.

Using (16), the lhs of (15) in Fock space reads

$$\begin{aligned} \mathcal{F}(x_1 \cdots x_{N+1}) &= (2\pi)^{-(N+1)/2} \sum_{j=0}^N S[(N-j)!(N+1)]^{-1} \\ &\times \int d^j \mathbf{y} \Psi^{(+)}(q_1 \cdots q_N | x_1 \cdots x_{N-j} y_1 \cdots y_j) \\ &\times g_{2j}(x_{N-j+1} \cdots x_{N+1} | \mathbf{y} | K). \end{aligned} \quad (18)$$

Here the symbol S denotes symmetrization in $x_1 \cdots x_{N+1}$. In what follows we consider the case

$$x_1 < x_2 < \cdots < x_N < x_{N+1} \quad (19)$$

which simplifies the considerations but imposes no restriction. Since $\Psi^{(+)}(\mathbf{q} | \mathbf{x})$ and $g_{2j}(\mathbf{x} | \mathbf{y} | K)$ are themselves symmetric, S may be expressed as a sum over ordered subsets from $j+1$ numbers out of $N+1$, i.e.,

$$\alpha_{N+1}^{j+1} = \{\alpha_1, \alpha_2, \dots, \alpha_{j+1}\}, \quad (20)$$

where $\alpha_i \in \{1, \dots, N+1\}$ and $\alpha_1 < \alpha_2 < \cdots < \alpha_{j+1}$. The ordered complement of α_{N+1}^{j+1} is denoted by

$$\beta(\alpha_{N+1}^{j+1}) = \{1, \dots, N+1\} / \alpha_{N+1}^{j+1} = \{\beta_1, \dots, \beta_{N-j}\} \quad (21)$$

and $\beta_1 < \beta_2 < \cdots < \beta_{N-j}$. We will write $\alpha_{N+1}^{j+1} \equiv \alpha$ since the dependence on j and N is obvious in what follows. Using (16), (17), and (9), we may write for the lhs of (15)

$$\mathcal{F}(x_1, \dots, x_{N+1}) = \sum_{P \in \mathcal{Y}^N} \sum_{j=0}^N \sum_{\alpha} A(P) I_j(\alpha, P), \quad (22)$$

where

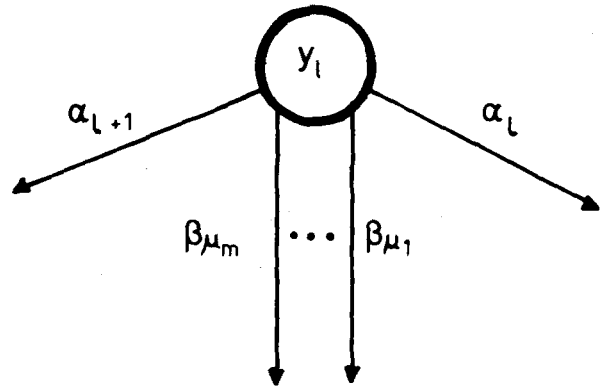


FIG. 1. Graphical element representing the integration over y_l . The various terms resulting from this integration are symbolized by arrows.

$$I_j(\alpha, P) = c^j \int_{x_{\alpha_1}}^{x_{\alpha_2}} dy_1 \cdots \int_{x_{\alpha_j}}^{x_{\alpha_{j+1}}} dy_j E \begin{pmatrix} x_{\beta_1} & \cdots & x_{\beta_{N-j}} & y_1 & \cdots & y_j \\ P_1 & \cdots & P_N & & & \end{pmatrix} \times \exp[-iK(x_{\alpha_1} + \cdots + x_{\alpha_{j+1}} - y_1 - \cdots - y_j)]. \quad (23)$$

$E(\cdots)$ stems from the N -particle Bethe function in (18), the exponential function and the limits of integration follow from g_{2j} .

We have to show that $\mathcal{F}(x_1, \dots, x_{N+1})$ is the $(N+1)$ -particle Bethe function times the factor $\prod [1 + Y(K, q_j)]$ appearing in (15). To facilitate the comparison, it is useful to manipulate the $(N+1)$ -particle Bethe function in the following way: Extract from $A(P)$ where $P \in \mathcal{Y}^{N+1}$ all factors containing $q_{N+1} \equiv K$ and rewrite the sum over permutations as

$$\sum_{P \in \mathcal{Y}^{N+1}} f(P) = \sum_{P \in \mathcal{Y}^N} \sum_{i=1}^{N+1} f(P\sigma_i), \quad (24)$$

where σ_i is a suitable transposition. Then the remaining $A(P)$ are identical to those appearing in (22). Finally we multiply by $\prod [1 + Y(K, q_j)]$ to obtain

$$\begin{aligned} &\prod_{j=1}^N [1 + Y(K, q_j)] \Psi^{(+)}(q_1 \cdots q_N K | \mathbf{x}) \\ &= \sum_{P \in \mathcal{Y}^N} \sum_{j=0}^N \sum_{\bar{\alpha}} \sum_{l=1}^{N+1} A(P) Y_{\bar{\alpha}_1} \cdots Y_{\bar{\alpha}_j} \cdot \text{sig}(\bar{\alpha} | l) \\ &\times E \begin{pmatrix} x_1 & \cdots & \overset{l}{\vee} & \cdots & x_{N+1} \\ P_1 & \cdots & P_N & & \end{pmatrix} e^{-iKx_l}, \end{aligned} \quad (25)$$

where we have abbreviated

$$Y_j = Y(K, q_j) \quad (26)$$

and $\overset{l}{\vee}$ means that the variable x_l does not occur in the symbol $E(\cdots)$, $\bar{\alpha}$ denotes the subsets of $\{1, \dots, N\}$, and

$$\text{sig}(\bar{\alpha} | l) = \begin{cases} 1 & \text{if the number of } \bar{\alpha}_j \text{ with } \bar{\alpha}_j \geq l \text{ is even,} \\ -1 & \text{otherwise.} \end{cases} \quad (27)$$

Comparing (22) and (25), we have reduced the problem to showing the identity




TABLE I. Rules for finding the analytical expression of a term that corresponds to a given descent.

(i) Write down

$$E \begin{pmatrix} x_{\beta_1} \dots x_{\beta_{j-1}}, y_1 \dots y_j \\ P_1 \dots P_N \end{pmatrix}$$

(ii) For each step in the descent, substitute and multiply according to the scheme below.

(iii) The resulting expressions has to be multiplied by $\exp[-iK(x_{\alpha_1} + x_{\alpha_2} + \dots + x_{\alpha_{j+1}})]$.

Element	Substitute	Multiply
	$y_k \rightarrow x_{\alpha_k}$	$Y_{\alpha_k} \exp(iKx_{\alpha_k})$
	$y_k \rightarrow x_{\beta_{\mu}}$	$(Y_{\beta_{\mu}} - Y_{\beta_{\mu}-1}) \exp(iKx_{\beta_{\mu}})$
	$y_k \rightarrow x_{\alpha_{k+1}}$	$-Y_{\alpha_{k+1}-1} \exp(iKx_{\alpha_{k+1}-1})$

$$\sum_{\alpha} I_j(\alpha, P) = \sum_{l=1}^{N+1} \sum_{\bar{\alpha}} Y_{\bar{\alpha}_1} \dots Y_{\bar{\alpha}_j} \text{sig}(\bar{\alpha}, l) \times E \begin{pmatrix} x_1 \dots \overset{l}{\vee} \dots x_{N+1} \\ P_1 \dots P_N \end{pmatrix} e^{-iKx_l} \quad (28)$$

i.e., the main problem consists of doing all the integrations appearing in $I_j(\alpha, P)$ [cf. (23)].

Every integration in (23) can be done in a straightforward manner, but due to the presence of the θ functions in $E(\dots)$ a lot of terms arise. It turns out, however, that every term can be uniquely identified by the elements of the subsets α and β .⁷ The terms resulting from an integration over say y_l may be represented symbolically as in Fig. 1, where $\alpha_l < \beta_{\mu_1} < \dots < \beta_{\mu_m} < \alpha_{l+1}$. Since the integrand is essentially an exponential function, the resulting terms are of the form $E(\dots)$ times a factor, where in $E(\dots)$ y_l has now been replaced by

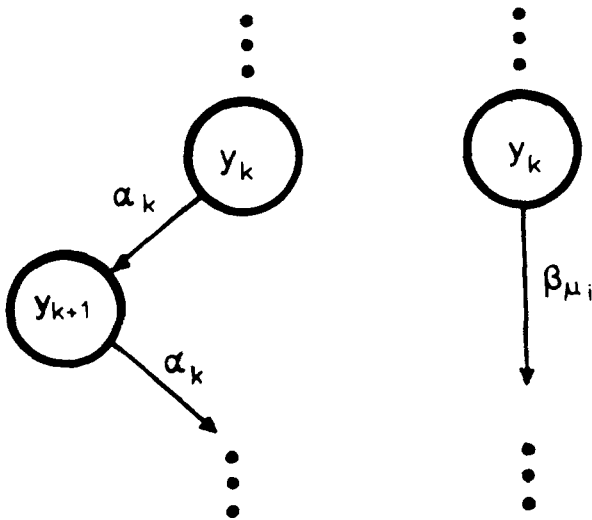


FIG. 2. Situations which lead to unwanted terms.

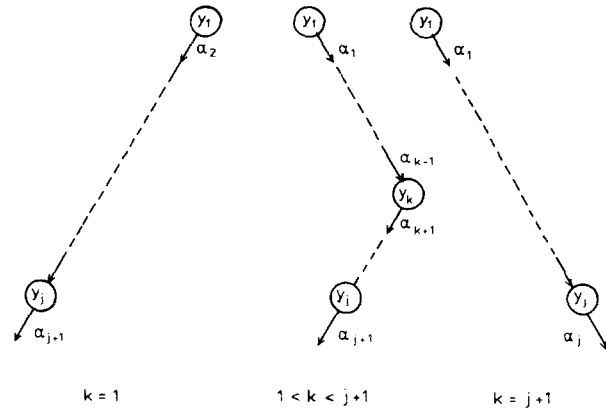


FIG. 3. Descents of the terms which contribute to the lhs of (28).

one of the x_j 's. The precise form of the substitutions and the factors occurring for each arrow in Fig. 1 can be found by the rules in Table I.

The carrying through of all integrations can be visualized as a chaining together of graphical elements given in Fig. 1. In this way a "genealogical tree" for all terms appearing in $I_j(\alpha, P)$ is obtained. Every term can be uniquely identified by its "descent," the corresponding analytical expression is given by applying the rules of Table I for every step in this descent.

We may write symbolically

$$I_j(\alpha, P) = \sum (\text{"last born" terms}). \quad (29)$$

Next we observe that in (29) terms occur which are "not wanted" in (28). These are all terms in whose descent at least one of the situations of Fig. 2 occurs.

Using the rules in Table I, one finds that this leads to a factor

$$\sim \exp[-i(q_m + q_n)x_{\alpha}], \quad (30)$$

i.e., the variable x_{α} occurs twice in $E(\dots)$. Using the techniques developed in this section, one may show that all terms of this kind vanish from (22), so they are not present in (28). We do not present here the complete procedure. In the Appendix we describe the essential ideas by a special example and outline the general case.

Now it is easy to see, that the graphical equivalent for all "wanted" terms are the $j+1$ cases depicted and characterized by an index k in Fig. 3.

The analytical expression for the sum of all these contributions is

$$\sum_{k=1}^{j+1} Y_{\alpha_1} \dots Y_{\alpha_{k-1}} Y_{\alpha_{k+1}-1} \dots Y_{\alpha_{j+1}-1} (-1)^{j+k-1} \times E \begin{pmatrix} x_1 \dots \overset{\alpha_k}{\vee} \dots x_{N+1} \\ P_1 \dots P_N \end{pmatrix} \exp(-iKx_{\alpha_k}) \quad (31)$$

Doing the sum over all α in (31), it is a matter of minor manipulations to arrive at the rhs of (28).

Hence (15) and consequently (12) are proven.

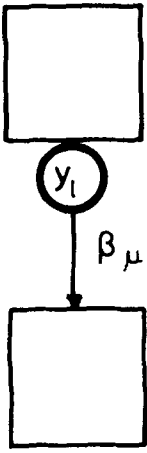


FIG. 4. A term where x_{β_μ} occurs twice in $E(\dots)$. Parts of the diagram are symbolized by boxes.

ACKNOWLEDGMENT

I would like to thank Professor J. Honerkamp for discussions and encouragement.

APPENDIX

We describe here the essential ideas how one can see that the "unwanted" terms cancel in (22). We discuss the case when only *one* variable, say x_{β_μ} , occurs twice in the symbol $E(\dots)$:

$$E \left(\begin{matrix} \dots x_{\beta_\mu} x_{\beta_\mu} \dots \\ P_1 \quad \dots \quad P_N \end{matrix} \right) \sim \exp[-i(q_{P_\mu} + q_{P'_\mu})x_{\beta_\mu}], \quad (\text{A1})$$

where we have identified $\beta = \beta_\mu$, $\beta' = \beta_\mu - 1$.

A term of this kind shows in its descent one of the situations given in Fig. 2. Specifying

$$\{\alpha, j, P\}, \quad (\text{A2})$$

we consider a term in whose descent the second situation of Fig. 2 occurs for the variable x_{β_μ} (see Fig. 4). Using Table I, one finds that it has the form

$$R(Y_\beta - Y_{\beta'}), \quad (\text{A3})$$

where R contains all factors which do not alter in what follows.

Next one looks for terms which lead to the same R . Because of (A1), R is symmetric in q_{P_β} and $q_{P_{\beta'}}$. Therefore, the diagram with the choice (A2) but P replaced by P' , where P' differs from P by the exchange $q_{P_\beta} \leftrightarrow q_{P_{\beta'}}$, again yields (A3).

Next, one associates with (A2) the following choice:

$$\{\tilde{\alpha} = \alpha \cup \{\beta_\mu\}, j+1, P\}, \quad (\text{A4})$$

where now the first situation of Fig. 2 occurs for x_{β_μ} but the remaining steps in the descent are unchanged (see Fig. 5). This term has the following form:

$$(-1)R Y_\beta Y_{\beta'}. \quad (\text{A5})$$

Again, the term with the same diagram but P replaced by P' yields (A5) too.

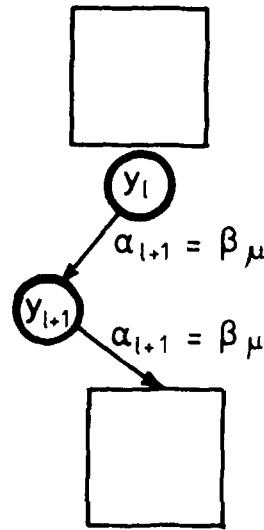


FIG. 5. The term associated with Fig. 4 via (A4). The boxes symbolize the same parts of the diagram as in Fig. 4.

These four terms in (22) can now be collected to yield the contribution

$$\{A(P) + A(P')\}R(Y_\beta - Y_{\beta'}) + \{A(P) + A(P')\}(-1)R Y_\beta Y_{\beta'}. \quad (\text{A6})$$

From the rules determining the $A(P)$'s,⁴ one finds

$$A(P') = \frac{i(q_{P_\beta} - q_{P_{\beta'}}) + c}{i(q_{P_\beta} - q_{P_{\beta'}}) - c} A(P); \quad (\text{A7})$$

therefore, (A6) reads

$$A(P)R \cdot \{Y_\beta - Y_{\beta'} - Y_\beta Y_{\beta'} + \frac{i(q_{P_\beta} - q_{P_{\beta'}}) + c}{i(q_{P_\beta} - q_{P_{\beta'}}) - c} [Y_\beta - Y_{\beta'} - Y_\beta Y_{\beta'}]\}. \quad (\text{A8})$$

It is easy to see that the curly bracket vanishes.

In the general case there are several variables occurring twice in $E(\dots)$, but one proceeds in the same way: One starts with a term of the type of Fig. 4 (with several x_{β_μ} 's) and considers all possible replacements $P \rightarrow P'$ and from Fig. 4 to Fig. 5, as above, where, of course, many combinations are possible. Then the corresponding "curly bracket" is shown to vanish. The main work to be done consist merely in formulating and bookkeeping. The graphical description presented above is of great use for this work.

¹E. K. Sklyanin, Dokl. Akad. Nauk **244**, 1337 (1979); L. D. Faddeev, Leningrad Preprint P-2-1979; H. B. Thacker and D. Wilkinson, Phys. Rev. D **19**, 3660 (1979).

²J. Honerkamp, P. Weber, and A. Wiesler, Nucl. Phys. B **152**, 266 (1979).

³V. E. Zakharov and A. B. Shabat, Zh. Eksp. Teor. Fiz. **61**, 118 (1971) [Sov. Phys. JETP **34**, 62 (1972)].

⁴E. H. Lieb and W. Lininger, Phys. Rev. **130**, 1605 (1963).

⁵C. N. Yang, Phys. Rev. Lett. **19**, 1312 (1967); Phys. Rev. **168**, 1920 (1968).

⁶H. A. Bethe, Z. Phys. **71**, 205 (1931).

⁷A. Wiesler, thesis, Freiburg, June 1979.

Hypervirial relations and orthogonalization conditions

Francisco M. Fernández and Eduardo A. Castro

INIFTA, Sección Química Teórica, Sucursal 4—Casilla de Correo 16, La Plata 1900, Argentina

(Received 16 July 1980; accepted 7 November 1980)

Through the imposition of orthogonalization conditions, a number of results are derived for wavefunctions and eigenvalues in the framework of hypervirial methodology. Some previous results are extended and various similar developments of off-diagonal Hellmann–Feynman and sum rule formulas are indicated.

PACS numbers: 03.65.Ob

I. INTRODUCTION

Practical applications of diagonal hypervirial formulas began with communications of Hirschfelder,¹ Epstein and Hirschfelder,² and Hirschfelder and Coulson.³ The relationship between diagonal hypervirial formulas and the variational theorem was analyzed in Refs. 2 and 3. Later, Chen⁴ introduced the formalism of unitary operators to study hypervirial relations. Robinson^{5,6} showed the importance of diagonal hypervirial relations within the context of perturbational theory. The usefulness of off-diagonal hypervirial formulas to compute spectroscopic constants was pointed out by Chen.⁷ Coulson⁸ introduced the off-diagonal hypervirial relations as an alternative method for obtaining approximate eigenfunctions and eigenvalues. Bradley and Hughes^{9,10} extended such methodology for periodic potentials and general hypervirial operators. The present authors¹¹ generalized the analysis given in Ref. 9 and 10. Furthermore, the concerted utilization of diagonal and off-diagonal hypervirial relations was given from a practical point of view,¹² and then it was discussed from a theoretical standpoint.¹¹ Klein and DeVries¹³ gave sufficient conditions for variational wavefunctions and energies to satisfy some off-diagonal hypervirial formulas which exact wavefunctions and associated eigenvalues satisfy.

Let H be an Hermitian operator which is defined over the Hilbert space V , and let \mathcal{V}^A , \mathcal{V}_A ($\mathcal{V}^A \supseteq \mathcal{V}_A$), and \mathcal{V}_B be three closed subspaces of V . If X is a linear operator and $|a\rangle \in \mathcal{V}_A$, $|b\rangle \in \mathcal{V}_B$, such that

$$\mathcal{P}^A(H - E_a)|a\rangle = 0, \quad (1)$$

$$\mathcal{P}_B(H - E_b)|b\rangle = 0, \quad (2)$$

$$\mathcal{V}^A \supseteq X\mathcal{V}_B, \quad X^+\mathcal{V}_A \subseteq \mathcal{V}_B. \quad (3)$$

Then, Klein and DeVries showed that the hypervirial relations

$$\langle a|[H, X]|b\rangle = (E_a - E_b)\langle a|X|b\rangle \quad (4)$$

are satisfied. \mathcal{P}^A and \mathcal{P}_B are projection operators onto \mathcal{V}^A and \mathcal{V}_B respectively. It is interesting to note in Eq. (4) that neither $|a\rangle$ nor $|b\rangle$ is required to be eigenket to H . The result (4) can be obtained when, from given X and \mathcal{V}_A , one chooses

$$\mathcal{V}_B = X^+\mathcal{V}_A, \quad (5)$$

and through an appropriate selection of linear parameters, Eq. (1)–(2) are satisfied; and from nonlinear parameters the

condition

$$\mathcal{P}_L(H - E_a)|a\rangle = 0, \quad L = \mathcal{V}^A - \mathcal{V}_A \quad (6)$$

is fulfilled. Furthermore, it is well known that in order that the functional energy

$$E(\psi) = \langle \psi|H|\psi\rangle / \langle \psi|\psi\rangle \quad (7)$$

be stable regarding a variation $\delta\psi$, then the equation

$$\langle (H - E)\psi|\delta\psi\rangle = 0 \quad (8)$$

must be satisfied. The previous results are generalized for two wavefunctions ψ_1 and ψ_2 in the following way:

For

$$\delta\psi_1 = W\psi_2, \quad \delta\psi_2 = W\psi_1, \quad W^+ = W \quad (9)$$

and

$$\langle (H - E_1)\psi_1|W\psi_2\rangle = \langle (H - E_2)\psi_2|W\psi_1\rangle = 0, \quad (10)$$

then

$$\langle \psi_1|[H, W]|\psi_2\rangle = (E_1 - E_2)\langle \psi_1|W|\psi_2\rangle. \quad (11)$$

From Eqs. (1), (2), (6), (8), and (10) we can see the key role played by the fact that the vector $(H - E)\psi$ is allowed to be orthogonal to a given subspace. When ψ is an eigenfunction of H with eigenvalue E , then

$$(H - E)\psi \perp V, \quad \text{or} \quad (H - E)\psi = 0. \quad (12)$$

Exact eigenfunctions of H and associated eigenvalues can not be obtained for general cases of physical interest. As a consequence, we shall restrict ourselves to the case

$$(H - E)\psi \perp U, \quad (13)$$

where U is an arbitrary closed subspace of V , and E is given by Eq. (7).

We present here the results for different choices of U , as well as the connection with other previous conclusions given by different authors. It will be assumed that operators are defined on the total space V throughout all the following in order to avoid difficulties regarding domain and range.

II. DIAGONAL HYPERVIRIAL RELATIONS AND ORTHOGONALITY CONDITIONS

Let V be a Hilbert space, U a closed subspace of V , and H an Hermitian operator on V . Given $\psi \in V$, we search for $\lambda \in R$ such that

$$\mathcal{P}_U(H - \lambda)\psi = 0, \quad (14)$$

where \mathcal{P}_U is a projection operator onto U . Then, the following known results are obtained:

(a) For $\delta\psi \in U$ a variation over ψ and $\lambda = E(\psi)$ given by Eq. (7), then

$$\begin{aligned} \delta E \langle \psi | \psi \rangle &= [\delta\psi | (H - E)\psi] + \langle (H - E)\psi | \delta\psi \rangle \\ &= \langle \delta\psi | \mathcal{P}_U (H - E)\psi \rangle + \text{c.c.} = 0, \end{aligned} \quad (15)$$

where c.c. denotes the complex conjugate term. Eq. (15) means that under preceding assumptions, E is an extremum for the variations $\psi \rightarrow \psi + \delta\psi$.

(b) If, in addition, $\psi \in U$, then

$$\mathcal{P}_U H \mathcal{P}_U \psi = E\psi. \quad (16)$$

When U is finitely generated by the basis set $\{\phi_i, i = 1, 2, \dots, n\}$, then

$$\mathcal{P}_U = \sum_{i=1}^n |\phi_i\rangle \langle \phi_i|$$

and Eq. (16) gives us the variational method of Rayleigh-Ritz

(c) For $\psi \in V$ and $H\psi \in U$, Eq. (14) gives

$$\langle H\psi | (H - \lambda)\psi \rangle = \|H\psi\|^2 - \lambda \langle \psi | H\psi \rangle = 0, \quad (17)$$

so $\lambda \in \mathbb{R}$. The condition $\lambda = E$ is fulfilled iff $H\psi = E\psi$, which is proven as follows:

(\Leftarrow) If $\lambda = E$, then

$$\langle \psi | (H - E)\psi \rangle = 0,$$

so

$$\begin{aligned} \langle H\psi | (H - E)\psi \rangle &= \langle (H - E)\psi | (H - E)\psi \rangle \\ &= \|(H - E)\psi\|^2 = 0. \end{aligned}$$

(\Rightarrow) It is immediate that if $H\psi = E\psi$, then $\lambda = E$.

(d) if v is an anti-Hermitian operator over V and $W = v + r$, $r \in \mathbb{R}$, then

$$\begin{aligned} \langle (H - \lambda)\psi | W\psi \rangle + \text{c.c.} &= \langle H\psi | v\psi \rangle + r \langle (H - \lambda)\psi | \psi \rangle \\ &+ \text{c.c.} \end{aligned} \quad (18)$$

From Eq. (18) we can see that if conditions $\lambda = E$ and $(H - E)\psi \perp W\psi$ are satisfied, then the diagonal hypervirial formula

$$\langle \psi | [H, v] | \psi \rangle = 0 \quad (19)$$

is valid.

(e) the analysis of orthogonality conditions carry us to obtain the results previously deduced by Robinson^{5,6} in a wholly independent way, too. Let $\phi^0(a, X) \in V$ be a function which depends on a parameter a , and H^0, H two Hermitian operators over V , such that

$$H^0 \phi^0 = E^0 \phi^0, \quad (20)$$

$$H = H^0 + H^{(1)}. \quad (21)$$

Then, the first order correction equation for the wavefunction is

$$(H^0 - E^0)\phi^{(1)} = (E^{(1)} - H^{(1)})\phi^0. \quad (22)$$

Equation (22) is valid for any a -value. For the choice $\lambda = E = E^0 + E^{(1)}$, and the parameter a chosen in such a way that the orthogonality condition

$$(H - E)\phi^0 \perp \chi, \quad \chi \in V \quad (23)$$

is satisfied, then, from Eqs. (22)–(23)

$$\langle (H^{(1)} - E^{(1)})\phi^0 | \chi \rangle = 0, \quad (24)$$

or, equivalently

$$\langle (H^0 - E^0)\phi^{(1)} | \chi \rangle = 0. \quad (25)$$

If the functions ϕ^0 and $\phi = \phi^0 + \phi^{(1)}$ are normalized, then

$$\langle \phi^0 | \phi^{(1)} \rangle + \text{c.c.} = 0. \quad (26)$$

Let us consider another Hermitian operator L over V , which satisfies the condition

$$(H^0 - E^0)\chi = (L - L^0)\phi^0, \quad (27)$$

with $L^0 = \langle \phi^0 | L \phi^0 \rangle$.

The first-order improvement for the expectation value of L in the E -eigenstate of H is

$$\begin{aligned} L^{(1)} = \langle \phi^0 | L \phi^{(1)} \rangle + \text{c.c.} &= \langle (L - L^0)\phi^0 | \phi^{(1)} \rangle + \text{c.c.} \\ &= \langle (H^0 - E^0)\chi | \phi^{(1)} \rangle + \text{c.c.} \\ &= 0. \end{aligned} \quad (28)$$

When $\chi = \partial\phi^0/\partial a$, we obtain the results presented in Ref. 5, while for $\chi = W\phi^0$, W defined as in (d), the more general results of Ref. 6 are gotten. These results are extended without difficulty if ϕ^0 depends on n parameters a_1, a_2, \dots, a_n . In this case such parameters are determined through the condition (14) when $\lambda = E$ and U is spanned by n linearly independent vectors $\{\chi_1, \dots, \chi_n\}$. Eqs. (24), (25), (27) adopt the more general following form:

$$\langle (H^{(1)} - E^{(1)})\phi^0 | \chi_i \rangle = 0, \quad i = 1, 2, \dots, n, \quad (29)$$

$$\langle (H^0 - E^0)\phi^{(1)} | \chi_i \rangle = 0, \quad i = 1, 2, \dots, n, \quad (30)$$

$$(H^0 - E^0)\chi_i = (L_i - L_i^0)\phi^0, \quad i = 1, 2, \dots, n, \quad (31)$$

and, as particular cases:

$$\chi_i = \partial\phi^0/\partial a_i, \quad i = 1, 2, \dots, n, \quad (32)$$

or

$$\chi_i = W_i \phi^0, \quad i = 1, 2, \dots, n. \quad (33)$$

III. OFF-DIAGONAL HYPERVIRIAL RELATIONS AND ORTHOGONALITY CONDITIONS

Let V be a Hilbert space, U and U' two closed subspaces of V , V an Hermitian operator, and W a linear operator. H and W are defined onto V . If, for given $\psi \in U$, and $\psi' \in U'$ there exist λ and $\lambda' \in \mathbb{R}$ such that

$$(H - \lambda)\psi \perp WU', \quad (34)$$

$$(H - \lambda')\psi' \perp W + U, \quad (35)$$

then

$$\begin{aligned} \langle \psi | [H, W] | \psi' \rangle &= \langle H\psi | W\psi' \rangle - \langle W^+ \psi | H\psi' \rangle \\ &= (\lambda - \lambda') \langle \psi | W\psi' \rangle. \end{aligned} \quad (36)$$

Various results emerge from Eq. (36):

(a) Let us suppose that U and U' are spanned by the basis set $\{\phi_1, \dots, \phi_n\}$ and $\{\phi'_1, \dots, \phi'_n\}$ respectively. Furthermore we assume that W fulfills the condition

$$\dim(WU') = \dim(W + U) = n.$$

Then, for

$$\psi = \sum_{i=1}^n c_i \phi_i,$$

$$\psi' = \psi = \sum_{i=1}^n c'_i \phi'_i \quad (37)$$

Eqs. (34)–(35) give us

$$\sum_{i=1}^n c_i \{ \langle H \phi_i | W \phi'_j \rangle - \lambda \langle \phi_i | W \phi'_j \rangle \} = 0, \quad j = 1, 2, \dots, n, \quad (38)$$

$$\sum_{i=1}^n c'_i \{ \langle H \phi'_i | W^+ \phi_j \rangle - \lambda' \langle \phi'_i | W^+ \phi_j \rangle \} = 0, \quad j = 1, 2, \dots, n. \quad (39)$$

If there exist $\lambda_1, \dots, \lambda_k, \lambda'_1, \dots, \lambda'_k \in \mathbb{R}$ corresponding to $\psi_1, \dots, \psi_k \in U, \psi'_1, \dots, \psi'_k \in U'$ which satisfy Eqs. (38–39); then

$$\langle \psi_i | [H, W] | \psi'_j \rangle = (\lambda_i - \lambda'_j) \langle \psi_i | W | \psi'_j \rangle; \quad i = 1, \dots, k, \quad j = 1, \dots, k'. \quad (40)$$

Conversely, if $\psi \in U, \lambda \in \mathbb{R}$ satisfy Eq. (34) and there exist $\lambda' \in \mathbb{R}, \psi' \in U'$ that satisfy Eq. (36), then it follows that $(H - \lambda')\psi' \perp W^+\psi$. Equations (34)–(36) or (38)–(40) are more general than usual hypervirial relations because neither ψ nor ψ' are required to be eigenfunctions of H .

(b) For $U' = W^+U$, then

$$(H - \lambda)\psi \perp WW^+U, \quad (41)$$

$$(H - \lambda')\psi' \perp U'. \quad (42)$$

From Eq. (4.2) we can see that ψ' is stable with respect to H restricted to U' , i.e.,

$$\mathcal{P}_{U'} H \mathcal{P}_{U'} \psi' = \lambda' \psi'. \quad (43)$$

But if ψ_i, ψ_j satisfy Eq. (41) with $\lambda_i, \lambda_j \in \mathbb{R}$, then

$$\langle \psi_i | [H, WW^+] | \psi_j \rangle = (\lambda_i - \lambda_j) \langle \psi_i | WW^+ | \psi_j \rangle. \quad (44)$$

Equation (44) corresponds to the results mentioned in Ref. 14.

(c) Previous results lead us to give the following

Theorem: If $\{\psi_1, \dots, \psi_n\}$ satisfy condition (41) and $\psi' \in U'$ fulfills hypervirial relations

$$\langle \psi_i | [H, W] | \psi' \rangle = (\lambda_i - \lambda') \langle \psi_i | W \psi' \rangle, \quad i = 1, \dots, n, \quad (45)$$

then ψ' is stable with respect to H restricted to U' , i.e., ψ' satisfies Eq. (43).

Demonstration: From Eq. (45) we have

$$\begin{aligned} \langle \psi_i | [H, W] | \psi' \rangle &= \langle H \psi_i | W \psi' \rangle - \langle W^+ \psi_i | H \psi' \rangle \\ &= \lambda_i \langle \psi_i | W \psi' \rangle - \langle W^+ \psi_i | H \psi' \rangle \\ &= \lambda_i \langle \psi_i | W \psi' \rangle - \lambda' \langle \psi_i | W \psi' \rangle, \\ & \quad i = 1, 2, \dots, n. \end{aligned}$$

Then

$$\langle W^+ \psi_i | (H - \lambda') \psi' \rangle = 0, \quad i = 1, 2, \dots, n. \quad (46)$$

Corollary I. If $W^+U = V$, then $H\psi' = \lambda' \psi'$.

Corollary II. If U' is invariant under H then $H\psi' = \lambda' \psi'$.

The theorem and corollaries extend previous results of the authors.¹⁵

(d) When we add the condition

$$\mathcal{P}_{U'} H \mathcal{P}_{U'} \psi = \lambda \psi \quad (47)$$

to those of (b), then ψ and ψ' will satisfy Eq. (36) and λ, λ' will be the eigenvalues of Eqs. (43)–(44), respectively. This case was discussed at full length by Klein and DeVries.¹³

IV. DISCUSSION

We have seen that the only considerations of orthogonality allow us to obtain a number of interesting general results for wavefunctions and eigenvalues. Particularly significant are those discussed in (c) because it is not necessary to know in advance exact wavefunctions as it is required in previous treatments.^{8–11} Moreover, the proposed methodology could be extended in a natural and feasible way to the analysis of off-diagonal Hellmann–Feynman and sum rule formulas. Such formulas, as well as hypervirial relations, are necessary in the computation and study of electronic state changing transitions, polarizability and Rayleigh and Raman photon scattering.

¹J. O. Hirschfelder, J. Chem. Phys. **33**, 1462 (1960).

²S. T. Epstein and J. O. Hirschfelder, Phys. Rev. **123**, 1496 (1961).

³J. O. Hirschfelder and C. A. Coulson, J. Chem. Phys. **36**, 941 (1962).

⁴J. C. Y. Chen, J. Chem. Phys. **39**, 3167 (1963).

⁵P. D. Robinson, Proc. R. Soc. London **82**, 659 (1963).

⁶P. D. Robinson, Proc. R. Soc. London, Sect. A **283**, 229 (1965).

⁷J. C. Y. Chen, J. Chem. Phys. **38**, 283 (1963); **40**, 615 (1964).

⁸C. A. Coulson, Quart. J. Math. Oxford, **16**, 279 (1965).

⁹C. J. Bradley and D. E. Hughes, Int. J. Quantum Chem. **1S**, 687 (1967).

¹⁰C. J. Bradley and D. E. Hughes, Int. J. Quantum Chem. **3**, 699 (1969).

¹¹F. M. Fernández and E. A. Castro, Int. J. Quantum Chem. **17**, 609 (1980).

¹²E. A. Castro, Int. J. Quantum Chem. **14**, 231 (1978).

¹³D. J. Klein and P. L. DeVries, J. Chem. Phys. **68**, 160 (1978).

¹⁴J. O. Hirschfelder, W. B. Brown, and S. T. Epstein, Adv. Quantum Chem. **1**, 264 (1964).

¹⁵F. M. Fernández and E. A. Castro, J. Chem. Phys. **73**, 4711 (1980).

A semiclassical treatment of path integrals for the spin system

Hiroshi Kuratsuji and Yutaka Mizobuchi

Department of Physics, Kyoto University, Kyoto 606, Japan

(Received 24 June 1980; accepted for publication 26 November 1980)

Starting with path integrals in the SU(2) coherent state representation, the semiclassical approximation of the propagator for the spin system is investigated. By extending the idea of the semiclassical expansion method, which was developed in the usual phase-space path integrals, to the path integrals in the curved phase space, which is characteristic of the SU(2) coherent states, we obtain a closed form for the semiclassical propagator. As an application, we discuss the semiclassical quantization condition for the spin system.

PACS numbers: 03.65.Ob, 03.65.Sq

1. INTRODUCTION

One of the significant advantages of a path integral formulation of quantum mechanics lies in the fact that the propagator (or transition amplitude) expressed in the form of a path integral can be systematically calculated within the semiclassical approximation. Among numerous works on this subject, one of the most noteworthy is the semiclassical analysis developed by Gutzwiller¹ for bound state problems which has been extensively applied to nonlinear field theory models by Dashen, Hasslacher, and Neveu.²

The path integral formulation has so far been carried out by using two different approaches: One is the Lagrangian formulation and the other is the Hamiltonian (or phase space) formulation. Besides these conventional ones, there is another approach, namely the "coherent state path integral" formulation, which was pioneered by Klauder and proved to be particularly suited for describing quantum dynamics of Bose systems.³ Although the coherent state path integral is formally considered as an alternative to the conventional phase-space path integral, the former has a crucial advantage since it can be extended to wider class of physical systems through the use of "generalized coherent states."⁴ One of the simplest but important examples is the spin system, the quantum dynamics of which is successfully described by the path integrals in the SU(2) (or spin) coherent state representation.⁵ In Ref. 5(b) it has been shown that the propagator joining the SU(2) coherent states is cast into the path integral in a generalized phase space and that in the classical limit one arrives at a classical dynamics in a "curved phase space" (\simeq two-dimensional sphere S^2), which is a natural generalization of the usual Hamiltonian dynamics.

As for a path integral for spin, Schulman formulated it as a Lagrangian path integral on the group manifold SU(2).⁶ The key concept of his theory is that the configuration space giving rise to a spin is the group SU(2) itself and the spin propagator is completely given by the geodesics (the extrema of the action) on the group manifold SU(2) (\simeq three-dimensional sphere S^3).

The purpose of the present paper is to put forward a semiclassical analysis of the propagator for the spin system expressed by the path integral for the SU(2) coherent states. We shall treat the semiclassical approximation of the path integral on the "phase space" SU(2)/U(1), which is the homogeneous space of SU(2), in contrast to the path integral on

the group manifold. To investigate this problem would be worthwhile since it would provide us with an analytic device for obtaining the bound state spectra of the spin and/or analogous systems which one encounters frequently in actual physical problems. The main aim is to get a closed form for the semiclassical propagator which consists of the dominant part with the classical action and the "reduced propagator" coming from the second variation of the action functional around the classical path in the curved phase space. To do this, we extend the idea of the "path expansion method", which was developed by Levit and Smilansky⁷ in the semiclassical analysis of the conventional Hamiltonian path integrals, to path integrals in curved phase space. The essential point of the calculational procedure lies in that the reduced propagator for the curved phase space is transcribed into the one for the *flat* phase space by a simple transformation and we can thereby utilize the technique developed in Ref. 7.

In the next section we will recapitulate the essence of the path integral formulation in the SU(2) coherent states and discuss its classical limit. In Sec. 3 we will derive the formula for the semiclassical propagator. Section 4 is devoted to the derivation of the semiclassical quantization condition for the spin system as an application of the formula obtained in Sec. 3. In Sec. 5 we will give additional remarks.

2. PATH INTEGRAL FORMULATION FOR THE SPIN SYSTEM

We recapitulate the necessary ingredients for the path integral representation of the propagator for the spin system expressed in the SU(2) coherent states, which has been formulated in Ref. 5(b).

We start with the SU(2) coherent states defined by

$$|\zeta\rangle = \exp(\mu\hat{J}_+ - \mu^*\hat{J}_-)|0\rangle \\ = (1 + \zeta^*\zeta)^{-J} \exp[\zeta\hat{J}_+]|0\rangle \quad (J = \frac{1}{2}, 1, \dots) \quad (2.1)$$

with

$$\zeta = (\mu/|\mu|)\tan|\mu|,$$

for arbitrary complex numbers ζ , where $|0\rangle$ denotes the eigenstate of \hat{J}_z with the minimum eigenvalue $-J$, and $\hat{J}_\pm (= \hat{J}_x \pm i\hat{J}_y)$ and \hat{J}_z are the SU(2) generators. The system $\{|\zeta\rangle\}$ has the completeness relation

$$\int |\zeta\rangle d\mu(\zeta) \langle\zeta| = 1, \quad (2.2)$$

which holds for an irreducible representation of each J . The SU(2)-invariant measure is given by

$$d\mu(\zeta) = \frac{2J+1}{\pi} \frac{d\text{Re}\zeta d\text{Im}\zeta}{(1+|\zeta|^2)^2}, \quad (2.3)$$

The overlap of two coherent states is given by

$$\langle \zeta | \zeta' \rangle = \left[\frac{(1 + \zeta^* \zeta')^2}{(1 + |\zeta|^2)(1 + |\zeta'|^2)} \right]^J. \quad (2.4)$$

Time evolution of the spin system is described by the transition amplitude which joins two coherent states $|\zeta'\rangle$

and $|\zeta''\rangle$:

$$K(\zeta'', t'' | \zeta', t') = \langle \zeta'' | \exp[-i\hat{H}(t'' - t')/\hbar] | \zeta' \rangle, \quad (2.5)$$

where the Hamiltonian \hat{H} is given by a polynomial form of \hat{J}_\pm and \hat{J}_z which obeys a prescribed convention of ordering. Making use of the completeness relation (2.3), the propagator (2.5) is cast into the path integral form

$$K = \lim_{N \rightarrow \infty} \int \prod_{k=1}^{N-1} d\mu(\zeta_k) \exp[iS^{(N)}/\hbar], \quad (2.6a)$$

$(N\epsilon = t'' - t')$

$$S^{(N)} = \sum_{k=1}^N \epsilon \left(\frac{iJ\hbar}{1+|\zeta_k|^2} \left(\zeta_k^* \frac{\zeta_k - \zeta_{k-1}}{\epsilon} - \zeta_k \frac{\zeta_k^* - \zeta_{k-1}^*}{\epsilon} \right) - \langle \zeta_k | \hat{H} | \zeta_k \rangle \right), \quad (2.6b)$$

which is formally written as

$$K = \int \mathcal{D}\mu[\zeta(t)] \exp[iS/\hbar]. \quad (2.7)$$

The action functional $S[\zeta(t)]$ is given by

$$S[\zeta(t)] = \int_{t'}^{t''} \left[\frac{iJ\hbar}{1+|\zeta|^2} (\zeta^* \dot{\zeta} - \dot{\zeta}^* \zeta) - \mathcal{H} \right] dt, \quad (2.8)$$

with

$$\mathcal{H} = \langle \zeta | \hat{H} | \zeta \rangle.$$

Classical limit

In the limit of $\hbar \rightarrow 0$, the dominant contribution to the path integral (2.7) comes from the path which makes the action functional stationary (the stationary phase approximation). The dominant path obeys thus the variation principle $\delta S = 0$, which yields the equations of motion

$$\begin{aligned} \dot{\zeta} &= \frac{(1+|\zeta|^2)^2}{2iJ\hbar} \frac{\partial \mathcal{H}}{\partial \zeta^*}, \\ \dot{\zeta}^* &= -\frac{(1+|\zeta|^2)^2}{2iJ\hbar} \frac{\partial \mathcal{H}}{\partial \zeta}, \end{aligned} \quad (2.9)$$

whose solutions are subject to the end point conditions $\zeta(t') = \zeta'$ and $\zeta(t'') = \zeta''$. For later purposes it is convenient to rewrite (2.9) in terms of real variables X and Y ($\zeta = X + iY$):

$$\begin{aligned} \dot{X} &= \frac{(1+X^2+Y^2)^2}{4J\hbar} \frac{\partial \mathcal{H}}{\partial Y}, \\ \dot{Y} &= -\frac{(1+X^2+Y^2)^2}{4J\hbar} \frac{\partial \mathcal{H}}{\partial X}. \end{aligned} \quad (2.10)$$

The equations of motion (2.9) [or (2.10)] can be regarded as an extension of the usual canonical equations to those for the "curved phase space" (\simeq two dimensional sphere). In fact one can define the Poisson bracket (PB) as

$$\{A, B\} = \frac{(1+X^2+Y^2)^2}{4J\hbar} \left(\frac{\partial A}{\partial X} \frac{\partial B}{\partial Y} - \frac{\partial B}{\partial X} \frac{\partial A}{\partial Y} \right), \quad (2.11)$$

where the factor in front of the bracket on the rhs reflects the metric of the curved phase space. By means of PB (2.11), the

equations of motion are expressed as

$$\begin{aligned} \dot{X} &= \{X, \mathcal{H}\}, \\ \dot{Y} &= \{Y, \mathcal{H}\}. \end{aligned} \quad (2.12)$$

In the special case $A = X$ and $B = Y$, we have

$$\{X, Y\} = (1+X^2+Y^2)^2/4J\hbar, \quad (2.13)$$

which implies that X and Y are canonically conjugate to each other in the curved phase space.

3. SEMICLASSICAL APPROXIMATION

We investigate the semiclassical analysis starting with the propagator (2.7). [Hereafter we use the real form of (2.7) through the relation $\zeta = X + iY$.] In order to do this, one has to specify the classical path around which the semiclassical expansion is performed. As was suggested by Klauder,^{4,5(a)} the equations of motion (2.10) do not always have any solution satisfying the arbitrary boundary conditions $(X(t'), Y(t')) = (X', Y')$ and $(X(t''), Y(t'')) = (X'', Y'')$, that is, the number of boundary conditions is excessive compared with the number of the equations of motion. In the limit $\hbar \rightarrow 0$, however, the propagator (2.7) takes a value, $K^{\text{cl}} \sim \exp(iS^{\text{cl}}/\hbar)$, if the classical path starting with the initial point (X', Y') passes through the final point (X'', Y'') ; otherwise it vanishes. This is seen from the fact that the classical action S^{cl} is defined only for the path given above. In order to get the classical path, it is actually sufficient to specify boundary values of only one variable (X', X'') [or (Y', Y'')], since the boundary values of the other, (Y', Y'') [or (X', X'')], are automatically determined by the equations of motion.⁸ Thus we shall perform the semiclassical expansion about the so determined path.

A. Semiclassical propagator

Let $[X_{\text{cl}}(t), Y_{\text{cl}}(t)]$ be the classical paths satisfying the boundary conditions

$$\begin{aligned} X_{\text{cl}}(t') &= X', \\ X_{\text{cl}}(t'') &= X''. \end{aligned} \quad (3.1)$$

There are in general several paths which satisfy (3.1), and we assume, hereafter, that these paths are far enough apart from

one another to treat separately their contributions to the path integral (2.7). We introduce the path variation from the classical path:

$$\begin{aligned}\xi &= X - X_{cl}, \\ \eta &= Y - Y_{cl},\end{aligned}\quad (3.2)$$

which satisfies

$$\xi(t') = \xi(t'') = 0. \quad (3.3)$$

The action functional is then expanded up to second order

$$S = S_{cl} + S^{(2)}. \quad (3.4)$$

S_{cl} is the action for the classical path and the second variation of the action functional is calculated as

$$S^{(2)} = \int_{t'}^{t''} \frac{2J\hbar}{(1 + X_{cl}^2 + Y_{cl}^2)^2} \times [(\eta\dot{\xi} - \xi\dot{\eta}) - (A\xi^2 + 2B\xi\eta + C\eta^2)] dt. \quad (3.5)$$

The second term of (3.5) is called the "secondary Hamiltonian" and the coefficients are given by

$$\begin{aligned}A(t) &= \left[\frac{\partial}{\partial X} \left\{ \frac{(1 + X^2 + Y^2)^2}{4J\hbar} \left(\frac{\partial}{\partial X} \right) \right\} \right]_{cl}, \\ B(t) &= \frac{1}{2} \left[\frac{\partial}{\partial X} \left\{ \frac{(1 + X^2 + Y^2)^2}{4J\hbar} \left(\frac{\partial \mathcal{H}}{\partial Y} \right) \right\} \right. \\ &\quad \left. + \left[\frac{\partial}{\partial Y} \left\{ \frac{(1 + X^2 + Y^2)^2}{4J\hbar} \left(\frac{\partial \mathcal{H}}{\partial X} \right) \right\} \right]_{cl} \right], \quad (3.6) \\ C(t) &= \left[\frac{\partial}{\partial Y} \left\{ \frac{(1 + X^2 + Y^2)^2}{4J\hbar} \left(\frac{\partial \mathcal{H}}{\partial Y} \right) \right\} \right]_{cl},\end{aligned}$$

where the subscript "cl" denotes the value along the classical path. In the derivation of (3.6), we have used the equations of motion (2.10) together with integration by parts to eliminate the derivatives of the classical path \dot{X}_{cl} and \dot{Y}_{cl} .

The propagator (2.7) thus can be approximated as

$$K^{sc} = \tilde{K} \exp(iS_{cl}/\hbar). \quad (3.7)$$

\tilde{K} (which may be called the *reduced propagator*) is given by

$$\tilde{K} = \int \exp(i\tilde{S}^{(2)}/\hbar) \prod_{t' < t < t''} \left(\frac{d\xi(t)d\eta(t)}{(1 + X_{cl}^2 + Y_{cl}^2)^2} \frac{2J + 1}{\pi} \right), \quad (3.8)$$

where the path measure is obtained as a consequence of the stationary phase approximation, i.e., the weight factor in the original path differential in (2.7) is replaced by the value along the classical path.

In order to carry out the functional integral (3.8), we introduce the following transformation:

$$\begin{aligned}x &= \frac{2\sqrt{J\hbar}}{1 + X_{cl}^2 + Y_{cl}^2} \xi, \\ y &= \frac{2\sqrt{J\hbar}}{1 + X_{cl}^2 + Y_{cl}^2} \eta.\end{aligned}\quad (3.9)$$

Hence, using the functional Jacobian

$$\frac{\delta[\xi, \eta]}{\delta[x, y]} = \prod_{t' < t < t''} \frac{(1 + X_{cl}^2 + Y_{cl}^2)^2}{4J\hbar},$$

the reduced propagator is written as

$$\tilde{K} = \int \exp(i\tilde{S}^{(2)}/\hbar) \prod_{t' < t < t''} \left(\frac{2J + 1}{2J} \frac{dx(t)dy(t)}{2\pi\hbar} \right), \quad (3.10)$$

where

$$\tilde{S}^{(2)}[x, y] = \int_{t'}^{t''} [\frac{1}{2}(y\dot{x} - x\dot{y}) - \mathcal{H}^{(2)}(x, y)] dt,$$

and

$$\mathcal{H}^{(2)}(x, y) = \frac{1}{2}(Ax^2 + 2Bxy + Cy^2). \quad (3.11)$$

$\tilde{S}^{(2)}$ is written as a discretized form by using the boundary condition $x(t') = x(t'') = 0$,

$$\begin{aligned}\tilde{S}_N^{(2)} &= \sum_{k=1}^N \left(\frac{y_k + y_{k-1}}{2} \right) (x_k - x_{k-1}) \\ &\quad - \frac{1}{2} \sum_{k=1}^N \epsilon (A_k x_k^2 + 2B_k x_k y_k + C_k y_k^2).\end{aligned}\quad (3.12)$$

We can replace the mean value $(y_k + y_{k-1})/2$ by y_k ; then the first term of (3.12) turns out to be $\sum_{k=1}^N y_k (x_k - x_{k-1})$. Thus (3.10) becomes

$$\tilde{K} = \lim_{N \rightarrow \infty} \tilde{K}_N = \lim_{N \rightarrow \infty} \left(\frac{2J + 1}{2J} \right)^{N-1} \int \exp\left(\frac{\tilde{S}_N^{(2)}}{\hbar} \right) \prod_{k=1}^{N-1} \frac{dx_k dy_k}{(2\pi\hbar)}. \quad (3.13)$$

\tilde{K} is of a similar form to the conventional phase-space path integral, but the path measure is different from the usual sort, i.e., the integration over y_N is missing. This missing y_N can be recovered by taking an average of (3.13) over y_N :

$\tilde{K}_N \rightarrow V_0^{-1} \int dy_N \tilde{K}_N / (2\pi\hbar)$ with $V_0 = \int dy_N / (2\pi\hbar)$. The factors before the integral are considered as simple constants (though infinite) and do not play any physical roles; hence we can discard them by performing the normalization

$\tilde{K} / \lim_{N \rightarrow \infty} [(2J + 1)/2J]^{N-1} V_0^{-1} \rightarrow \tilde{K}$. Thus the reduced propagator is just transcribed into the one for the usual flat phase space path integral:

$$\tilde{K} = \int \exp(i\tilde{S}^{(2)}/\hbar) \mathcal{D}[x(t), y(t)], \quad (3.14)$$

with

$$\mathcal{D}[x(t), y(t)] = \lim_{N \rightarrow \infty} \prod_{k=1}^N dy_k \prod_{k=1}^{N-1} dx_k (2\pi\hbar)^{-N}.$$

B. Calculation of \tilde{K}

The evaluation of the Gaussian path integral in (3.14) can be most readily carried out by the path expansion method.⁷ In order to do this, we write $\tilde{S}^{(2)}$ in a bilinear form of x and y :

$$\tilde{S}^{(2)} = \frac{1}{2} \int_{t'}^{t''} \Phi \Lambda \Phi dt, \quad (3.15)$$

where Φ and Λ are defined by

$$\Phi = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \Lambda = \begin{pmatrix} -A & -B - d/dt \\ -B + d/dt & -C \end{pmatrix}. \quad (3.16)$$

The operator A is called the Dirac type operator in the literature.⁹ Let us consider the eigenvalue problem

$$A\Phi_k = \lambda_k \Phi_k, \quad x_k(t') = x_k(t'') = 0. \quad (3.17)$$

It is noted that the eigenvalues λ_k are real since A is Hermitian. The system of eigenfunctions $\{\Phi_k\}$ is orthonormalized:

$$\int_{t'}^{t''} \Phi_k(t) \Phi_l(t) dt = \int_{t'}^{t''} [x_k(t)x_l(t) + y_k(t)y_l(t)] dt = \delta_{kl}. \quad (3.18)$$

An arbitrary $\Phi(t)$ can be expanded as

$$\Phi(t) = \sum_k a_k \Phi_k(t), \quad (3.19)$$

with the coefficients a_k determined by

$$a_k = \int_{t'}^{t''} \Phi_k(t) \Phi(t) dt = \int_{t'}^{t''} [x_k(t)x(t) + y_k(t)y(t)] dt. \quad (3.20)$$

By means of (3.19) and (3.17) the path integral (3.14) is cast into

$$\tilde{K} = \int \dots \int \exp\left[\frac{i}{2\hbar} \sum_{k=1}^{\infty} \lambda_k a_k^2\right] \Delta \cdot \prod_{k=1}^{\infty} da_k, \quad (3.21)$$

which is readily calculated as

$$\tilde{K} = \Delta \cdot \prod_{k=1}^{\infty} [2\pi i \hbar / \lambda_k]^{1/2}. \quad (3.22)$$

The factor Δ in (3.21) is the functional determinant defined by

$$\mathcal{D}[x(t), y(t)] = \Delta \cdot \prod_{k=1}^{\infty} da_k. \quad (3.23)$$

We see that Δ is independent of the choice of the secondary Hamiltonian. Let (\tilde{x}, \tilde{y}) be another variation associated with an alternative secondary Hamiltonian; then (\tilde{x}, \tilde{y}) can be connected with (x, y) by an appropriate canonical transformation,¹⁰ which remains the measure $\mathcal{D}[x(t), y(t)]$ invariant.

Using the invariant property of Δ , \tilde{K} is rewritten through a reference propagator $\tilde{K}^{(\alpha)}$:

$$\tilde{K} = \left[\prod_{k=1}^{\infty} \lambda_k^{(\alpha)} / \prod_{k=1}^{\infty} \lambda_k \right]^{1/2} \cdot \tilde{K}^{(\alpha)}, \quad (3.24)$$

where $\lambda_k^{(\alpha)}$'s are eigenvalues for the operator $A^{(\alpha)}$ corresponding to a reference Hamiltonian. As the reference Hamiltonian, we can choose the simplest case $A = B = 0$, $C = \alpha$ with a real parameter $\alpha > 0$; then $\tilde{K}^{(\alpha)}$ reads as

$$\tilde{K}^{(\alpha)} = \int \exp\left\{ (i/2\hbar) \int_{t'}^{t''} [(y\dot{x} - x\dot{y}) - \alpha y^2] dt \right\} \times \mathcal{D}[x(t), y(t)], \quad (3.25)$$

which is just the propagator for a free particle which goes through $x(t') = x(t'') = 0$. The integral (3.25) is evaluated as

$$\tilde{K}^{(\alpha)} = [2\pi i \hbar \alpha (t'' - t')]^{-1/2}. \quad (3.26)$$

The eigenvalue equation for the reference Hamiltonian reads

$$-\frac{dy}{dt} = \lambda^{(\alpha)} x, \quad \frac{dx}{dt} = (\lambda^{(\alpha)} - \alpha) y, \quad (3.27)$$

with

$$x(t') = x(t'') = 0.$$

We immediately obtain the eigenvalues

$$\lambda_{n+1}^{(\alpha)} = \frac{1}{2} \left\{ \alpha \pm \left[\alpha^2 + [2n\pi / (t'' - t')]^2 \right]^{1/2} \right\} \quad (n = 0, 1, 2, \dots). \quad (3.28)$$

In particular we have two eigenvalues $\lambda_1 = 0$, α for $n = 0$; but $\lambda_1 = 0$ should be discounted, for the corresponding eigenfunction is trivial: $x(t) = y(t) \equiv 0$. Thus, isolating $\lambda_1 = \alpha$ in (3.24), we get

$$\tilde{K} = \left[\prod_{k=2}^{\infty} \lambda_k^{(\alpha)} / \prod_{k=1}^{\infty} \lambda_k \right]^{1/2} \cdot \alpha^{1/2} \cdot [2\pi i \hbar \alpha (t'' - t')]^{-1/2} = [2\pi i \hbar (t'' - t')]^{-1/2} \cdot \left[\prod_{k=2}^{\infty} \lambda_k^{(\alpha)} / \prod_{k=1}^{\infty} \lambda_k \right]^{1/2}. \quad (3.29)$$

Finally, in the limiting case of $\alpha \rightarrow 0$ (i.e., the identically vanishing secondary Hamiltonian), (3.29) approaches¹¹

$$\tilde{K} = [2\pi i \hbar (t'' - t')]^{-1/2} \cdot \left[\prod_{k=2}^{\infty} \lambda_k^{(0)} / \prod_{k=1}^{\infty} \lambda_k \right]^{1/2}, \quad (3.30)$$

where

$$\lambda_k^{(0)} = \pm (k-1)\pi / (t'' - t') \quad (k = 2, 3, 4, \dots).$$

1. Absolute value of \tilde{K}

In order to handle the expression (3.30), it is convenient to investigate its absolute value and phase separately. First, we study the absolute value. We assume (without mathematical rigor) that the ratio of the finite products in (3.30) converges uniformly. According to the theorem proved in Ref. 7, the absolute value of the infinite products in (3.30) is compactly written as

$$\left| \prod_{k=2}^{\infty} \lambda_k^{(0)} / \prod_{k=1}^{\infty} \lambda_k \right| = (t'' - t') / |\phi(t'')|, \quad (3.31)$$

so that

$$|\tilde{K}| = (2\pi \hbar)^{-1/2} \cdot |\phi(t'')|^{-1/2}. \quad (3.32)$$

$\phi(t)$ is the solution of the initial value problem

$$A\chi = 0, \quad \chi = \begin{pmatrix} \phi \\ \psi \end{pmatrix}, \quad \phi(t') = 0, \quad \psi(t') = 1. \quad (3.33)$$

The proof of (3.31) [or (3.32)] is based on some special techniques on the spectral theory of the Dirac type boundary problem. Alternatively we can verify this formula by directly evaluating the Gaussian path integral (3.14) with the aid of the conventional discretization procedure (see Appendix).

In order to construct the solution of (3.33), we perform a transformation

$$f = (1 + X_{cl}^2 + Y_{cl}^2) \phi / 2\sqrt{J\hbar}, \quad g = (1 + X_{cl}^2 + Y_{cl}^2) \psi / 2\sqrt{J\hbar}, \quad (3.34)$$

which is the same transformation as (3.9). Then, after simple calculations, we get

$$\begin{aligned}
\dot{f} &= \frac{\partial}{\partial X_{cl}} \left[\frac{(1+X^2+Y^2)^2}{4J\hbar} \frac{\partial \mathcal{H}}{\partial Y} \right]_{cl} f \\
&+ \frac{\partial}{\partial Y_{cl}} \left[\frac{(1+X^2+Y^2)^2}{4J\hbar} \frac{\partial \mathcal{H}}{\partial Y} \right]_{cl} g, \\
-\dot{g} &= \frac{\partial}{\partial X_{cl}} \left[\frac{(1+X^2+Y^2)^2}{4J\hbar} \frac{\partial \mathcal{H}}{\partial X} \right]_{cl} f \\
&+ \frac{\partial}{\partial Y_{cl}} \left[\frac{(1+X^2+Y^2)^2}{4J\hbar} \frac{\partial \mathcal{H}}{\partial X} \right]_{cl} g \quad (3.35)
\end{aligned}$$

with the initial conditions

$$f(t') = 0, \quad g(t') = (1+X'^2+Y'^2)/2\sqrt{J\hbar}. \quad (3.36)$$

Equations (3.35) may be regarded as the "Jacobi equations" for our curved phase space. Alternatively, these correspond to the variational equations in the stability theory of orbits (cf. Whittaker¹²). In fact, as is easily verified, Eqs. (3.35) can be derived by taking variations of the equations of motion (2.10) up to first order.

We can construct the solutions of Eqs (3.35) in terms of the classical orbits. We consider a family of classical paths parametrized by some real number α [i.e., $X_{cl}(t, \alpha)$, $Y_{cl}(t, \alpha)$], and the "response" associated with the variation with respect to α (i.e., $\partial X_{cl}/\partial \alpha$, $\partial Y_{cl}/\partial \alpha$) (we can assume that this family is differentiable with respect to α). Differentiating Eqs. (2.10) with respect to α , we get

$$\begin{aligned}
\frac{d}{dt} \left(\frac{\partial X_{cl}}{\partial \alpha} \right) &= \frac{\partial}{\partial X_{cl}} \left[\frac{(1+X^2+Y^2)^2}{4J\hbar} \frac{\partial \mathcal{H}}{\partial Y} \right]_{cl} \frac{\partial X_{cl}}{\partial \alpha} \\
&+ \frac{\partial}{\partial Y_{cl}} \left[\frac{(1+X^2+Y^2)^2}{4J\hbar} \frac{\partial \mathcal{H}}{\partial Y} \right]_{cl} \frac{\partial Y_{cl}}{\partial \alpha}, \\
-\frac{d}{dt} \left(\frac{\partial Y_{cl}}{\partial \alpha} \right) &= \frac{\partial}{\partial X_{cl}} \left[\frac{(1+X^2+Y^2)^2}{4J\hbar} \frac{\partial \mathcal{H}}{\partial X} \right]_{cl} \frac{\partial X_{cl}}{\partial \alpha} \\
&+ \frac{\partial}{\partial Y_{cl}} \left[\frac{(1+X^2+Y^2)^2}{4J\hbar} \frac{\partial \mathcal{H}}{\partial X} \right]_{cl} \frac{\partial Y_{cl}}{\partial \alpha}, \quad (3.37)
\end{aligned}$$

which means that the response functions form a set of solutions of (3.35). Especially we can choose the initial values $X_{cl}(t') = X'$, $Y_{cl}(t') = Y'$ as parameters α . Taking account of the initial conditions (3.36), we obtain

$$\begin{aligned}
f(t) &= \frac{1+X'^2+Y'^2}{2\sqrt{J\hbar}} \frac{\partial X_{cl}(t)}{\partial Y'}, \\
g(t) &= \frac{1+X'^2+Y'^2}{2\sqrt{J\hbar}} \frac{\partial Y_{cl}(t)}{\partial Y'}. \quad (3.38)
\end{aligned}$$

Making use of transformation (3.34), we get

$$\begin{aligned}
\phi(t) &= \frac{1+X'^2+Y'^2}{1+X_{cl}(t)^2+Y_{cl}(t)^2} \frac{\partial X_{cl}(t)}{\partial Y'}, \\
\psi(t) &= \frac{1+X'^2+Y'^2}{1+X_{cl}(t)^2} \frac{\partial Y_{cl}(t)}{\partial Y'}. \quad (3.39)
\end{aligned}$$

Thus, we finally arrived at the closed form

$$\begin{aligned}
|\tilde{K}| &= (2\pi\hbar)^{-1/2} \\
&\times \left| \frac{1+X'^2+Y'^2}{1+X'^2+Y'^2} \left(\frac{\partial X_{cl}(t)}{\partial Y'} \right)_{t=t'}^{-1} \right|^{1/2}. \quad (3.40)
\end{aligned}$$

The formula (3.40) indicates that the reduced propagator (and hence the semiclassical propagator (3.10)) is completely given in terms of the classical orbit in the curved phase space. It is of a similar form to the reduced propagator for the usual phase space,⁷ $|\tilde{K}| = (2\pi\hbar)^{-1/2} |\partial q''/\partial p'|^{-1/2}$. However, there are two differences between the cases: (i) the factor $(1+X'^2+Y'^2)/(1+X'^2+Y'^2)$, which reveals nothing but the curved nature of the phase space, appears in our case, and (ii) in the case of the usual phase space \tilde{K} is reduced to the well-known Van Vleck determinant through the relation $p' = -\partial S/\partial q'$, whereas it is not written in such a simple form in our case, by virtue of the fact that the pair (X, Y) cannot be regarded as a canonical pair of the usual sort.

Here we remark that the semiclassical propagator obtained by (3.40) is quite different from the propagator on the SU(2) group manifold,⁶ which is a natural consequence of the DeWitt expression for the propagator in the Riemannian manifold.¹³ The discrepancy lies in the point that in our case the manifold on which the propagator is defined is the coset space SU(2)/U(1) and the action functional [i.e., (2.8)] has a rather different form from the one used in the conventional Feynman path integral.

2. Phase of \tilde{K}

Now we discuss the phase of \tilde{K} . We first note that the Dirac type operator has the twofold eigenvalues $\lambda_{k, \pm}$ for $k = 2, 3, \dots$, except one single eigenvalue which we denote as λ_1 . With this in mind, we separate the ratio of infinite products of eigenvalues into two groups labeled by $i = \pm$:

$$\frac{1}{\lambda_1} \prod_{i=\pm} \prod_{k=2}^{\infty} \frac{\lambda_{k,i}^{(0)}}{\lambda_{k,i}}, \quad (3.41)$$

where $\lambda_{k, \pm}^{(0)} = \pm(k-1)\pi/(t''-t')$ for $k = 2, 3, \dots$, and the eigenvalue λ_1 is isolated which tends to $\lambda_1^{(0)} = 0$ in the limit of the vanishing secondary Hamiltonian. We observe that for a sufficiently large k the sign of each ratio $\lambda_{k,i}^{(0)}/\lambda_{k,i}$ is positive for both groups; therefore the total sign is determined by the ratios for only small k 's. In fact we can see that for a small k the ratio for each group may exhibit an opposite sign to that for the other group, and we denote the number of such k 's as ν . Then the total sign becomes $(-)^{\nu}$. The phase of the reduced propagator \tilde{K} is given by $\exp(-i\nu\pi/2)$.

The number ν corresponds to the "Morse index" of the classical path, which is precisely defined in the variational analysis in the large.¹⁴ In the variational analysis, it is known that the Morse index is closely related to the notion of focal points. In our case, the focal point is defined as the point at which the solution $\phi(t)$ of Eqs. (3.35) vanishes along the classical path and the Morse index ν is given by the number of such focal points.

C. Simple model

In order to illustrate the results obtained in the preceding subsections, we consider the motion of the spin system described by the simple model Hamiltonian

$$\hat{H} = -\omega \hat{j}_z, \quad (3.42)$$

(e.g., a single spin in a homogeneous magnetic field). Making

use of the formula in Ref. 5(b), the classical Hamiltonian is given by

$$\mathcal{H} = \omega \hbar J \frac{1 - X^2 - Y^2}{1 + X^2 + Y^2}. \quad (3.43)$$

The equations of motion are immediately obtained as

$$\dot{X}(t) = -\omega \cdot Y(t), \quad \dot{Y}(t) = \omega \cdot X(t), \quad (3.44)$$

solutions of which are subject to the end-point conditions $X(t') = X'$ and $X(t'') = X''$. By solving (3.44), we get the classical orbit

$$\begin{aligned} X_{cl}(t) &= X' \cdot \cos\omega(t - t') - Y' \cdot \sin\omega(t - t'), \\ Y_{cl}(t) &= X' \cdot \sin\omega(t - t') + Y' \cdot \cos\omega(t - t'), \end{aligned} \quad (3.45)$$

where $Y' \equiv Y_{cl}(t')$, which depends on X' and X''

The eigenvalue problem (3.17) now reads

$$\begin{aligned} \frac{dx_k}{dt} + \omega y_k &= \lambda_k y_k, \\ -\frac{dy_k}{dt} + \omega x_k &= \lambda_k x_k, \end{aligned} \quad (3.46)$$

$$x_k(t') = x_k(t'') = 0,$$

from which the eigenvalues are readily obtained as

$$\lambda_k = \omega + n_k \pi / (t'' - t'), \quad (3.47)$$

where $n_k = \pm 1, \pm 2, \dots$, for $k > 1$ and $n_1 = 0$. On the other hand, the eigenvalues $\lambda_k^{(0)}$ turn out to be

$$\lambda_k^{(0)} = n_k \pi / (t'' - t'). \quad (3.48)$$

Thus the reduced propagator is given by

$$\begin{aligned} \tilde{K} &= [2\pi i \hbar \omega (t'' - t')]^{-1/2} \\ &\times \left\{ \prod_{n=1}^{\infty} \left[1 - \left(\frac{\omega(t'' - t')}{n\pi} \right)^2 \right] \right\}^{-1/2}, \end{aligned}$$

which becomes, with the aid of the Euler formula $(\sin x)/x = \prod_{n=1}^{\infty} [1 - (x/n\pi)^2]$,

$$\tilde{K} = [2\pi i \hbar \omega (t'' - t')]^{-1/2}. \quad (3.49)$$

The result (3.49) can be readily deduced from the initial value problem (3.33). In fact Eqs. (3.33) are written as

$$\begin{aligned} \dot{\phi}(t) &= -\omega \psi(t), \quad \dot{\psi}(t) = \omega \phi(t), \\ \phi(t') &= 0, \quad \psi(t') = 1, \end{aligned} \quad (3.50)$$

whose solutions are

$$\begin{aligned} \phi(t) &= \frac{\partial X_{cl}(t)}{\partial Y'} = -\sin\omega(t - t'), \\ \psi(t) &= \frac{\partial Y_{cl}(t)}{\partial Y'} = \cos\omega(t - t'). \end{aligned} \quad (3.51)$$

By substituting $|\phi(t'')|$ into the formula (3.32), we arrive at the expression (3.49). The index ν is given by

$$\nu = [\omega(t'' - t')/\pi]_G, \quad (3.52)$$

where $[]_G$ is the Gauss symbol.

4. SEMICLASSICAL QUANTIZATION CONDITION

In this section, as an application of the semiclassical propagator obtained in the previous section, we shall examine the semiclassical quantization condition of the spin sys-

tem, which leads to bound state spectra.

In the case of stationary bound states, the orbit appearing in the semiclassical propagator becomes a closed curve and lies on the constant energy surface $\mathcal{H}(X, Y) = E$. The classical action is now given by

$$\begin{aligned} S_{cl} &= 2J\hbar \int_0^{t'} \frac{Y\dot{X} - X\dot{Y}}{1 + X^2 + Y^2} dt - E \cdot t \\ &= 2J\hbar \int_{(X', Y')}^{(X'', Y'')} \frac{YdX - XdY}{1 + X^2 + Y^2} - E \cdot t. \end{aligned} \quad (4.1)$$

where we put $t'' = t$ and $t' = 0$ for the sake of simplicity, and (X, Y) is used to mean the classical path. The last integral in (4.1) means the line integral along the classical trajectory passing through the end points (X', Y') and (X'', Y'') . The semiclassical propagator thus reads

$$K(X'', Y'', t | X', Y', 0) = \tilde{K} \cdot e^{-iEt/\hbar}, \quad (4.2)$$

where \tilde{K} does not explicitly depend on t . The propagator (4.2) can be regarded as a *wavefunction* of the final point (X'', Y'', t) with the initial point $(X', Y', 0)$ being fixed, or vice versa; hence Eq. (4.2) corresponds to an ordinary wavefunction representing a stationary state of energy E . In the following we consider the case that the initial point (X', Y') is fixed.

Following Keller's idea,¹⁵ we derive the semiclassical quantization condition. The essential point of Keller's procedure is that the semiclassical wavefunction

$$\begin{aligned} \psi^{sc}(q, t) &= A(q, t) \exp[iS_{cl}(q, t)/\hbar] \\ &\equiv \exp\{i[S_{cl} + (\hbar/i)\log A]/\hbar\} \end{aligned} \quad (4.3)$$

should be "single-valued" with respect to the argument q . The single-valuedness makes a restriction on the change of the phase $\Delta S_{cl} + (\hbar/i)\Delta \log A$ when the system cycles along closed orbits, and this leads to a generalization of the Bohr-Sommerfeld quantization condition. We take over this idea into the present problem.

Let us write the semiclassical propagator as

$$K(X'', Y'', t | X', Y', 0) = \exp\{i[S_{cl} + (\hbar/i)\log \tilde{K}]/\hbar\}, \quad (4.4)$$

and evaluate the change of the phase, i.e., ΔS_{cl} and $(\hbar/i)\Delta \log \tilde{K}$, when the system goes around a closed loop starting from the point (X'', Y'') . First, from (4.1), ΔS_{cl} is given by the line integral

$$\Delta S_{cl} = 2J\hbar \int_C \frac{YdX - XdY}{1 + X^2 + Y^2}, \quad (4.5)$$

or using Stokes' theorem

$$\Delta S_{cl} = 4J\hbar \oint_S \frac{dX \wedge dY}{(1 + X^2 + Y^2)^2}, \quad (4.5')$$

where the integral is taken over the area encircled by the classical orbit. Next, the change $\Delta \log \tilde{K}$ is deduced from the number or singularities of \tilde{K} along the classical orbit [see also Keller¹⁵; it should be noted that the reduced propagator \tilde{K} plays the role of the amplitude A in (4.3)]. Recalling that the trajectory goes through a focal point, \tilde{K} changes $\sqrt{-1}$ and $\Delta \log \tilde{K}$ becomes $i\pi/2$. If the index is ν , the total change of $\Delta \log \tilde{K}$ yields $i\nu\pi/2$. Thus, by the single-valuedness condition $\exp[i(\Delta S_{cl} + (\hbar/i)\Delta \log \tilde{K})/\hbar] = 1$, we get

$$4J \int_S \frac{dX \wedge dY}{(1+X^2+Y^2)^2} = (2n + \frac{\nu}{2}) \cdot \pi, \quad (4.6)$$

where $n = 0, 1, 2, \dots$. With the aid of the stereographic projection

$$\begin{aligned} X &= \cot(\vartheta/2) \cdot \cos\varphi, \\ Y &= \cot(\vartheta/2) \cdot \sin\varphi, \end{aligned} \quad (4.7)$$

(4.6) reads

$$J \int_S \sin \vartheta d\vartheta d\varphi = (2n + \nu/2)\pi. \quad (4.6')$$

Equation (4.6) is just regarded as a quantization condition à la Bohr–Sommerfeld in the curved phase space S^2 . However, in contrast to the usual (flat) phase space quantization condition, the integer n does not take arbitrary positive values because the integral (4.6) is just proportional to the area on a unit sphere encircled by the closed orbit and is bounded by 4π .

Finally we examine the formula (4.6) for the simplest Hamiltonian $\bar{H} = -\hbar\omega J_z$, which was taken up in the previous section. The classical orbit (3.45) is written in angle variables as

$$\vartheta = \vartheta_0 = \text{const}, \quad \varphi = \omega t + \varphi_0, \quad (4.8)$$

which just describes a circle on a unit sphere. The energy is given by

$$E = -\omega \hbar J_z = -\omega \hbar J \cos\vartheta_0. \quad (4.9)$$

The index ν is given by 2 for the solution (4.8), since the singularity of the semiclassical propagator, as is seen from Eq. (3.49), appears twice per period, i.e., $\sin \omega t$ vanishes at two points $t = \pi/\omega, 2\pi/\omega$. Thus Eq. (4.6') yields

$$2\pi J (1 - \cos\vartheta_0) = (2n + 1)\pi. \quad (4.10)$$

Using the relation $J_z = J \cos\vartheta_0$, we get

$$J_z = J - (n + \frac{1}{2}).$$

From (4.11), J_z takes an integer or half-integer value between $-(J - 1/2)$ and $J - 1/2$. Hence we obtain an energy spectrum

$$E_m = -\omega \hbar m \quad [m = -(J - \frac{1}{2}), -(J - \frac{3}{2}), \dots, J - \frac{1}{2}].$$

This spectrum clearly differs from the exact result of the quantum theory, i.e., the magnitude of the spin is reduced by $\hbar/2$. This discrepancy originates from the index $\nu = 2$. It suggests that in the semiclassical approximation one makes a replacement $J \rightarrow J + 1/2$. This is correct in the large J limit, where the semiclassical picture becomes accurate.

5. CONCLUDING REMARKS

In this paper we have investigated a semiclassical analysis of the spin system and obtained a closed form of the semiclassical propagator. The formula obtained here is useful for the approximate calculation of the energy spectra of an asymmetric top and the many-body systems which can be described by the quasispin formulation, etc.¹⁶

The essential point of our treatment is that the semiclassical propagator in the curved phase space (\simeq two-dimensional sphere), which appears at first sight rather complicated, can be handled on the same footing as that in the usual

flat phase space. In this way, the present method would provide us a promising tool for the semiclassical analysis of the path integral in a curved phase space with more complicated geometrical structure. As an immediate extension, we can consider the spin systems with many degrees of freedom, e.g., a spin chain.

The other problem is the extension to the path integral in the coherent states for the unitary group of higher dimension which has been proposed elsewhere¹⁷ in connection with the quantization of nuclear collective motions. The present method would provide us with a useful basis for this subject. As for a path integral on the group manifold $SU(n)$, Dowker found a compact formula for the propagator by applying the DeWitt formula for the Riemannian space as the case of $SU(2)$ propagator,¹⁸ whereas the path integral form in Ref. 17 is given by the functional integral on the homogeneous space $U(m+n)/U(m) \times U(n)$; hence the semiclassical propagator may result in a quite different form from the one for the group manifold $SU(n)$.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Tōru Suzuki for several discussions which are helpful for this work. Their thanks are also due to the members of the Nuclear Theory Group of Kyoto University. One of the authors (Y.M.) acknowledges the Japan Society for the Promotion of Science for financial support.

APPENDIX

In this appendix we evaluate the reduced propagator (3.14) by adopting the discretization procedure for functional Gaussian integral.¹⁹ The N th approximation for \tilde{K} reads

$$\begin{aligned} \tilde{K}_N &= (2\pi\hbar)^{-N} \int \prod_{k=1}^{N-1} dx_k \prod_{k=1}^N dy_k \\ &\times \exp \left[(i/\hbar) \sum_{k=1}^N \{ y_k (x_k - x_{k-1}) \right. \\ &\quad \left. - \frac{1}{2} \epsilon (A_k x_k^2 + 2B_k x_k y_k + C_k y_k^2) \} \right], \end{aligned} \quad (A1)$$

with $\epsilon \equiv (t'' - t')/N$. By performing the integration over y variables, we get

$$\begin{aligned} \tilde{K}_N &= (2\pi\hbar)^{-N} (-2\pi i \hbar)^{N/2} \prod_{k=1}^N (C_k \epsilon)^{-1/2} \\ &\int \prod_{k=1}^{N-1} dx_k \exp \left[\frac{i}{\hbar} L_N^{(2)} \right]. \end{aligned} \quad (A2)$$

$L_N^{(2)}$ is the N th approximation of the Lagrangian

$$L^{(2)} = \int_{t'}^{t''} \left[\frac{1}{2C} \dot{x}^2 - \frac{B}{C} x \dot{x} + \frac{1}{2} \left(\frac{B^2}{C} - A \right) x^2 \right] dt. \quad (A3)$$

By using the integration by part and noting the boundary $x_0 = x_N = 0$, the N th approximation of $L^{(2)}$ becomes

$$L_N^{(2)} = \sum_{k=1}^N \left\{ \frac{(x_k - x_{k-1})^2}{2C_k \epsilon} - \frac{1}{2} \left[A - \frac{B^2}{C} - \left(\frac{B}{C} \right)' \right]_k x_k^2 \cdot \epsilon \right\}, \quad (A4)$$

where $()'$ denotes the differentiation with respect to t and $()_k$ the value at the time $t_k = t' + k\epsilon$. Noting that the (A4) is

written as the quadratic form, and, using the Gaussian integral formula, we get for \tilde{K}_N

$$\tilde{K}_N = (2\pi i \hbar)^{-1/2} \left\{ (C_1 \epsilon) \cdot \det \left[\begin{pmatrix} C_N \epsilon & & 0 \\ & \ddots & \\ 0 & & C_2 \epsilon \end{pmatrix} \hat{a} \right] \right\}^{-1/2}, \quad (\text{A5})$$

where \hat{a} is the "Jacobi matrix" of order $N-1$:

$$\hat{a} = \begin{pmatrix} a_{N-1} & -\frac{1}{C_{N-1}\epsilon} & & 0 \\ -\frac{1}{C_{N-1}\epsilon} & a_{N-2} & \ddots & \\ & \ddots & \ddots & -\frac{1}{C_2\epsilon} \\ 0 & & -\frac{1}{C_2\epsilon} & a_1 \end{pmatrix}, \quad (\text{A6})$$

$$a_{k-1} = \frac{1}{C_{k-1}\epsilon} + \frac{1}{C_k\epsilon} + \left(A - \frac{B^2}{C} - \left(\frac{B}{C} \right)' \right)_{k-1} \cdot \epsilon.$$

Following the procedure by Gel'fand and Yaglom,¹⁹ we evaluate the limit $\tilde{K} = \lim \tilde{K}_N$. Let us introduce

$$D_k = C_1 \epsilon \cdot \det \left[\begin{pmatrix} C_k \epsilon & & 0 \\ & \ddots & \\ 0 & & C_2 \epsilon \end{pmatrix} \hat{a}_k \right] \quad (k = 2, \dots, N), \quad (\text{A7})$$

where \hat{a}_k is the matrix obtained by the replacement $N \rightarrow k$. The matrix in (A7) is also the Jacobi type matrix (but not symmetric). Making use of the recurrence relation for the determinant of the Jacobi type matrix, we obtain the difference equation

$$D_{k+2} = \left[1 + \frac{C_{k+2}}{C_{k+1}} + C_{k+2} \epsilon^2 \left(A - \frac{B^2}{C} - \left(\frac{B}{C} \right)' \right)_{k+1} \right] \cdot D_{k+1} - \frac{C_{k+2}}{C_{k+1}} D_k. \quad (\text{A8})$$

Putting $D_k \equiv D(k\epsilon)$, (A8) is rewritten as

$$\begin{aligned} & \frac{D((k+2)\epsilon) - 2D((k+1)\epsilon) + D(k\epsilon)}{\epsilon^2} \\ &= \left(\frac{C_{k+2}}{C_{k+1}} - 1 \right) \frac{D((k+1)\epsilon) - D(k\epsilon)}{\epsilon^2} \\ &+ C_{k+2} \left(A - \frac{B^2}{C} - \left(\frac{B}{C} \right)' \right)_{k+1} D((k+1)\epsilon). \end{aligned} \quad (\text{A9})$$

In the limit $N \rightarrow \infty$ ($\epsilon \rightarrow 0$), this reduces to the differential equation

$$\frac{d^2 D}{d\tau^2} - \frac{1}{C} \frac{dC}{d\tau} \frac{dD}{d\tau} - C \left(A - \frac{B^2}{C} - \frac{d}{d\tau} \left(\frac{B}{C} \right) \right) \cdot D = 0. \quad (\text{A10})$$

The initial conditions to determine the solution of (A10) are given by

$$D_2 = C_1 \epsilon \left[1 + \frac{C_2}{C_1} - C_2 \left(A - \frac{B^2}{C} - \left(\frac{B}{C} \right)' \right)_1 \cdot \epsilon^2 \right]$$

and

$$\begin{aligned} \frac{D_3 - D_2}{\epsilon} &= C_1 \left[1 + \frac{C_3}{C_2} - C_3 \left(A - \frac{B^2}{C} - \left(\frac{B}{C} \right)' \right)_2 \cdot \epsilon^2 \right] \\ &\times \left[1 + \frac{C_2}{C_1} - C_2 \left(A - \frac{B^2}{C} - \left(\frac{B}{C} \right)' \right)_1 \cdot \epsilon^2 \right] - \frac{C_3}{C_2} \\ &- \left[1 + \frac{C_2}{C_1} - C_2 \left(A - \frac{B^2}{C} - \left(\frac{B}{C} \right)' \right)_1 \cdot \epsilon^2 \right]. \end{aligned}$$

Noting $C_2/C_1, C_3/C_2 \rightarrow 1$ for $\epsilon \rightarrow 0$, these reduce to

$$D(t') = 0 \quad \text{and} \quad \frac{dD}{d\tau} \Big|_{\tau=t'} = C(t'). \quad (\text{A11})$$

Thus we get for the reduced propagator \tilde{K}

$$\tilde{K} = (2\pi i \hbar)^{-1/2} [D(t'')]^{-1/2}, \quad (\text{A12})$$

where $D(t'')$ is the value at $t = t''$ of the solution of (A10) with the initial conditions (A11). We can show that $D(\tau)$ becomes the solution $x(\tau)$ of the initial value problem:

$$\begin{aligned} \frac{dx}{d\tau} &= Bx + Cy, \\ \frac{dy}{d\tau} &= -Ax - By, \end{aligned} \quad (\text{A13})$$

with $x(t') = 0, y(t') = 1$, which are just Eqs. (3.33). This is verified by observing that the elimination of y from (A13) yields (A10).

¹M. Gutzwiller, *J. Math. Phys.* **8**, 1979 (1967); **10**, 1004 (1969); **11**, 1791 (1970); **12**, 343 (1971).

²R. F. Dashen, B. Hasslacher, and A. Neveu, *Phys. Rev. D* **10**, 4114, 4130, 4138, (1974); **D 11**, 3424 (1975); **D 12**, 2443 (1975).

³J. R. Klauder, *Ann. Phys. (N.Y.)*, **11**, 123 (1960). See also S. S. Schweber, *J. Math. Phys.* **3**, 831 (1962).

⁴J. R. Klauder, in *Path Integrals*, Proceedings of the NATO Advanced Summer Institute, edited by G. J. Papadopoulos and J. T. Devreese (Plenum, New York, 1978), p. 5.

⁵(a) J. R. Klauder, *Phys. Rev. D* **19**, 2349 (1979); (b) H. Kuratsuji and T. Suzuki, *J. Math. Phys.* **21**, 472 (1980).

⁶L. Schulman, *Phys. Rev.* **176**, 1538 (1969).

⁷S. Levit and U. Smilansky, *Ann. Phys. (N.Y.)* **108**, 165 (1977).

⁸In this respect, Klauder considered more general cases [Refs. 4, 5(a)].

Allowing all the boundary values to be complex and extending the classical path space to the complex region, he obtained a solution which connects arbitrary initial and final points. We note that our case may be considered as a special case of his solutions, i.e., the boundary values (X', Y') and (X'', Y'') are all real.

⁹B. M. Levitan and I. S. Sargsian, *Introduction to the Spectral Theory* (Moscow, 1970).

¹⁰Such a canonical transformation can be constructed by using the solutions of the equation $\Lambda_X = 0$ [cf. Eq. (3.33)] with suitable initial conditions.

¹¹This expression is just the same as one obtained in Ref. 7, in which the explicit form of Δ is calculated with the aid of rather detour technique, but in our case one does not need an explicit form of Δ .

¹²E. T. Whittaker, *A Treatise on the Analytical Dynamics* (Cambridge U.P., London, 1937), Chap. XV.

¹³B. S. DeWitt, *Rev. Mod. Phys.* **29**, 377 (1957).

¹⁴J. W. Milnor, *Morse Theory* (Princeton U.P., Princeton, N.J., 1962); M. Morse, *Variational Analysis* (Wiley, New York, 1973).

¹⁵J. B. Keller, *Ann. Phys. (N.Y.)* **4**, 180 (1958).

¹⁶H. Kleinert, *Phys. Lett. B* **69**, 9 (1977); he proposed an alternative method of the functional approach to a model of the many-fermion system.

¹⁷H. Kuratsuji and T. Suzuki, *Phys. Lett. B* **92**, 19 (1980).

¹⁸J. S. Dowling, *J. Phys. A* **3**, 451 (1970); *Ann. Phys. (N.Y.)* **62**, 361 (1971).

¹⁹I. M. Gel'fand and A. M. Yaglom, *J. Math. Phys.* **1**, 48 (1960).

Semiclassical approximations at positive temperatures in stochastic physics ^{a)}

Steven M. Moore

Departamento de Física, Universidad de los Andes, Apartado Aéreo 4976, Bogotá, Colombia

(Received 2 April 1980; accepted for publication 28 August 1980)

Semiclassical approximations are developed for stochastic mechanics and stochastic field theory at positive temperatures. The tunneling phenomena of Euclidean quantum mechanics is seen to have a statistical interpretation. A semiclassical algorithm for calculating the generating functional of the moments of the positive-temperature process is developed. Positive-temperature fluctuations around a scalar soliton and in the pure SU (2) Yang-Mills theory are also briefly considered.

PACS numbers: 03.65.Sq, 03.70. + k, 05.30. — d, 11.10.Np

I. INTRODUCTION

In previous papers¹⁻⁵ Nelson's stochastic model of quantum mechanics⁶ has been extended to a more complete description of particle motion and fields. Although mathematical methods used in stochastic descriptions of microscopic phenomena still have a long way to go before they can compete with the highly successful mathematical algorithms of quantum mechanics, the program of developing them is interesting if only for the basic differences in interpretation that are obtained.

In a recent paper³ path-integral methods were developed for treating problems in stochastic mechanics (SM) and stochastic field theory (SFT) at both zero and positive temperatures. Many of the formulas of Euclidean quantum mechanics (EQM) were given a real-time interpretation in the spirit of Guerra and Ruggiero.⁷ However, at the level of interpretation it was pointed out that SM and SFT are to be distinguished from EQM and, in particular, stochastic fields are essentially different than Euclidean fields.³ Other interpretational problems, such as that of the $\beta\hbar$ -periodicity in time for equilibrium states (KMS condition), have also been treated.⁵

A natural approximation to use for path integrals is the semiclassical one, i.e. the expansion around the path (or paths) that make stationary the action functional. Indeed, this technique has been mentioned by Yasue for the zero-temperature case in SM⁸ and SFT.⁹ In particular, the path-integral formulas of his second paper⁹ can be made rigorous using the nonstandard analysis approach of the present author,^{3,4} an approach also initiated by Yasue.¹⁰ Yasue has also linked these semiclassical methods with stochastic control theory.¹¹ Another essentially different approach has recently advanced by Jona-Lasinio,¹² based on the semiclassical approximation in stochastic differential equations developed by Venttsel' and Freidlin.¹³

In this paper the objective is to study semiclassical methods in stochastic physics in some detail using the path-integral methods developed in the previous paper.³ These

have been treated extensively in EQM at zero temperatures (see the excellent review by Coleman¹⁴ for basic concepts and references), but the positive-temperature case has not been so treated. An analysis has been given for some specific examples at positive temperatures,¹⁵⁻¹⁷ but there still remains much to be done. Stochastic physics, being similar to EQM, has some things in common with the latter, but it also brings some surprises. Moreover, stochastic physics, with its direct probability interpretation, may be at least a definite aid to EQM and possibly a useful alternative to it.

Since the treatment of the zero-temperature semiclassical approximation can be considered as the limit of the positive-temperature one, attention is given here to positive-temperature phenomena. In the next section particle mechanics is considered, since this forms the basis for the treatment of field theoretical problems.^{1-3,5} The tunneling phenomena of EQM is seen to have a statistical interpretation. A semiclassical algorithm for calculating the generating functional of the moments of the positive-temperature process is developed. In the last section positive-temperature fluctuations around a scalar soliton and in the pure SU (2) Yang-Mills theory are also briefly considered.

II. PARTICLE MECHANICS

The basic postulates of SM were given in previous papers.^{1-3,5} They generalize Nelson's formulation,⁶ both in form and in their interpretation. One assumes that the physical system is described by a stochastic process in \mathbb{R}^n satisfying the stochastic differential equation

$$dq(t) = b(q(t))dt + dw(t). \quad (1)$$

Here dw is the differential of a Wiener-like process corresponding to Wiener measure over periodic paths; i.e., it is Gaussian with the first moments given by

$$\langle dw(t) \rangle = 0 \quad (2)$$

$$\langle dw_i(t) dw_j(t') \rangle = \frac{\delta_{ij}}{\beta m} \sum_{n=-\infty}^{\infty} e^{i w_n(t-t')} dt dt', \quad (3)$$

where

$$w_n = 2\pi n / \beta \hbar. \quad (4)$$

(The zero temperature moment corresponding to (3) is sim-

^{a)}Research financed in part by Colciencias.

ply the limit when $\beta \rightarrow \infty$ and is the one used by Nelson.⁶⁾

Some interpretational problems associated with SM were mentioned in previous papers.^{1,3} They have been treated in a more recent paper.⁵ It is probably best to consider SM as a zero charge Markov limit of classical stochastic electrodynamics,¹⁸ although this concept is not well defined since there is more than one way to take the Markov limit.¹⁹ These interesting questions aside, one can summarize the results of the treatment of interpretational problems as follows:⁵ (1) q should be considered as a fluctuation around a classical equilibrium point (this holds for classical stochastic electrodynamics as well); (2) The $\beta\hbar$ -periodicity (KMS condition) of q can be interpreted as an approximation to be used at low temperatures. To see this, one only needs to consider the spectrum of the positive-temperature process in classical stochastic electrodynamics. It is a Fourier integral and can therefore be approximated by a Fourier series using exponentials of period $\beta\hbar$. Thus the stochastic process can be approximated by one of period $\beta\hbar$, and, in the Markov limit, one obtains the $\beta\hbar$ -periodicity used previously.^{2,3} Of course, this would not make sense without (1); (3) The drift velocity b has the same form at all temperatures where the approximation is valid and is determined by the zero-point probability density:

$$b = (\hbar/2m)\nabla \ln \rho. \quad (5)$$

The last observation means that the temperature dependence of b in Eq. (1) comes from q itself through w . Also one must be careful to calculate averages since an average calculated with ρ only gives the zero-temperature average.

These observations about the interpretation of SM do not change the main results of the previous paper.³ For example, the passage from the Fokker-Planck equation for the transition probability,

$$\frac{\partial p}{\partial t} = -\nabla \cdot (bp) + (\hbar/2m)\Delta p \quad (6)$$

to the path integral formula,

$$p(x,t|x_0,t_0) = \left(\frac{\rho(x)}{\rho(x_0)}\right)^{1/2} \int d\mu_w\{q\} \times \exp\left[-\hbar^{-1} \int_{t_0}^t (U(q) - E_0) dt'\right] \quad (7)$$

is still correct, although E_0 is the zero-point average energy. U is now written instead of V , where $U(x) = V(x + \bar{x})$ and \bar{x} is the classical equilibrium point around which q fluctuates.⁵ b appears in (7) through ρ , but neither b nor E_0 appear in the generating functional for the moments,

$$G\{J\} = \int d\mu_w\{q\} \exp\left[-\hbar^{-1} \int_{-\beta\hbar/2}^{\beta\hbar/2} (U(q) + J \cdot q) dt'\right], \quad (8)$$

which was the basic tool used before.³ For the rest of this paper (7) will be written as

$$p(x,t|x_0,t_0) = \left(\frac{\rho(x)}{\rho(x_0)}\right)^{1/2} \int D\{q\}$$

$$\times \exp\left\{-\hbar^{-1} \int_{t_0}^t (\frac{1}{2}m\dot{q}^2 + U(q) - E_0) dt'\right\}, \quad (9)$$

with similar changes in (8). The use of a path integral measure $\mathcal{D}\{q\}$ instead of the Wiener measure $d\mu_w\{q\}$ is justified only by the ease of comparison with EQM.

The most probable fluctuation path q_0 will be the one that makes the action functional

$$S\{q\} \equiv \int_{t_0}^t (\frac{1}{2}m\dot{q}^2 + U(q)) dt' \quad (10)$$

stationary. Expanding $S\{q\}$ in a functional Taylor series around q_0 , one has that

$$S\{q\} = S\{q_0\} + \delta_{q_0} S\{Q\} + \delta_{q_0}^2 S\{Q,Q\} + \dots \quad (11)$$

The first variation of $S\{q\}$, namely $\delta_{q_0} S\{Q\}$, is zero by definition of q_0 . This is²⁰

$$\delta_{q_0} S\{Q\} = m\dot{q}_{0i} Q_i |'_{t_0} + \int_{t_0}^t (-m\ddot{q}_{0i} + U_i(q_0)) Q_i dt'. \quad (12)$$

As long as $t - t_0 < \beta\hbar$, one can guarantee that there exists a periodic path q_0 such that $q_0(t) = x$, $q_0(t_0) = x_0$. Hence the surface term in (12) is zero since $Q(t) = Q(t_0) = 0$. Thus, to have a stationary point, q_0 must satisfy

$$-m\ddot{q}_{0i} + U_i(q_0) = 0. \quad (13)$$

The fact that it has been assumed that $t - t_0 < \beta\hbar$ should not be passed over lightly. Indeed, the first surprise of SM (in comparison with EQM) is that for $t - t_0 < \beta\hbar$ there is no difference between the zero-temperature case and the positive-temperature one. Any path on $[t_0, t]$ can be made periodic with period $\beta\hbar$ if $t - t_0 < \beta\hbar$.²¹

It is amusing to consider a simple example of this equivalence. The example is any double well potential with equilibrium points a and b , $V(a) > V(b)$. Take q to be the fluctuation around the "false vacuum" $\bar{x} = a$. Let x_1 and x_2 be such that $a < x_1 < x_2 < b$ and $V(x_1) = V(x_2) = E_0$. (For the moment, it is assumed that such x_1 and x_2 exist.) Since the full path is $q + \bar{x}$, $q(t_0) = x_1 - a$, $q(t) = x_2 - a$. As a gross approximation, one has

$$p(x_2 - a, t|x_1 - a, t_0) \simeq \text{Const} \exp[-\hbar^{-1} S\{q_0\} + \hbar^{-1} E_0(t - t_0)]. \quad (14)$$

Equation (13) shows that

$$\frac{1}{2}m\dot{q}_0^2 - V(q_0 + a) = K \quad (15)$$

is a constant. This can be used to change variables in $S\{q_0\}$:

$$S\{q_0\} = -K(t - t_0) + \int_{x_1}^{x_2} \sqrt{2m(k + V(x))} dx. \quad (16)$$

Thus (14) is time independent if $K = -E_0$ and, in this case, it reduces to

$$\text{Const} \exp\left[-\hbar^{-1} \int_{x_1}^{x_2} \sqrt{2m(V(x) - E_0)} dx\right]. \quad (17)$$

Now it may happen that E_0 is so low that x_1 does not exist. Then there is no solution for q_0 with $K = -E_0$.²² The best one can do is take $K = -V(a)$. In this case (14) reduces to

$$\text{Const exp} \left[-\hbar^{-1} \int_{x_1}^{x_2} \sqrt{2m(K + V(x))} dx \right] \times \exp[\hbar^{-1}(E_0 + K)(t - t_0)]. \quad (18)$$

Since $K < -E_0$, one has exponential decay with increasing t .

These results can be considered as a generalization of Yasue's.⁸ As approximate calculations, their limitations should be recognized. First, the fact that (17) is the WKB approximation of EQM is interesting, but misleading. One could, in general, take K to be larger than $-E_0$ [but less than $-V(a)$], although $K = -E_0$ is, in some sense, a "minimum fluctuation." Hence the time independence of (17) is illusory. Moreover, one knows very well that (17) is not correct if $t - t_0 = \beta\hbar$, since

$$\lim_{t \rightarrow t_0} p(x, t | x_0, t_0) = \delta(x - x_0) \quad (19)$$

implies

$$\lim_{t - t_0 \rightarrow n\beta\hbar} p(x, t | x_0, t_0) = \delta(x - x_0). \quad (20)$$

The same observation shows that the exponential decay in (18) can only go so far, up till $t - t_0 = \beta\hbar$.

Nevertheless, (17) and (18) show that the phenomenon of "tunneling" exists in stochastic physics, although the name here is perhaps inappropriate. What happens is that the fluctuations can be so large that there is a finite probability that the particle "climbs the hill" and arrives at the "true vacuum". However, once it gets there, it does not necessarily stay there as it would in classical physics, as one can easily show by calculating the transition probability for the fluctuation around $\bar{x} = b$; it can go back. This provides a real-time realization of instantons.^{14,15,17}

When $t - t_0 = \beta\hbar$, (20) shows that the only interesting p to calculate is the one for equal endpoints. In this case, it is more interesting to calculate $G\{J\}$, which is related to $p(0, \beta\hbar/2 | 0, -\beta\hbar/2)$ (see Ref. 3). In fact, once it has been shown that tunneling exists, $G\{J\}$ is the interesting object, since it determines the moments. Unlike EQM, however, SM has no direct method for determining the partition function.

Up to the second variation of S^J , where

$$S^J \equiv \int_{-T}^T (\frac{1}{2}m\dot{q}^2 + U(q) + J \cdot q) dt, \quad (21)$$

($T = \beta\hbar/2$ to ease notation) one has

$$G\{J\} \simeq \text{Const exp} \left[-\hbar^{-1} S^J\{q_0^J\} \right] D\{Q\} \times \exp \left[-\hbar^{-1} \int_{-T}^T (\frac{1}{2}m\dot{Q}^2 + \frac{1}{2}U_{ij}(q_0^J)Q_i Q_j) dt \right]. \quad (22)$$

$q_0^J(\pm T) = 0$ would seem to leave $Q(\pm T)$ arbitrary, but $Q(\pm T) = 0$ is seen to be correct by considering $q_0^J(\pm T) = \epsilon$ and taking the limit $\epsilon \rightarrow 0$. q_0^J is the path making S^J stationary i.e.

$$-mq_{0i}^J + U_i(q_0^J) + J_i = 0. \quad (23)$$

Note that there is no problem with arbitrary endpoint condi-

tions as in EQM.^{15,17} This is because $G\{J\}$ can only be indirectly related to the partition function through the average energy when this can be calculated from the moments,^{3,5} although both SM and EQM produce the same partition function when it can be calculated in the former.

The oscillator calculation mentioned in the previous paper³ is really a semiclassical approximation made exact by the fact that the path integral term in (22) does not depend on J . Hence one finds that

$$G\{J\} = \text{Const exp} \left[-\frac{1}{2m\hbar} \int_{-T}^T \int_{-T}^T ds J(t)G(t-s)J(s) \right], \quad (24)$$

where the second G is the Green's function for the differential operator $d^2/dt^2 - \omega_0^2$ with periodic boundary conditions. This enters directly in the calculation of q_0^J , which is why it appears in (24), i.e., the exponential is just $\exp[-S_0^J\{q_0^J\}/\hbar]$.

The oscillator example in fact shows how to carry out the first step in the algorithm for calculating $G\{J\}$. From (23), one calculates the Green's function and determines q_0^J . Hence $S^J\{q_0^J\}$ and $U_{ij}(q_0^J)$ are known. For the oscillator one needs go no further, since U_{ij} is independent of q_0^J , but for the general case, one must calculate the path integral. This is done by shifting techniques.²³ Choose the matrix N so that

$$\dot{N}_{ij} = m^{-1}U_{ik}(q_0^J)N_{kj}. \quad (25)$$

Then (22) reduces to

$$G\{J\} \simeq \text{Const exp} \left[-\hbar^{-1} S^J\{q_0^J\} \right] \times |N(T)N(-T)|^{1/2} \left| \int_{-T}^T N_{ki}^{-1}(t)N_{ki}^{-1}(t)dt \right|. \quad (26)$$

This completes the algorithm for the semiclassical approximation.

Thus the problem of solving for $G\{J\}$ in the semiclassical approximation reduces to one of solving the differential equations (23) and (25) and calculating the integrals and determinants in (26). Because J is arbitrary, these steps are not too amenable for solving with a computer, but at least in (26) there is no longer any reference to the path integral.

III. FIELD THEORY

The semiclassical approximation and tunneling are particularly important in SFT when there exist topological conservation laws. These appear in classical nonabelian gauge theories, for example, so one should expect that the nontrivial structures found there (e.g., the monopole of 't Hooft and Polyakov²⁴ and the instanton of Polyakov and collaborators²⁵) become important in SFT.

Given a solitary wave solution (soliton) that is independent of time, one can expand around this solution to find the stochastic fluctuations. Yasue has already done this for the zero-point scalar field.²⁶ It is a simple matter to see that the only change for positive temperature is in the second moment, which can be written

$$\begin{aligned} & \langle (\varphi(x, t) - \varphi_0(x, t))(\varphi(y, t') - \varphi_0(y, t')) \rangle \\ &= \frac{1}{2\pi} \int d^3\mathbf{k} \beta^{-1} \sum_{n=-\infty}^{\infty} (\omega_k^2 + \omega_n^2)^{-1} e^{i\omega_n(t-t')}. \end{aligned} \quad (27)$$

But here the semiclassical approximation is not needed, so details will not be considered.

More interesting, perhaps, are Yang-Mills fields. Here the semiclassical paths, called calorons in the context of EQM,¹⁶ are generalizations of the Polyakov instanton.²⁵ The same case of a pure Yang-Mills SU(2) gauge field will be considered here. The results will be a positive-temperature generalization of those of Yasue for zero temperature.⁹

The $A_0 = 0$ gauge is appropriate for SFT since it allows the separation of a potential term (see below). The fact that it is not Lorentz invariant is irrelevant since the Lorentz frame where the ensemble is defined is privileged, even for zero temperature.^{3,5} Moreover, the $A_0 = 0$ gauge is the one used in most instanton calculations.

The standard Lagrangian is

$$\mathcal{L} = \frac{1}{4} F_{\mu\nu}^a F^{a\mu\nu}, \quad (28)$$

where

$$F_{\mu\nu}^a \equiv \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + \epsilon^{abc} A_\mu^b A_\nu^c. \quad (29)$$

(Latin indices are the SU(2) indices and Greek indices are the space-time ones. The ϵ^{abc} are the structure constants of the SU(2) Lie algebra.) Define electric and magnetic gauge fields by

$$E_i^a \equiv F_{0i}^a, \quad B_i^a \equiv \frac{1}{2} \epsilon^{ijk} F_{jk}^a. \quad (30)$$

Then

$$\mathcal{L} = \frac{1}{2} (E_i^a E_i^a - B_i^a B_i^a) = \frac{1}{2} (\dot{A}_i^a \dot{A}_i^a - B_i^a B_i^a). \quad (31)$$

B^a depends on the vector potential A^a but not on \dot{A}^a , so one can separate kinetic and potential energy terms. Classical equilibrium points are determined by time independent field configurations for which the potential energy term, $\int \frac{1}{2} (B^a)^2 d^3\mathbf{x}$, is zero. These are pure gauge fields and are divided up in homotopy classes labeled by the Pontryagin index. For the purposes of this paper, one such field will be selected from each homotopy class and will be denoted by ${}^r A^a$, where r is the Pontryagin index.

Any one of the ${}^r A^a$ can be used as a base point for discussing stochastic fluctuations around it. According to the path-integral description of the previous paper,³ the transition probability from ${}^r A^a$ at time t_0 to ${}^{r+s} A^a$ at time t is given by

$$\begin{aligned} p({}^{r+s} A^a, t | {}^r A^a, t_0) &= \sqrt{\rho({}^{r+s} A^a) / \rho({}^r A^a)} \int \mathcal{D}\{A^a\} \\ &\times \exp \left[- \int_{t_0}^t \left(\frac{1}{2} \dot{A}^a \right)^2 + \frac{1}{2} (B^a)^2 dt d^3\mathbf{x} \right] e^{E_i^a(t-t_0)}. \end{aligned} \quad (32)$$

Here the semiclassical approximation must be handled with care. There is no classical field with finite action which connects the two vacua. Nevertheless, because the right-hand side of (32) is well defined by a judicious handling of nonstandard quantities,^{3,4} an infinite-action path is permissible in principle. This will now be shown to be true in practice.

Let A_0^a be such a path that connects ${}^r A^a$ to ${}^{r+s} A^a$. Write

it in terms of its standard components³:

$$A_0^a(\mathbf{x}, t) = \left[\sum_{n < N} \mathbf{q}_{0n}^{a(N)}(t) e_n(\mathbf{x}) \right]. \quad (33)$$

Since $\ddot{A}_0^a - \Delta A_0^a = 0$, the equation for $\mathbf{q}_{0n}^{a(N)}$ is

$$\mathbf{q}_{0n}^{a(N)} - k_n^2 \mathbf{q}_{0n}^{a(N)} = 0. \quad (34)$$

Thus the explicit form for (32) has the right-hand side in the semiclassical approximation equal to

$$\begin{aligned} & \left[\sqrt{\frac{\rho^{(N)}(\mathbf{q}_0^{a(N)}(t))}{\rho^{(N)}(\mathbf{q}_0^{a(N)}(t_0))}} \exp[E_0^{(N)}(t-t_0)] \right. \\ & \times \exp \left(- \int_{t_0}^t \left(\frac{1}{2} (\mathbf{q}_0^{a(N)})^2 + \frac{1}{2} (k_n \mathbf{q}_0^{a(N)})^2 + \frac{1}{2} (k_n \mathbf{q}_r^{a(N)})^2 \right) dt' \right) \left. \right] \end{aligned} \quad (35)$$

where $E_0^{(N)}$ is the average zero-point energy of the N th component process. Each N th component of (35) is now finite and the equivalence class therefore defines a nonstandard real number. Note that the base point for the expansion does not appear in Eq. (34) due to the linear force, but it does appear in (35) in the potential. However, this just introduces the constant term

$$\exp \left[- \frac{1}{2} \int_{t_0}^t (k_n \mathbf{q}_r^{a(N)})^2 dt' \right], \quad (36)$$

where the $\mathbf{q}_r^{a(N)}$ are the time-independent components of the base field ${}^r A^a$.

Hence it is possible to give a more rigorous interpretation to Yasue's Sec. V⁹ where the extension to positive temperatures has been shown to be direct as long as $t - t_0 < \beta \hbar$.

One could now proceed to construct the generating functional $G\{J\}$. However, it is easier to calculate the field directly using the construction introduced by Guerra and Ruggiero⁷ and employed by the present author in order to construct the electromagnetic field.² The technique is basically the same for this case. The fluctuation field has moments

$$\langle A_i^a(\mathbf{x}, t) \rangle = 0, \quad (37)$$

$$\begin{aligned} \langle A_i^a(\mathbf{x}, t) A_j^b(\mathbf{y}, t') \rangle &= \delta_{ab} \int d^3\mathbf{k} \beta^{-1} \sum_{n=-\infty}^{\infty} (\omega_k^2 + \omega_n^2)^{-1} \\ &\times e^{i\omega_n(t-t')} (\delta_{ij} - k_i k_j / |\mathbf{k}|^2). \end{aligned} \quad (38)$$

Hence the complete field corresponding to ${}^r A^a$ plus fluctuations has average ${}^r A^a$ and covariance equal to the right-hand side of (38). In the limit $\beta \rightarrow \infty$, one has the zero-point field. This should be compared to Eq. (4.10) of Ref. 9. The only difference is the polarization vector sum.

ACKNOWLEDGMENTS

The author wishes to thank Professors A. M. Rodriguez-Vargas, A. Rueda, and J. Vargas for stimulating discussions. He also wishes to express his gratitude to Professor K. Yasue for interesting correspondence and for sending him prints and reprints. Many of the results in this paper are generalizations of his results in the positive-temperature case.

- ¹S. M. Moore, *Found. Phys.* **9**, 237 (1979).
²S. M. Moore, *Lett. Nuovo Cimento* **24**, 284 (1979).
³S. M. Moore, *J. Math. Phys.* **21**, 2102 (1980).
⁴S. M. Moore, *Rev. Colombiana Mat.* **14**, 73 (1980).
⁵S. M. Moore, UNIANDES Preprint 28-6-80.
⁶E. Nelson, *Phys. Rev.* **150**, 1079 (1966); *Dynamical Theories of Brownian Motion* (Princeton U. P., Princeton, N. J., 1967).
⁷F. Guerra and P. Ruggiero, *Phys. Rev. Lett.* **31**, 1022 (1973).
⁸K. Yasue, *Phys. Rev. Lett.* **40**, 665 (1978).
⁹K. Yasue, *Phys. Rev. D* **18**, 532 (1978).
¹⁰K. Yasue, *J. Math. Phys.* **19**, 1892 (1978).
¹¹K. Yasue, *J. Math. Phys.* **20**, 1861 (1979); "Quantum mechanics and stochastic control theory" (preprint).
¹²G. Jona-Lasinio, Università di Roma preprint no. 138.
¹³A. D. Ventsel' and M. I. Freidlin, *Russian Math. Surveys* **25**, 1 (1970).
¹⁴S. Coleman, Ettore Majorana lectures, 1977.
¹⁵B. J. Harrington, *Phys. Rev. D* **18**, 2982 (1978).
¹⁶B. J. Harrington and H. K. Shepard, *Phys. Rev. D* **17**, 2122 (1978); **D 18**, 2990 (1978).
¹⁷L. Dolan and J. Kiskis, *Phys. Rev. D* **20**, 505 (1979).
¹⁸L. De la Peña and A. M. Cetto, *J. Math. Phys.* **18**, 1612 (1977); **20**, 469 (1979); *Found. Phys.* **8**, 191 (1978).
¹⁹See, for example, M. Lax, *Rev. Mod. Phys.* **38**, 541 (1966).
²⁰In order to simplify notation, U_i denotes the vector $\partial U / \partial x_i$, and U_{ij} denotes the tensor $\partial^2 U / \partial x_i \partial x_j$.
²¹Let ϵ be such that $t - t_0 + \epsilon = \beta \hbar$. Extend q continuously to the left from $q(t_0) = x_0$ to $q(t_0 - \epsilon/2) = 0$, and to the right from $q(t) = x$ to $q(t + \epsilon/2) = 0$.
²²This is because the machinery only works for equilibrium processes. An unbounded q_0 takes the system out of equilibrium.
²³I. M. Gel'fand and A. M. Yaglom, *J. Math. Phys.* **1**, 48 (1960).
²⁴G. 't Hooft, *Nucl. Phys. B* **79**, 276 (1974); A. M. Polyakov, *JETP Lett.* **20**, 194 (1974).
²⁵A. A. Belavin, A. M. Polyakov, A. S. Schwartz, and Yu. S. Tyupkin, *Phys. Lett. B* **59**, 85 (1975).
²⁶K. Yasue, *Phys. Lett. B* **73**, 302 (1978).

On the infinite volume limit of the strongly coupled Yukawa₂ model

Guy A. Battle

Mathematics Department, Texas A&M University, College Station, Texas 77843

Lon Rosen

Mathematics Department, University of British Columbia, Vancouver, B.C., V6T1Y4, Canada

(Received 8 September 1980; accepted for publication 21 November 1980)

Using the FKG inequality, we construct infinite volume expectations of products of boson fields and fermi currents $(\bar{\psi}\psi)_{\text{ren}}$ for the scalar Yukawa₂ model with arbitrary coupling constant. These expectations satisfy the Osterwalder–Schrader axioms with the possible exception of clustering.

PACS numbers: 03.70. + k

1. INTRODUCTION

This paper is a sequel to our paper¹ on the FKG inequality for the Yukawa₂ (Y_2) quantum field model. Our main motivation for establishing the FKG inequality was to be able to carry over to the Y_2 model with arbitrary coupling the method used by Fröhlich and Simon² to construct the infinite volume $P(\phi)_2$ model. (The infinite volume limit for Y_2 has already been controlled for weak coupling,^{3,4} for large external field⁵ and for very large coupling in the pseudoscalar case.⁶) Actually, since the FKG inequality is a statement about functions of the boson field only (see Theorem 1.1), we are able to control the infinite volume limit for expectations of boson fields but not of general fermion fields. However, as we show in Sec. 4, our control on the boson subtheory can be extended to certain fermi currents via the boson field equation.

As usual,⁷ the starting point of a rigorous analysis of the Euclidean Y_2 model is the Matthews–Salam–Seiler formulation where the fermions have been “integrated out.” With a volume cutoff $\Lambda \subset \mathbb{R}^2$, the expectation of a function $F(\phi)$ of the boson field ϕ is

$$\langle F \rangle_\Lambda = \int F d\nu_\Lambda,$$

where $d\nu_\Lambda$ is a probability measure on $\mathcal{S}'(\mathbb{R}^2)$ of the form $d\nu_\Lambda = \rho_\Lambda d\mu$, where $d\mu$ is the free boson measure⁸ with mass $m_b > 0$ and (formally)

$$\rho_\Lambda(\phi) = \det(1 - K_\Lambda) e^{\text{Tr} K_\Lambda - \frac{1}{2} \text{Tr} K_\Lambda^* K_\Lambda - e_\Lambda}, \quad (1.1)$$

In (1.1), $K_\Lambda = \lambda S \phi \chi_\Lambda \Gamma$ where λ is the coupling constant, S is the free Fermi two-point function,

$$S(x, y) = \frac{1}{(2\pi)^2} \int d^2 p e^{ip(x-y)} \begin{pmatrix} m_f - ip_1 & -ip_0 \\ -ip_0 & m_f + ip_1 \end{pmatrix}^{-1}, \quad (1.2)$$

χ_Λ is the characteristic function of λ , $\Gamma = 1$ for scalar Y_2 and $\Gamma = \gamma_5 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ for pseudoscalar Y_2 , and $m_f \geq 0$ is the bare Fermi mass. The term

$$\frac{1}{2} \text{Tr}: K_\Lambda^* K_\Lambda : = \frac{1}{2} \delta m_b^2 \int_\Lambda \phi^2(x) dx,$$

is the boson mass counterterm, δm_b^2 and e_Λ being infinite

constants. When properly interpreted, (1.1) makes rigorous mathematical sense. We refer the reader to the papers cited above^{1,3-7} for a more detailed discussion of (1.1). Our main result in Ref. 1 was the following.

Theorem 1.1 (FKG inequality): Consider the scalar Y_2 model with $m_f \geq 0$ or the pseudoscalar Y_2 model with $m_f = 0$ with volume cutoff the rectangle Λ . Let \mathcal{S}_Λ be the class of increasing functions of the boson field,

$$\mathcal{S}_\Lambda = \{F(\phi(h_1), \dots, \phi(h_n)) | F: \mathbb{R}^n \rightarrow \mathbb{R} \text{ continuous and increasing; } h_j \in L^2(\Lambda), h_j \geq 0\}.$$

Then for $f, g \in \mathcal{S}_\Lambda$

$$\langle fg \rangle_\Lambda \geq \langle f \rangle_\Lambda \langle g \rangle_\Lambda, \quad (1.3)$$

provided f, g , and fg are $d\nu_\Lambda$ -integrable.

In Sec. 2, (1.3) is the main input for constructing the pure boson expectations for the infinite volume scalar Y_2 theory with $m_f > 0$. Briefly, we start with the large external field modification of the Y_2 interaction; the point is that the infinite volume limit of such a theory has already been shown to exist⁵ because the model may be transformed to a weakly coupled model in a way analogous to Spencer’s treatment⁹ of the $P(\phi)_2$ model with large external field. The weakly coupled model can then be controlled by the cluster expansion.^{3,4} Following Fröhlich and Simon,² we “turn off” the large external field by means of the FKG inequality (1.3). However, in our case there is a technical complication concerning boundary conditions: While the FKG inequality holds for the $P(\phi)_2$ model for quite general boundary conditions,¹⁰ (1.3) has been proven for the scalar Y_2 model only in the case of free boundary conditions; on the other hand the model transformation of Ref. 5 is based on *periodic* B.C.

In Sec. 3 we investigate whether the infinite volume states constructed in Sec. 2 satisfy clustering. Following Fröhlich and Simon’s² strategy for $P(\phi)_2$, we can reduce the question of clustering to a conjecture about the pressure (Conjecture 3.3). Unfortunately, we have been unable to verify this conjecture because the techniques available in the $P(\phi)_2$ case, such as conditioning, do not seem to have a useful analogue in the Y_2 case.

In Sec. 4 we extend our boson sub-theory for scalar Y_2 to a theory including both the boson field and the fermi current that is the nonhomogeneous term in the (Euclidean) boson field equation. We accomplish this by showing by a shift of the boson field that such an extension of the theory is not really an extension at all!

¹Research partially supported by the National Research Council of Canada.

In Sec. 5 we extend our analysis to the case where $m_f = 0$ and hence to the case of pseudoscalar Y_2 with $m_f = 0$.

We note also that our results continue to hold if we modify the Y_2 interaction by adding an additional $P(\phi)_2$ term.

Although the results of this paper show that correlation inequalities, which have been so powerful in the case of pure boson theories, also can play a useful role in the case of models involving fermions, there are several unfortunate limitations. The first, mentioned above, is that our proof of clustering is incomplete. A second limitation is that our construction is essentially limited to pure boson expectations. With the uniform bounds that are available for the Schwinger functions of Y_2 ^{11,12} one could also construct infinite volume fermi expectations by a compactness argument, but only the boson sub-theory of such a theory would satisfy Euclidean covariance. If one tries to remedy this problem by averaging the theory over the Euclidean group, the important OS positivity property is destroyed in general (see the remarks in Ref. 2).

A third limitation is that we should like to duplicate Theorem 5.2 of Ref. 2, i.e., we want differentiability of the "pressure" to imply that the infinite volume theory is unique and independent of the classical boundary conditions chosen. It seems that the stability of the FKG inequality for $P(\phi)_2$ under changes in boundary conditions is truly critical for the method of Ref. 2. This stability is not available to us. Finally, the nonlocality of the Y_2 interaction effectively rules out DLR equations of the type satisfied by $P(\phi)_2$ theories.^{2,13}

Finally, we wish to comment on the striking similarity between the results of this paper and (unpublished) results of Fröhlich and Park.¹⁴ Instead of using the Matthews–Salam–Seiler formalism, they "bosonize" the fermi interaction currents by representing them in terms of an auxiliary (massless) boson field σ . Roughly speaking, the representation is

$$\bar{\psi}\psi \equiv (1/2\pi) \cdot \cos(\sqrt{4\pi}\sigma), \quad \bar{\psi}\gamma^5\psi \equiv (1/2\pi) \cdot \sin(\sqrt{4\pi}\sigma), \quad (1.4)$$

where ψ is the massless Euclidean field. For example, for scalar Y_2 (with $m_f = 0$),

$$\det(1 - K_\Lambda) e^{\text{Tr}K_\Lambda} = \left\langle \exp \left[(\lambda/2\pi) \int_\Lambda \phi \cdot \cos(\sqrt{4\pi}\sigma) dx \right] \right\rangle_\sigma, \quad (1.5)$$

where $\langle \cdot \rangle_\sigma$ denotes $\int d\mu(\sigma)$. They then obtain Griffiths inequalities involving \cos such as

$$\langle : \cos(\sqrt{4\pi}\sigma(x)) : \rangle_\sigma \langle : \cos(\sqrt{4\pi}\sigma(y)) : \rangle_\sigma \geq \langle : \cos(\sqrt{4\pi}\sigma(x)) : \rangle_\sigma \langle : \cos(\sqrt{4\pi}\sigma(y)) : \rangle_\sigma. \quad (1.6)$$

In this way they can (formally) construct the infinite volume limit of expectations of products of ϕ 's for scalar Y_2 or pseudoscalar Y_2 with $m_f = 0$. Like our results, theirs seem restricted to $d = 2$ since the transformation (1.4) is peculiar to two dimensions! (However, there is a technical difficulty concerning the removal of a regularization in their approach, which, to the best of our knowledge, has not been solved.) The similarity between the two sets of results is not

surprising when one realizes that the basic inequality we used¹ to establish the FKG inequality,

$$\frac{\delta^2}{\delta\phi(x)\delta\phi(y)} \log \det(1 - K_\Lambda) \geq 0, \quad x \neq y,$$

is, by (1.5), nothing but (1.6)!

2. EXISTENCE OF THE BOSON SUBTHEORY

In this section we apply the FKG inequality for the scalar Y_2 model with space cutoff to the problem of constructing infinite volume measures for the boson subtheory. Although the arguments of Fröhlich and Simon² along these lines for the $P(\phi)_2$ models are of a fairly general nature, there is a technical difficulty to be overcome in our case: Periodic (or half-periodic) boundary conditions are required for the transformation⁵ that expresses a theory with a large external field in terms of a weakly coupled theory; on the other hand, free boundary conditions are required for the FKG inequality. To overcome this mismatch, we shall impose *two* volume cutoffs Λ and Λ' on the transformed theory with $\Lambda' \subset \Lambda$ and with free boundary conditions on Λ' and half-periodic boundary conditions on Λ . We remove both of these cutoffs via the cluster expansion, taking care to remove the half-periodic cutoff first. This gives us the FKG inequality for the transformed model and also for the original model.

We begin by recapitulating the model transformation of Ref. 5. Let $E_{\pm, \Lambda}$ denote the pure boson expectation functional for the scalar Y_2 model with boson mass $m_b > 0$, fermion mass $m_f > 0$, external field $\pm \mu_\infty$, interaction volume the $l_0 \times l_1$ rectangle Λ , and half-periodic BC; i.e.,

$$E_{\pm, \Lambda}(F) = \frac{\int F(\phi) e^{\pm \mu_\infty \phi(x_\Lambda)} \rho_\Lambda^{\text{HP}}(\phi) d\mu_\Lambda^{\text{P}}}{\int e^{\pm \mu_\infty \phi(x_\Lambda)} \rho_\Lambda^{\text{HP}}(\phi) d\mu_\Lambda^{\text{P}}}, \quad (2.1)$$

where $d\mu_\Lambda^{\text{P}}$ is the free boson measure with periodic BC on Λ , and (formally)

$$\rho_\Lambda^{\text{HP}}(\phi) = \text{const} \det(1 - K_\Lambda^{\text{P}}) e^{\text{Tr}K_\Lambda - 1; \text{Tr}K_\Lambda^{\text{P}} K_\Lambda}, \quad (2.2)$$

where $K_\Lambda = S\phi\chi_\Lambda$ as in (1.1), whereas $K_\Lambda^{\text{P}} = S_\Lambda^{\text{P}}\phi\chi_\Lambda$ with S_Λ^{P} the Fermi two-point function with periodic BC,

$$S_\Lambda^{\text{P}} = \frac{1}{|\Lambda|} \sum_{p \in \Lambda^*} e^{ip(x-y)} \begin{pmatrix} m_f - ip_1 & -ip_0 \\ -ip_0 & m_f + ip_1 \end{pmatrix}^{-1}, \quad (2.3)$$

where Λ^* is the lattice $(2\pi/l_0)\mathbb{Z} \times (2\pi/l_1)\mathbb{Z}$. Note that we have set $\lambda = 1$ and that the Wick dots in (2.2) refer to Wick ordering with respect to $d\mu_\Lambda^{\text{P}}$ (although we could equally well use $d\mu$).

Let \tilde{E}_Λ denote the expectation functional for the scalar Y_2 model with boson mass \tilde{m}_b , zero external field, and half-periodic B.C in Λ . For $h \in L^2_0(\mathbb{R}^2)$, the space of L^2 functions of compact support, define

$$\phi_c(h) = \phi(h) + c \int h(x) dx.$$

The model transformation is given by the following.

Lemma 2.1: (a) Let

$$\Delta(m_f, \tilde{m}_f) = \frac{\tilde{m}_f(\tilde{m}_f^2 - m_f^2)}{2\pi^2} \int \frac{d^2p}{(p^2 + m_f^2)(p^2 + \tilde{m}_f^2)}.$$

For given m_b, m_f, μ_∞ , define $\tilde{m}_f = m_f - c_\pm$ and

$$\tilde{m}_b = [m_b^2 + 2\Delta(m_f, \tilde{m}_f)/\tilde{m}_f]^{1/2},$$

where c_{\pm} is chosen to solve

$$\pm \mu_{\infty} - c_{\pm} m_b^2 + \Delta(m_f, m_f - c_{\pm}) = 0. \quad (2.4)$$

Then for $h_1, \dots, h_n \in L^2(\Lambda)$

$$E_{\pm, \Lambda} \left(\prod_{i=1}^n \phi(h_i) \right) = \tilde{E}_{\Lambda} \left(\prod_{i=1}^n \phi_{c_{\pm}}(h_i) \right). \quad (2.5)$$

(b) For given m_b and m_f , \tilde{m}_b and $|\tilde{m}_f|$ can be made arbitrarily large by choosing μ_{∞} sufficiently large.

Remarks: (1) This lemma is essentially Lemma 2 of Ref. 5, but we have replaced the periodic BC used there by half-periodic BC. Half-periodic BC are more natural to work with. In particular, the transformation does not produce a small Λ -dependent boson self-interaction as in the case of periodic BC.

(2) The relation (2.4) holds for functions of ϕ more general than products (see Lemma 2.3 below).

(3) \tilde{E}_{Λ} depends on the choice of \pm sign since \tilde{m}_f and \tilde{m}_b do, but we suppress this dependence in the notation.

For our purposes we must also consider \tilde{E}_{Λ} with an additional space cutoff Λ' (which we take to be a rectangle) imposed by replacing ϕ by $\phi\chi_{\Lambda'}$. We denote this doubly cut-off expectation by \tilde{E}_{Λ}' . We now fix μ_{∞} sufficiently large so that by Lemma 2.1b the cluster expansion is applicable in the \tilde{E}_{Λ} theory. Then the limits

$$\tilde{E} \equiv \lim_{\Lambda \rightarrow \mathbb{R}^2} \tilde{E}_{\Lambda}, \quad (2.6)$$

$$\tilde{E}^{\Lambda'} \equiv \lim_{\Lambda \rightarrow \mathbb{R}^2} \tilde{E}_{\Lambda}', \quad (2.7)$$

and, by Lemma 2.1a,

$$E_{\pm} = \lim_{\Lambda \rightarrow \mathbb{R}^2} E_{\pm, \Lambda}, \quad (2.8)$$

exist. [If we do not specify otherwise, then we assume that the arguments of the various expectation functionals are of the form $\prod_{i=1}^n \phi(h_i)$ and that $h_i \in L^2_0(\mathbb{R}^2)$.] The following theorem is implicit in Ref. 5.

Theorem 2.2: Let c_{\pm} , $\tilde{E}^{\Lambda'}$, and E_{\pm} be given by (2.4), (2.7), and (2.8), respectively. Then

$$E_{\pm} \left[\prod \phi(h_i) \right] = \lim_{\Lambda' \rightarrow \mathbb{R}^2} \tilde{E}^{\Lambda'} \left[\prod \phi_{c_{\pm}}(h_i) \right]. \quad (2.9)$$

Proof: The convergence (2.7) of \tilde{E}_{Λ}' to $\tilde{E}^{\Lambda'}$ is uniform in Λ' , since the cluster expansion provides an error estimate with exponential decay in the distance from $\partial\Lambda$ to $\cup \text{supp } h_i$, uniformly in Λ' . Therefore, we may interchange limits:

$$\begin{aligned} \lim_{\Lambda'} \tilde{E}^{\Lambda'} &= \lim_{\Lambda'} \lim_{\Lambda} \tilde{E}_{\Lambda}'^{\Lambda'} = \lim_{\Lambda} \lim_{\Lambda'} \tilde{E}_{\Lambda}'^{\Lambda'} \\ &= \lim_{\Lambda} \tilde{E}_{\Lambda} \quad (\text{since } \tilde{E}_{\Lambda}'^{\Lambda'} = \tilde{E}_{\Lambda} \text{ if } \Lambda \subset \Lambda') \\ &= \tilde{E}. \end{aligned} \quad (2.10)$$

Equation (2.9) now follows from (2.5). \square

Theorem 2.2 overcomes the BC mismatch referred to above. For $\tilde{E}^{\Lambda'}$ is the expectation for the Y_2 model with boson mass \tilde{m}_b , fermi mass \tilde{m}_f , zero external field, interaction in Λ' , and *free* BC (the half-periodic cutoff has been removed). Accordingly, the FKG inequality of Theorem 1.1

applies to the right side of (2.9) and hence to the left side. [Note that the cutoff Λ' on \tilde{E} does not correspond to a cutoff on E_{\pm} but that we do not require such a cutoff in view of (2.9).]

In order for the FKG inequality to be effective we must check that (2.9) holds for a rich enough supply of increasing functions of ϕ . We merely sketch the arguments. [Notation: we shall drop the subscripts \pm on E and c in (2.9) and we shall write $F_c(\phi)$ for $F(\phi_c)$.]

We first note that (2.9) holds

$$E(F) = \lim_{\Lambda' \rightarrow \mathbb{R}^2} \tilde{E}^{\Lambda'}(F_c), \quad (2.11)$$

for functions of the form

$$F(\phi) = \exp \left[\phi(h) - \sum_{i=1}^n \kappa_i \phi(g_i)^2 \right], \quad (2.12)$$

where $h, g_1, \dots, g_n \in L^2_0$ and $\text{Re} \kappa_i \geq 0$. This follows, for $|\kappa_i|$ small, from expanding the exponential in (2.12) and appealing to Theorem 2.2 and the bound

$$\left| \tilde{E}^{\Lambda'} \left(\prod_{i=1}^n \phi(h_i) \right) \right| \leq C^n \sqrt{n!} \prod_{i=1}^n \|h_i\|_{-1}, \quad (2.13)$$

where $\|h\|_{-1}^2 \equiv \int |\hat{h}(k)|^2 (k^2 + \tilde{m}_b^2)^{-1} d^2k$ and C is a constant independent of Λ' , n , $\|h_1\|_{-1}, \dots, \|h_n\|_{-1}$ (but dependent on $\cup_i \text{supp } h_i$). The bound (2.13) is a consequence of cluster expansion estimates.^{3,4} But it follows from (2.13) that $\tilde{E}^{\Lambda'}(F(\phi_c))$ is analytic in $\kappa_i > 0$ and so, by the Vitali Convergence Theorem and the fact that $\Pi_i \{\kappa_i | \text{Re} \kappa_i > 0, |\kappa_i| < \epsilon\}$ is a determining set in the sense of analytic functions, we deduce the desired convergence.

Another type of random variable that will be of interest to us is the truncation of the field¹⁵

$$\sigma_c(h) = \begin{cases} -1, & \text{if } \phi_c(h) \leq -1, \\ 1, & \text{if } \phi_c(h) \geq 1, \\ \phi_c(h), & \text{otherwise.} \end{cases}$$

We claim that (2.11) holds for F of the form

$$F(\phi) = \prod_{i=1}^m \sigma_{c_i}(h_i). \quad (2.14)$$

To see this, we approximate F by

$$G(\phi) = \prod_{i=1}^m \sigma_{c_i}(h_i) e^{-\kappa_i \phi_{c_i}(h_i)^2},$$

with $\kappa_i > 0$. The error is controlled by the estimates

$$|\sigma_{c_i} - \sigma_{c_i} \exp[-\kappa_i \phi_{c_i}(h_i)^2]| \leq \kappa_i \sigma_{c_i}(h_i)^2,$$

and (2.13). To establish (2.11) for G , we apply the Stone-Weierstrass theorem to approximate G in *sup norm* by random variables

$$\prod_i P(\sigma_{c_i}(h_i)) e^{-\kappa_i \phi_{c_i}(h_i)^2},$$

where P is a polynomial. Since we know (2.11) holds for such random variables, it also holds for F given by (2.14).

Clearly, the above argument applies to any $F(\phi(h_1), \dots, \phi(h_m))$, where $F: \mathbb{R}^m \rightarrow \mathbb{R}$ is bounded and continuous. Moreover, it applies to such an F multiplied by an exponential and polynomial in the smeared fields. Accordingly, let

$$\mathcal{C} = \{F(\phi(h_1), \dots, \phi(h_m)) \mid m = 0, 1, 2, \dots; h_i \in L_0^2\};$$

$$F(\mathbf{x}) = P(\mathbf{x})G(\mathbf{x})e^{a \cdot \mathbf{x}};$$

P a polynomial, G bounded and continuous, $a \in \mathbb{R}^m$.

Obviously \mathcal{C} is closed under multiplication. We have checked the following.

Lemma 2.3: (2.11) holds for any $F \in \mathcal{C}$.

As in Ref. 2, we introduce the expectations

$$E_{\pm, \mu, \Lambda}(F) = E_{\pm}(F e^{(\mu \mp \mu_{\infty})\phi(\chi_{\Lambda})}) / E_{\pm}(e^{(\mu \mp \mu_{\infty})\phi(\chi_{\Lambda})}), \quad (2.15)$$

where taking the limit $\lim_{\Lambda \rightarrow \mathbb{R}^2} E_{\pm, 0, \Lambda}$ corresponds to "turning off the large external field." Combining Theorem 1.1 with Lemma 2.3, we obtain the key result of this section:

Theorem 2.4: Let $f, g \in \mathcal{C} \cap \mathcal{S}$ where the class of increasing function \mathcal{S} is defined like \mathcal{S}_{Λ} in Theorem 1.1, except that the condition $h_j \in L^2(\Lambda)$ is replaced by $h_j \in L_0^2$. Then

$$E_{\pm, \mu, \Lambda}(fg) \geq E_{\pm, \mu, \Lambda}(f)E_{\pm, \mu, \Lambda}(g). \quad (2.16)$$

Proof: We pass to a cutoff version $\tilde{E}_{\pm, \mu, \Lambda}^{\Lambda'}$ of (2.15) where c_{\pm} , \tilde{m}_f , and \tilde{m}_g are defined as in Lemma 2.1:

$$\tilde{E}_{\pm, \mu, \Lambda}^{\Lambda'}(F) \equiv \tilde{E}^{\Lambda'}(F e^{(\mu \mp \mu_{\infty})\phi(\chi_{\Lambda})}) / \tilde{E}^{\Lambda'}(e^{(\mu \mp \mu_{\infty})\phi(\chi_{\Lambda})}).$$

By Lemma 2.3, for $F \in \mathcal{C}$,

$$E_{\pm, \mu, \Lambda}(F) = \lim_{\Lambda'} \tilde{E}_{\pm, \mu, \Lambda}^{\Lambda'}(F c_{\pm}). \quad (2.17)$$

Now the FKG inequality holds for $\tilde{E}^{\Lambda'}$ and hence for $\tilde{E}_{\pm, \mu, \Lambda}^{\Lambda'}$ since the additional interaction term $e^{(\mu \mp \mu_{\infty})\phi(\chi_{\Lambda})}$ does not affect the proof of the FKG inequality. Moreover, if $F \in \mathcal{S}$, then clearly so does $F c_{\pm}$. Consequently, the inequality (2.16) holds for the expectations on the right of (2.17) and we obtain the theorem in the limit as $\Lambda' \rightarrow \mathbb{R}^2$. \square

We are now in a position to carry out the Fröhlich–Simon construction (Theorem 4.1 of Ref. 2) for the boson expectations of the scalar Y_2 model. First we recall some definitions.² If $\{d\nu_{t,l}\}$ is a two-parameter family of measures on $\mathcal{S}'(\mathbb{R}^2)$, then we say that $d\nu_{t,l}$ converges to the measure $d\nu$ by iteration if

$$d\nu = \lim_{l \rightarrow \infty} (\lim_{t \rightarrow \infty} d\nu_{t,l}),$$

where the limits are taken in the sense of characteristic functions. If $d\nu_1$ and $d\nu_2$ are two probability measures on $\mathcal{S}'(\mathbb{R}^2)$, then we say that

$$d\nu_1 \leq d\nu_2 \quad (\text{FKG}),$$

if for all $f \in \mathcal{C} \cap \mathcal{S}$

$$\int f d\nu_1 \leq \int f d\nu_2. \quad (2.18)$$

[No regularity properties are specified in Ref. 2 on the f 's occurring in (2.18), but in the case of the more involved Y_2 model we have at least shown that the various integrals make sense for $f \in \mathcal{C}$. An inspection of the proofs in Ref. 2 shows that actually only functions f in \mathcal{C} are ever used.]

We denote the probability measure⁸ on $\mathcal{S}'(\mathbb{R}^2)$ corresponding to $E_{\pm, \mu, \Lambda}$ by $d\nu_{\pm, \mu, \Lambda}$.

Theorem 2.5: (a) The limits $d\nu_{\pm, \mu} \equiv \lim_{\Lambda \rightarrow \mathbb{R}^2} d\nu_{\pm, \mu, \Lambda}$ exist

where the limit is taken in the sense of characteristic functions or in the sense of moments.

(b) If $\Lambda_{t,l} = \{(x_0, x_1) \mid |x_0| < t/2, |x_1| < l/2\}$, then $d\nu_{\pm, \mu, \Lambda_{t,l}} \rightarrow d\nu_{\pm, \mu}$ by iteration.

(c) The pure boson theories defined by the probability measures $d\nu_{\pm, \mu}$ satisfy all the Osterwalder–Schrader axioms¹⁶ with the possible exception of clustering.

(d) The Schwinger functions and Schwinger generating functionals for $d\nu_{\pm, \mu}$ are continuous in μ from the right and left, respectively.

(e) If $\mu' < \mu$, then $d\nu_{\pm, \mu'} \ll d\nu_{\pm, \mu}$ (FKG).

Discussion: Given Theorem 2.4, the proof of this theorem follows as in the $P(\phi)_2$ case.² In particular, for $\mu < \mu_{\infty}$ and $h > 0$, $E_{\pm, \mu, \Lambda}(e^{\phi(h)})$ is monotone decreasing in Λ by the inequality (2.16). The Vitali arguments work as in Ref. 2; the analyticity of the function

$$f(\lambda) = E_{\pm, \mu, \Lambda}(e^{\lambda\phi(h)}),$$

follows from the bound (2.13), and the necessary uniform bounds in Λ are implicit in the arguments of Seiler and Simon¹¹ (see the proof of Lemma 3.4 below for the essential idea).

We have been unable to prove that the infinite volume measures $d\nu_{\pm, \mu}$ constructed by this procedure are independent of the choice of μ_{∞} , as in the $P(\phi)_2$ case.²

3. CLUSTERING?

Let $E_{\pm, \mu}$ be the boson expectation functional corresponding to the measure $d\nu_{\pm, \mu}$ constructed in Theorem 2.5. $E_{\pm, \mu}$ satisfies all the OS axioms with the possible exception of clustering. By virtue of the FKG inequality and a theorem of Simon, the clustering axiom reduces² to the clustering of the two-point function $E_{\pm, \mu}(\phi(x)\phi(y))$. Fröhlich and Simon deduced this clustering for $P(\phi)_2$ from an identity for the pressure (Theorem 3.1 of Ref. 2). We have been unable to establish the corresponding identity for Y_2 but we conjecture it is true. Assuming this identity (Conjecture 3.3), we shall show that the Fröhlich–Simon strategy for proving clustering works for Y_2 . The reasoning is similar to the $P(\phi)_2$ case but we shall be careful with some of the details since the chessboard estimates of Ref. 2 are proven in a way that depends on the local nature of the interaction in the $P(\phi)_2$ model.

Consider the finite volume model corresponding to the expectation $\tilde{E}^{\Lambda'}$ of (2.7), i.e., the scalar Y_2 model with large fermi and boson masses \tilde{m}_f and \tilde{m}_b , no external field, and free BC on the boundary of the cutoff region Λ' . Now introduce the external field $(\mu \mp \mu_{\infty})\chi_{\Lambda}$, and define the pressure as follows:

$$\tilde{\alpha}_{\pm}(\mu) \equiv \lim_{\Lambda' \rightarrow \mathbb{R}^2} \frac{1}{|\Lambda'|} \log \int e^{(\mu \mp \mu_{\infty})\phi(\chi_{\Lambda'})} d\tilde{\nu}_{\Lambda'}, \quad (3.1)$$

where $d\tilde{\nu}_{\Lambda'}$ is defined as in (1.1) but with masses \tilde{m}_f and \tilde{m}_b . As an immediate consequence of Theorems 6.1 and 7.2 of Ref. 11, we have the following.

Lemma 3.1: (a) $\tilde{\alpha}_{\pm}(\mu)$ exists.

(b) $\tilde{\alpha}_{\pm}(\mu)$ is convex and hence continuous in μ .

(c) For $f \in C_0^{\infty}(\mathbb{R}^2)$,

$$\tilde{E}(e^{\phi(f)}) \ll \exp \left\{ \int dx [\tilde{\alpha}_{\pm}(\pm \mu_{\infty} + f(x)) - \tilde{\alpha}_{\pm}(\pm \mu_{\infty})] \right\}. \quad (3.2)$$

By (2.11) and (2.10)

$$E_{\pm}(e^{\phi^{\nu}}) = \tilde{E}(e^{\phi^{\nu}})e^{c_{\pm} \int f}, \quad (3.3)$$

so that if we set

$$\alpha_{\pm}(\mu) \equiv \tilde{\alpha}_{\pm}(\mu) + c_{\pm} \mu, \quad (3.4)$$

then (3.2) reads

$$E_{\pm}(e^{\phi^{\nu}}) \leq \exp \left\{ \int dx [\alpha_{\pm}(\pm \mu_{\infty} + f(x)) - \alpha_{\pm}(\pm \mu_{\infty})] \right\}. \quad (3.5)$$

Since we shall be using Lemma 6.4 of Ref. 11 in our estimates, we state it here for convenience.

Lemma 3.2: Let $f(x)$ be convex in the region $1 \leq x < \infty$

and linearly bounded from above. Then, $a = \lim_{x \rightarrow \infty} f(x)/x$ exists; for any $l_0 > 0$, $f(x) - f(x - l_0) \rightarrow a l_0$ as $x \rightarrow \infty$; and $f(x) - ax$ is monotone decreasing.

Let $\chi_{t,l}$ denote the characteristic function of the rectangle $\Lambda_{t,l}$. Our conjecture is the following

Conjecture 3.3:

$$\lim_{l \rightarrow \infty} \left[\lim_{t \rightarrow \infty} \frac{1}{lt} \log \tilde{E}(e^{(\mu \mp \mu_{\infty}) \phi(\chi_{t,l})}) \right] = \tilde{\alpha}_{\pm}(\mu) - \tilde{\alpha}_{\pm}(\pm \mu_{\infty}). \quad (3.6)$$

Remarks: (1) The identity (3.6) seems reasonable if one writes the left side as

$$\begin{aligned} & \lim_{\Lambda \rightarrow \mathbb{R}^2} \frac{1}{|\Lambda|} \log \tilde{E}(e^{(\mu \mp \mu_{\infty}) \phi(\chi_{\Lambda})}) \\ &= \lim_{\Lambda} \lim_{\Lambda'} \frac{1}{|\Lambda|} \log \tilde{E}^{\Lambda'}(e^{(\mu \pm \mu_{\infty}) \phi(\chi_{\Lambda})}) \\ &= \lim_{\Lambda} \lim_{\Lambda'} \frac{1}{|\Lambda|} \left[\log \int e^{(\mu \mp \mu_{\infty}) \phi(\chi_{\Lambda})} d\tilde{\nu}_{\Lambda'} - \log \int d\tilde{\nu}_{\Lambda'} \right], \end{aligned}$$

and considers the limit with $\Lambda = \Lambda'$. As a matter of fact, the inequality lhs (3.6) \leq rhs (3.6) is an immediate corollary of (3.2). It is the reverse inequality which is problematic. In the $P(\phi)_2$ case,² the reverse inequality is proved by conditioning with Dirichlet BC. But we are doubtful that there is a useful analog of conditioning for the Y_2 model because of the infinite counterterms involved.

(2) Perhaps it would have been more natural to formulate our conjecture in terms of the pressure for the $E_{\pm, \Lambda}$ model, but it is technically easier to work with free rather than half-periodic BC. By (3.3) and (3.4) we can state (3.6) equivalently in terms of E_{\pm} :

$$\lim_{l \rightarrow \infty} \left[\lim_{t \rightarrow \infty} \frac{1}{lt} \log E_{\pm}(e^{(\mu \mp \mu_{\infty}) \phi(\chi_{t,l})}) \right] = \alpha_{\pm}(\mu) - \alpha_{\pm}(\pm \mu_{\infty}). \quad (3.7)$$

For the remainder of this section we assume Conjecture 3.3. The next lemma supplies the necessary exponential bounds for applying the methods of Ref. 2 to prove clustering. The basic technique of the proof is to use OS positivity in the manner of Seiler and Simon.¹¹

Lemma 3.4: For $a \in \mathbb{R}$ and $\Delta \subset \mathbb{R}^2$ a rectangle

$$E_{\pm, \mu}(e^{a\phi(\chi_{\Delta})}) \leq \exp \{ |\Delta| [\alpha_{\pm}(\mu + a) - \alpha_{\pm}(\mu)] \}. \quad (3.8)$$

Proof: We take Δ to be the rectangle Λ_{l_0, l_1} centered at the origin. For convenience we shall concentrate on $E_{+, \mu}$ and drop the subscript $+$; the corresponding proof for $E_{-, \mu}$ is

identical. By the definition (2.14) and Theorem 2.5b

$$\begin{aligned} E_{\mu}(e^{a\phi(\chi_{\Delta})}) &= \lim_l \lim_t E(e^{a\phi(\chi_{\Delta}) + (\mu - \mu_{\infty}) \phi(\chi_{t,l})}) / E(e^{(\mu - \mu_{\infty}) \phi(\chi_{t,l})}) \\ &= \lim_l \lim_t F(l_0, l_1; t, l) / Z(t, l), \end{aligned} \quad (3.9)$$

where

$$\begin{aligned} F(l_0, l_1; t, l) &\equiv E(\exp[a\phi(\chi_{l_0, l_1}) + (\mu - \mu_{\infty}) \phi(\chi_{t,l})]) \\ Z(t, l) &\equiv E(e^{(\mu - \mu_{\infty}) \phi(\chi_{t,l})}). \end{aligned}$$

Now the state E satisfies OS positivity so that in particular

$$F(l_0, l_1; t, l) \leq F(2l_0, l_1; t + l_0, l) Z(t - l_0, l)^l, \quad (3.10)$$

where we have reflected about the line $x_0 = l_0/2$ and then appealed to the translation invariance of E to translate back in the x_0 direction by $l_0/2$. We repeat this process on the first factor in (3.10) by reflecting about the line $x_0 = l_0$, and so on. After m steps we obtain

$$F(l_0, l_1; t, l) \leq F(2^m l_0, l_1; t + (2^m - 1)l_0, l)^{2^{-m}} Z(t - l_0, l)^{1 - 2^{-m}}.$$

Reflecting n times in the x_1 direction, we have

$$\begin{aligned} F(l_0, l_1; t, l) &\leq F(2^m l_0, 2^n l_1; t + (2^m - 1)l_0, l + (2^n - 1)l_1)^{2^{-m-n}} \\ &\quad \cdot Z(t + (2^m - 1)l_0, l - l_1)^{2^{-m} - 2^{-m-n}} Z(t - l_0, l)^{1 - 2^{-m-n}}. \end{aligned} \quad (3.11)$$

We apply (3.5) to the first factor in (3.11) to dominate it by (assuming $l_0 < t$, $l_1 < l$)

$$\begin{aligned} & \exp \{ 2^{1/(m+n)} [2^{m+n} l_1 (\alpha(a + \mu) - \alpha(\mu_{\infty}))] \\ & \quad + ((t + (2^m - 1)l_0)(l + (2^n - 1)l_1) - 2^{m+n} l_0 l_1) \\ & \quad \times (\alpha(\mu) - \alpha(\mu_{\infty})) \}, \end{aligned}$$

which approaches, in the limit $m, n \rightarrow \infty$,

$$\exp[|\Delta| (\alpha(a + \mu) - \alpha(\mu_{\infty}))]. \quad (3.12)$$

By (3.5) the function $t \rightarrow \log Z(t, l)$ is linearly bounded, and by OS positivity the function is convex, so that by Lemma 3.2 there is an E_l such that as $t \rightarrow \infty$

$$Z(t, l)^{1/t} \rightarrow e^{E_l}, \quad (3.13)$$

and

$$Z(t - l_0, l) / Z(t, l) \rightarrow e^{-l_0 E_l}. \quad (3.14)$$

Moreover, $h \rightarrow E_l$ is also linearly bounded and convex so that by Lemma 3.2

$$\begin{aligned} \lim_{l \rightarrow \infty} (E_l - E_{l-l_1}) &= l_1 \lim_{l \rightarrow \infty} E_l / l \\ &= l_1 (\alpha(\mu) - \alpha(\mu_{\infty})), \end{aligned} \quad (3.15)$$

where in the last equality we have used the conjecture (3.7). By (3.13) the second factor in (3.11) approaches $\exp(l_0 E_{l-l_1})$ in the limit $m, n \rightarrow \infty$. Hence by (3.9), (3.11), and (3.12)

$$\begin{aligned} E_{\mu}(e^{a\phi(\chi_{\Delta})}) &\leq e^{|\Delta| (\alpha(a + \mu) - \alpha(\mu_{\infty}))} \lim_l \exp(l_0 E_{l-l_1}) \lim_t \frac{Z(t - l_0, l)}{Z(t, l)} \\ &= e^{|\Delta| (\alpha(a + \mu) - \alpha(\mu_{\infty}))} \lim_l \exp[l_0 (E_{l-l_1} - E_l)] \quad [\text{by (3.14)}] \\ &= e^{|\Delta| (\alpha(a + \mu) - \alpha(\mu_{\infty}))} e^{-|\Delta| (\alpha(\mu) - \alpha(\mu_{\infty}))} \quad [\text{by (3.15)}] \\ &= e^{|\Delta| (\alpha(a + \mu) - \alpha(\mu))}. \end{aligned}$$

The following theorem is the analog of Theorem 4.2 of Ref. 2 and, given Lemma 3.4, the proof is practically the same.

Theorem 3.5: $E_{\pm, \mu}(\phi(0)) = D^{\pm} \alpha_{\pm}(\mu)$, where D^{\pm} denotes the right and left derivative, respectively.

We have now developed all the machinery needed for the following clustering theorem whose proof is identical to Theorem 4.4 of Ref. 2.

Theorem 3.6: Assuming (3.7), the states $E_{\pm, \mu}$ satisfy all of the Osterwalder–Schrader axioms, including clustering.

4. FERMION CURRENTS

In this section we show how our control over the boson subtheory of the scalar Y_2 model can be extended to include certain fermion currents as well as boson fields. Let $\psi, \bar{\psi}$ be (formal) Euclidean fermion fields. We do not control general fermion currents of the form $(\bar{\psi}\Gamma\psi)_{\text{ren}}$ but only the current with $\Gamma = 1$ that occurs in the boson field equation, for it is actually the boson field equation that provides this control.

To motivate the objects that we consider below, we first work formally with the Fermion fields $\psi, \bar{\psi}$. Denoting the free Fermion expectation with periodic BC on Λ by $\langle \cdot \rangle_{\epsilon, \Lambda}$, we have

$$\rho_{\Lambda}^{\text{HP}}(\phi) = \langle e^{-U_{\Lambda}(\phi)} \rangle_{\epsilon, \Lambda}, \quad (4.1)$$

where

$$U_{\Lambda}(\phi) = \int_{\Lambda} \bar{\psi}(x)\psi(x) :_{\epsilon} \phi(x) dx + \frac{1}{2} \delta m_b^2 \int_{\Lambda} \phi(x)^2 dx + e_{\Lambda},$$

$:_{\epsilon}$ denotes Wick ordering with respect to the free Fermion expectation $\langle \cdot \rangle_{\epsilon}$ with free BC, and $: :$ denotes Wick ordering with respect to $d\mu$. If $g_1, \dots, g_n \in C_0^{\infty}(\Lambda)$ and $\lambda_1, \dots, \lambda_n \in \mathbb{R}$, we write $\lambda \cdot g = \sum_{j=1}^n \lambda_j g_j$. Now clearly

$$U_{\Lambda}(\phi - \lambda \cdot g) = U_{\Lambda}(\phi) - \lambda \cdot \int_{\Lambda} g(x) j(x) dx + \frac{1}{2} \delta m_b^2 \int_{\Lambda} (\lambda \cdot g(x))^2 dx, \quad (4.2)$$

where j is the renormalized Fermion current

$$j(x) \equiv :_{\epsilon} \bar{\psi}\psi :_{\epsilon}(x) + \delta m_b^2 \phi(x). \quad (4.3)$$

For $f \in C_0^{\infty}(\Lambda)$, we write $j(f) = \int j(x) f(x) dx$ for the smeared current.

If we require that the g_1, \dots, g_n have disjoint supports,

$$\text{supp } g_i \cap \text{supp } g_j = \emptyset, \quad i \neq j, \quad (4.4)$$

then by (4.2)

$$\frac{\partial^n}{\partial \lambda_1 \dots \partial \lambda_n} e^{-U_{\Lambda}(\phi - \lambda \cdot g)} \Big|_{\lambda=0} = \prod_{k=1}^n j(g_k) e^{-U_{\Lambda}(\phi)}, \quad (4.5)$$

since the quadratic term in (4.2) makes no contribution at $\lambda = 0$. From (4.5) and (4.1) we have, assuming (4.4),

$$\begin{aligned} & \int \langle \prod_{k=1}^n j(g_k) e^{-U_{\Lambda}(\phi)} \rangle_{\epsilon, \Lambda} e^{\phi(f)} d\mu_{\Lambda}(\phi) \\ &= \frac{\partial^n}{\partial \lambda_1 \dots \partial \lambda_n} \int \rho_{\Lambda}^{\text{HP}}(\phi - \lambda \cdot g) e^{\phi(f)} d\mu_{\Lambda}(\phi) \Big|_{\lambda=0}. \end{aligned} \quad (4.6)$$

As we show below, the right side of (4.6) is well-defined, i.e.,

$$\mathcal{S}_{\Lambda}(f, g) \equiv Z_{\Lambda}^{-1} \int \rho_{\Lambda}^{\text{HP}}(\phi - g) e^{\phi(f)} d\mu_{\Lambda}(\phi), \quad (4.7)$$

with $Z_{\Lambda} = \int \rho_{\Lambda}^{\text{HP}} d\mu_{\Lambda}$, is a well-defined generating functional for expectations of products of ϕ 's and f 's. We note that a "natural" expression for the generating functional like

$$Z_{\Lambda}^{-1} \int \langle e^{j(g)} e^{-U_{\Lambda}(\phi)} \rangle_{\epsilon, \Lambda} e^{\phi(f)} d\mu_{\Lambda}(\phi),$$

does not make sense because expectations of products of $j(g_k)$'s with overlapping arguments contain uncanceled infinities. Although it will not be necessary for our purposes, it is possible to evaluate the derivatives in (4.6) directly and to control the resulting expressions in a finite volume. For example,

$$\begin{aligned} \frac{\partial}{\partial \lambda_1} \rho_{\Lambda}^{\text{HP}}(\phi - \lambda \cdot g) \Big|_{\lambda=0} &= \{ \text{Tr} [(1 - K_{\Lambda})^{-1} S_{\Lambda}^p g_1] \\ &+ \delta m_b^2 \phi(g_1) \} \rho_{\Lambda}^{\text{HP}}(\phi), \end{aligned} \quad (4.8)$$

which, for $g_1 \in D((- \Delta + 1)^{\frac{1}{2}})$, can be shown to be in $L^p(d\mu_{\Lambda})$ for any $p < \infty$ by the methods of Ref. 4. As a matter of fact, a frontal attack on (4.6) consists of proving that: (a) the right side of (4.6) exists for fixed Λ ; (b) in the presence of a large external field $\pm \mu_{\infty}$, (4.7) and its λ -derivatives converge as $\Lambda \rightarrow \mathbb{R}^2$; (c) the large external field can be switched off by the correlation inequalities.

The analogs of steps (a) and (b) were carried out for weakly coupled ϕ_3^4 by Feldman and Raczka,¹⁷ but in fact there really is nothing to be done since we simply have to shift the field, as follows.

Lemma 4.1: For $f, g \in C_0^{\infty}(\Lambda)$,

$$\begin{aligned} \int \rho_{\Lambda}^{\text{HP}}(\phi - g) e^{\phi(f)} d\mu_{\Lambda}(\phi) &= e^{(g, f)} \int \rho_{\Lambda}^{\text{HP}}(\phi) e^{\phi(f)} \\ &: e^{-\phi((- \Delta + m_b^2)g)} : d\mu_{\Lambda}(\phi). \end{aligned} \quad (4.9)$$

Proof: Equation (4.9) follows immediately from the equations

$$\begin{aligned} \int \rho_{\Lambda}^{\text{HP}}(\phi - g) e^{\phi(f)} d\mu_{\Lambda}(\phi) &= \int \rho_{\Lambda}^{\text{HP}}(\phi) e^{(\phi + g)(f)} d\mu_{\Lambda}(\phi + g), \\ d\mu_{\Lambda}(\phi + g) &= e^{-\phi((- \Delta + m_b^2)g)} e^{-((- \Delta + m_b^2)g, g)/2} d\mu_{\Lambda}(\phi), \end{aligned}$$

and

$$: e^{-\phi((- \Delta + m_b^2)g)} : = e^{-((- \Delta + m_b^2)g, g)/2}.$$

Note that since $\text{supp } g$ is strictly contained in Λ , we may use the infinite volume Laplacian in place of the Laplacian with periodic BC on $\partial\Lambda$ in these equations.

From (4.9) we see that $\mathcal{S}_{\Lambda}(f, g)$ of (4.7) is a well-defined generating functional for expectations of products of ϕ 's and f 's. Moreover, as in Sec. 2, we can pass to the infinite volume by introducing an external field $\pm \mu_{\infty} \chi_{\Lambda}$ and then turning it off by the FKGI inequality to obtain the infinite volume generating functionals,

$$\mathcal{S}_{\pm}(f, g) = E_{\pm, 0}(e^{\phi(f) + (g, f)} : e^{-\phi((- \Delta + m_b^2)g)} :), \quad (4.10)$$

where $f, g \in C_0^{\infty}$.

It may seem surprising that expectations formally corresponding to the lhs of (4.6) can be rewritten as pure boson expectations until one realizes the physical meaning of (4.9). Replacing g with λg , differentiating both sides of (4.9) with respect to λ and then setting $\lambda = 0$, we get (formally)

$$\int \left\langle \int dx g(x) [(-\Delta + m_b^2)\phi(x) + j(x)] e^{-U_\Lambda(\phi)} \right\rangle_{f,\Lambda} e^{\phi(U)} d\mu_\Lambda = 0, \quad (4.11)$$

provided that f and g have disjoint supports. Equation (4.11) is the Euclidean boson field equation for the spatially cutoff Y_2 model. [The term involving $j(x)$ in (4.11) can be given a rigorous meaning in the Matthews–Salam–Seiler formalism via equation (4.8)].

In conclusion, we incorporate the Fermi currents in the following way: we construct the Schwinger functions for a theory that involves the two field ϕ and j . Formally these Schwinger functions are given by

$$S_\pm(f_1, \dots, f_m; g_1, \dots, g_n) = \left\langle \prod_{i=1}^m \phi(f_i) \prod_{k=1}^n j(g_k) \right\rangle_{\pm, 0}, \quad (4.12)$$

where the g_1, \dots, g_n have disjoint supports, and where $\langle \cdot \rangle_{\pm, 0}$ denotes $E_{\pm, 0}$ formally extended to smeared Fermi currents. The rigorous definition of these Schwinger functions uses the generating functional (4.10):

$$S_\pm(f_1, \dots, f_m; g_1, \dots, g_n) \equiv \frac{\partial^{m+n}}{\partial \kappa_1 \dots \partial \kappa_m \partial \lambda_1 \dots \partial \lambda_n} \times \mathcal{L}_\pm(\kappa \cdot f, \lambda \cdot g) |_{\kappa = \lambda = 0}. \quad (4.13)$$

Actually (4.13) makes sense for arbitrary $f_i, g_k \in C_0^\infty$ but bears the interpretation (4.12) only when the g_1, \dots, g_n have disjoint supports. In any event, the Osterwalder–Schrader reconstruction theorem¹⁶ involves a knowledge of the Schwinger functions with noncoincident arguments, and since the Schwinger functions satisfy all of the OS axioms except possibly clustering, we can analytically continue to the Wightman functions¹⁸ of a theory involving two fields $\Phi(x, t)$ and $J(x, t)$. The Euclidean field equation (4.11) analytically continues to the corresponding relativistic wave equation

$$(\partial_t^2 - \partial_x^2 + m_b^2)\Phi(x, t) + J(x, t) = 0.$$

Remark: If we were able to construct the Schwinger functions involving products of fermi fields it would presumably not be hard to identify J in terms of the basic boson and fermion fields [as in the relation (4.3)]. Such an analysis could be carried out for the weakly coupled model, but we refrain from doing so.

5. CASE OF MASSLESS FERMIONS

In the previous sections we assumed that $m_f > 0$. We now wish to extend our results to the case $m_f = 0$. A reexamination of our approach with this in mind shows that two changes are needed in the definition of ρ_Λ^{HP} of (2.2): (i) S_Λ^P is defined as in (2.3) but with $p = 0$ excluded from the sum over $p \in \Lambda^*$; (ii) the mass counterterm in (2.2) is chosen as

$\delta m_b^2 S_\Lambda : \phi^2$: with

$$\delta m_b^2 = \frac{2}{(2\pi)^2} \int_{|p|>1} \left(\frac{d^2 p}{p^2} \right). \quad (5.1)$$

The choice (5.1) has already been analyzed in Sec. 7 of Ref. 1 for the case of free BC. The case of periodic BC can be treated as a perturbation of the free BC by applying Theorem 2.2 of Ref. 1. We omit the details, but the conclusion is that $\rho_\Lambda^{\text{HP}} \in L^p(d\mu_\Lambda^p)$ and the arguments of Secs. 2–4 go through as before.

At first glance, one might expect further difficulties since the existing proofs^{11,12} of exponential bounds for Y_2 rely heavily on the assumption that $m_f > 0$. However, our approach requires such bounds only for the transformed model [see (3.2)] which has a large fermi mass \tilde{m}_f .

As we noted in the introduction, the results for $m_f = 0$ hold also for the pseudoscalar Y_2 model (where we know the FKG inequality only when $m_f = 0$).

¹G. Battle and L. Rosen, *J. Stat. Phys.* **22** 123–192 (1980).

²J. Fröhlich and B. Simon, *Ann. Math.* **105** 493–526 (1977).

³J. Magnen and R. Sénéor, *Commun. Math. Phys.* **51** 297–313 (1976).

⁴A. Cooper and L. Rosen, *Trans. Am. Math. Soc.* **234** 1–88 (1977).

⁵L. Rosen, *J. Math. Phys.* **18** 891–897 (1977).

⁶J. Balaban and K. Gawędzki, “A Low Temperature Expansion for the Pseudoscalar Yukawa Model of Quantum Fields in Two Space–Time Dimensions,” preprint.

⁷E. Seiler, *Commun. Math. Phys.* **42** 163–182 (1975).

⁸B. Simon, *The $P(\phi)_2$ Euclidean (Quantum) Field Theory* (Princeton University, Princeton, N.J., 1974).

⁹T. Spencer, *Commun. Math. Phys.* **39** 63–76 (1974).

¹⁰F. Guerra, L. Rosen, and B. Simon, *Ann. Inst. Henri Poincaré* **25** 231–334 (1976).

¹¹E. Seiler and B. Simon, *Ann. Phys.* **97** 470–518 (1976).

¹²O. McBryan, *Contribution to International Colloquium on Mathematical Methods of Quantum Field Theory*, Marseille, 1975.

¹³F. Guerra, L. Rosen, and B. Simon, *Ann. Math.* **101** 111–259 (1975).

¹⁴J. Fröhlich and Y. M. Park (private communication); see also J. Fröhlich and Y. M. Park, *Commun. Math. Phys.* **59** 235–266 (1978); and J. Fröhlich and E. Seiler, *Helv. Phys. Acta* **49** 889–924 (1976).

¹⁵B. Simon, *Commun. Math. Phys.* **31** 127–136 (1973).

¹⁶K. Osterwalder and R. Schrader, *Commun. Math. Phys.* **42** 281–305 (1975).

¹⁷J. Feldman and R. Raczka, *Ann. Phys.* **108** 212–229.

¹⁸R. Streater and A. S. Wightman, *PCT, Spin and Statistics and All That* (Benjamin, New York, 1964).

The path integral formulation of the dynamical map ^{a)}

Mark S. Swanson ^{b)}

Theoretical Physics Institute, Department of Physics, University of Alberta, Edmonton, Alberta, Canada, T6G 2J1

(Received 21 February 1979; accepted for publication 8 February 1980)

A modified path integral form for the generating functional of the dynamical map is developed in terms of canonical field theory. The Yang–Feldman equation for arbitrary operator products is derived and a simple form of the boson theorem is proved. The effects of internal symmetry and broken symmetry upon the dynamical map are investigated. Simple applications to free fields and an interacting case are exhibited.

PACS numbers: 03.70. + k

I. INTRODUCTION

In recent papers¹ Matsumoto, Umezawa, *et al.*, have developed the boson method and applied it to numerous model systems. In their approach the central object of interest is the dynamical map,² which relates the interacting field, assumed to obey a nonlinear equation of motion, to asymptotic in- or out-fields which obey free field equations. In this way the Hilbert space over which the interacting field is defined may be made consistent with the equation of motion. Spontaneous breakdown of symmetry is reflected in the asymptotic fields whose particles comprise the measurable spectrum, while the original symmetry operation on the interacting field is induced by operations on these asymptotic fields. This phenomenon is known as dynamical rearrangement of symmetry.³

The topological singularities and extended objects occurring in the field theories under consideration are seen to arise from the condensation of Goldstone bosons into the ground state. This condensation is manifested through the boson theorem⁴ which states that the equation of motion for the interacting field maintains the same form when the asymptotic fields in the dynamical map are translated by c -number functions which satisfy the respective free equations of motion. Analysis of a similar nature may be found in the work of Klauder.⁵ His investigations were limited to ultralocal models where no spatial gradient terms are present and the effect of translating the field by a non- L^2 c -number function could be rigorously discussed. Further work on this subject can be found in Hammer and DeFazio.⁶

For simplicity of calculation previous works have limited evaluation of the vacuum expectation value of the interacting field to the tree approximation,⁷ which becomes exact only in the limit that Planck's constant \hbar tends to zero. Since the path integral formalism⁸ provides a convenient method for determining quantum corrections in field theory, it is reasonable to expect similar results for a path integral representation of the dynamical map.

It should be noted that functional methods have been applied previously in this problem.⁹ In the context of previous work they were used to examine the Ward–Takahashi

identities for the Green's functions. However, the results developed in Sec. II of this paper yield an explicit recipe for constructing the dynamical map of the field operator and its retarded products given in terms of a modified path integral. Although the final results derived from this approach must coincide with those previously derived, the modified path integral makes many results clearer in both their origin and application. In Sec. III the properties of the dynamical map are examined, including an alternative proof of the boson theorem and the method for generating identities of the dynamical map similar to the Ward–Takahashi identities. Section IV contains applications of the techniques developed in the previous sections. Section V contains the suggestions for extension of this work.

Throughout this paper consideration will be limited to simple scalar field theories with at most a continuous phase invariance. There are several reasons for this limitation, the first being that the Ward–Takahashi identities derived from the path integral formalism fail to give any axial anomaly in theories where it is known to be present. As a result, rather than modify the path integral structure, in itself a worthy project, theories which would manifest this defect will be avoided. The second reason lies in the fact that the additional degrees of freedom in more complicated theories create problems in identifying the asymptotic spectrum of the field operators due to the presence of ghosts and bound states.

A final caveat to the reader in regard to the validity of the path integral as a generator of field operator products must be made. It is still an open question whether the measure over the fields exists if interactions are present. It is often written that evaluating the path integral is equivalent to solving the operator formulation of the same theory. The truth of this statement is not obvious to the author, but examination of this problem will not be made here.

II. THE DYNAMICAL MAP

In the following work consideration will be limited to theories describing a single scalar field; the generalization of these results will be discussed later in this section. The interacting field ψ is assumed to satisfy, in the weak sense, some nonlinear equation of motion.

$$\lambda (\partial_x) \psi(x) = F[\psi(x)]. \quad (2.1)$$

^{a)}Work supported by the National Research Council of Canada.

^{b)}Current address: Department of Physics, University of Connecticut, Storrs, Connecticut 06268.

The field $\psi(x)$ has the weak-limit asymptotic form

$$\text{w-lim}_{t \rightarrow -\infty} \psi(x) = \phi_0(x), \quad (2.2)$$

where ϕ_0 is the asymptotic in-field which satisfies the free field equation

$$\lambda (\partial_x) \phi_0(x) = 0. \quad (2.3)$$

Here the usual wavefunction renormalization constants are being suppressed for simplicity.

The LSZ reduction formula,¹⁰ coupled with the assumed completeness of the asymptotic states, yields the usual form of the dynamical map²:

$$\begin{aligned} G[\psi] &= \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \int dx_1 \dots dx_n : \phi_0(x_1) \lambda(\partial_{x_1}) \dots \phi_0(x_n) \lambda(\partial_{x_n}) \\ &\times \langle 0 | R \{ G[\psi] \psi(x_1) \dots \psi(x_n) \} | 0 \rangle_c, \end{aligned} \quad (2.4)$$

where the in-field representation has been chosen and the subscript c in (2.4) denotes the connected part of the Green's functions.¹¹ Here $G[\psi]$ is some function for functional of the interacting field or its retarded products and the retarded form of the connected Green's functions has been employed to be consistent with the use of in-fields.

It is then assumed that the connected retarded Green's function can be obtained from a generating function $W_r[J]$ by functional differentiation, i.e.,

$$\langle 0 | R \{ \psi(x_1) \dots \psi(x_n) \} | 0 \rangle_c = (-i)^n \frac{\delta^n W_r[J]}{\delta J(x_1) \dots \delta J(x_n)} \Big|_{J=0}. \quad (2.5)$$

The connection to canonical field theory is made when $W_r[J]$ takes the form¹¹

$$W_r[J] = \ln Z_r[J], \quad (2.6)$$

and the functional $Z_r[J]$ has the path integral representation¹²

$$Z_r[J] = N \int [d\phi] \exp i \int dx [\mathcal{L}(\phi) + J(\phi)], \quad (2.7)$$

where $\mathcal{L}(\phi)$ is the Lagrangian density of the theory and N is a normalizing constant defined by

$$Z_r[J=0] = 1. \quad (2.8)$$

Normally (2.6), taken with (2.7), is interpreted as the generating functional for the connected time-ordered Green's functions of the theory.¹³ However, with appropriate boundary conditions it becomes the generator for retarded or advanced Green's functions. This same result is well known from the LSZ reduction theorem, where it is seen that the retarded and time-ordered product formulations of S -matrix elements differ only over a set of measure zero,¹⁴ so that the expansions of the S matrix elements in terms of either are equivalent.

As a final note in this respect, and for future reference, expression (2.7) can be evaluated explicitly in the case that $\mathcal{L}(\phi)$ is a quadratic form, i.e.,

$$\mathcal{L}(\phi) = \frac{1}{2} \phi \lambda (\partial) \phi. \quad (2.9)$$

If (2.9) holds, then (2.7) becomes

$$Z_r[J] = \exp - \frac{1}{2} i \int dx dy J(x) \Delta(x-y) J(y), \quad (2.10)$$

where the single restriction on Δ necessary to perform the path integration is

$$\lambda (\partial x) \Delta(x-y) = \delta(x-y). \quad (2.11)$$

In order to be consistent with (2.5) it is necessary to apply the boundary conditions and identify Δ as the retarded Green's function.¹⁵

The development of the path integral form of the dynamical map begins by rewriting (2.4) as

$$\begin{aligned} G[\psi] &= \exp \left[-i \int dx \phi_0(x) \lambda(\partial_x) \frac{\delta}{i\delta J(x)} \right] \\ &\times G \left[\frac{\delta}{i\delta J} \right] W_r[J] \Big|_{J=0}, \end{aligned} \quad (2.12)$$

Due to the translational nature of the first functional operator appearing in (2.12), it follows that

$$\begin{aligned} &\exp \left[-i \int dx \phi_0(x) \lambda(\partial_x) \frac{\delta}{i\delta J(x)} \right] W_r[J] \\ &= \ln \left\{ \exp \left[-i \int dx \phi_0(x) \lambda(\partial_x) \frac{\delta}{i\delta J(x)} \right] Z_r[J] \right\}. \end{aligned} \quad (2.13)$$

It is then assumed that $\mathcal{L}(\phi)$ may be written

$$\mathcal{L}(\phi) = \frac{1}{2} \phi \lambda (\partial) \phi + \mathcal{L}_{\text{int}}(\phi), \quad (2.14)$$

so that, from (2.7), it follows that

$$\begin{aligned} &\exp \left[-i \int dx \phi_0(x) \lambda(\partial_x) \frac{\delta}{i\delta J(x)} \right] Z_r[J] \\ &= \exp \left\{ i \int dy \mathcal{L}_{\text{int}} \left[\frac{\delta}{i\delta J(x)} \right] \right\} N \int [d\phi] \exp i \\ &\times \int dx \left\{ \frac{1}{2} \phi \lambda (\partial_x) \phi + [J - \phi_0 \lambda (\partial_x)] \phi \right\}. \end{aligned} \quad (2.15)$$

The right-hand side of (2.15) may be evaluated using (2.10), (2.11), and (2.3) to obtain

$$\begin{aligned} &\frac{N}{N'} \exp \left\{ i \int dy \mathcal{L}_{\text{int}} \left[\frac{\delta}{i\delta J(x)} \right] \right\} \exp - \frac{1}{2} i \\ &\times \int dx dz J(x) \Delta(x-z) J(z) \exp i \int dx J(x) \phi_0(x), \end{aligned} \quad (2.16)$$

where N' is the constant necessary to normalize (2.10). However, it is apparent from (2.10) that

$$\begin{aligned} &\exp - \frac{1}{2} i \int dx dy J(x) \Delta(x-y) J(y) \exp i \int dx J(x) \phi_0(x) \\ &= N' \int [d\phi] \exp i \int dx \left[\frac{1}{2} \phi \lambda (\partial_x) \phi + J(\phi + \phi_0) \right], \end{aligned} \quad (2.17)$$

so that expression (2.12) becomes

$$G[\psi] = :G \left[\frac{\delta}{i\delta J} \right] W_r[J, \phi_0] \Big|_{J=0} :, \quad (2.18)$$

where

$$W_r[J, \phi_0] = \ln Z_r[J, \phi_0] \quad (2.19)$$

and

$$Z_r[J, \phi_0] = N \int [d\phi] \exp i \int dx [\frac{1}{2} \phi \lambda (\partial_x) \phi + \mathcal{L}_{int}(\phi + \phi_0) + J(\phi + \phi_0)]. \quad (2.20)$$

Relations (2.18)–(2.20) give the path integral form of the dynamical map for this simple case. However, it is necessary to note that difficulties may arise in the formulation of the dynamical map when several particle types or internal symmetries are present in the initial Lagrangian. Bound states, ghosts, and Goldstone bosons may occur, and some or all of these must be included in the dynamical map in order for the asymptotic states to be complete. Therefore, the generalization of (2.20) is *not* straightforward and depends critically upon the form of the dynamics and upon whether a symmetric or broken symmetric solution is being sought. In general, carefully applying the Ward–Takahashi identities reveals the asymptotic fields which are necessary to achieve completeness as well as the general features of the dynamical map. In this way the theory is made self-consistent. Furthermore, more cogent examples of this may be found in the existing literature.¹

III. PROPERTIES OF THE DYNAMICAL MAP

In this section certain properties of the generating functional (2.19) will be examined. For the purpose of making contact with classical field theory, it is convenient to introduce Planck's constant \hbar into the dynamical map. This is accomplished by multiplying the action appearing in (2.7) by \hbar^{-1} . Expansion of the dynamical map of an operator in powers of \hbar is equivalent to expansion of the same operator in terms of multiloop graphs, each loop carrying a power of \hbar .¹⁶

A. The Yang–Feldman equation

If (2.18) is applied, assuming (2.14) is valid, then

$$\begin{aligned} \psi(x; \phi_0) &\equiv: \frac{\hbar}{i} \frac{\delta W_r}{\delta J(x)} \Big|_{J=0} : \\ &= \phi_0(x) - \int dy \Delta(x-y) : Z_r^{-1}[J, \phi_0] \\ &\quad \times \mathcal{L}'_{int} \left[\frac{\hbar}{i} \frac{\delta}{\delta J(y)} \right] Z_r[J, \phi_0] \Big|_{J=0} : , \end{aligned} \quad (3.1)$$

where

$$\mathcal{L}'_{int}(\phi) = \frac{\partial \mathcal{L}_{int}(\phi)}{\partial \phi}, \quad (3.2)$$

and the commutation relation

$$\left[\frac{\delta^n}{\delta J(x)^n}, J(y) \right] = n \delta(x-y) \frac{\delta^{n-1}}{\delta J(x)^{n-1}} \quad (3.3)$$

has been used.

Relation (3.1) has the form of the usual Yang–Feldman equation¹⁷ for ψ , but exact identification *appears to be prevented by the fact that the Z_r functional, rather than W_r , appears on the right-hand side.* As a result, relation (2.1) does not seem to be generalized by (3.1).

The solution to this problem is more conceptual than mathematical in nature and lies in the idea that the quan-

tized nonlinear equation of motion should become the classical nonlinear equation of motion in the event that the vacuum expectation value is taken and \hbar is taken to zero. This is consistent with the correspondence principle and prevents certain conceptual difficulties associated with the inverse process, i.e., quantization of a classical field theory.

To see how this idea works *vis à vis* (3.1), it is easiest to examine a specific case. Suppose that

$$i\hbar^{-1} \mathcal{L}'_{int} \left[\frac{\hbar}{i} \frac{\delta}{\delta J} \right] = \left(\frac{1}{3} \right)^{\lambda \hbar^2} \frac{\delta^3}{\delta J^3}. \quad (3.4)$$

Then, in terms of W_r , (3.1) becomes

$$\begin{aligned} \psi(x; \phi_0) &= \phi_0 + \lambda \int dy \Delta(x-y) : \frac{\hbar^2}{i^2} \frac{\delta^2 W_r}{J(y)^2} \Big|_{J=0} \\ &\quad + \left[\frac{\hbar}{i} \frac{\delta W_r}{\delta J(y)} \Big|_{J=0} \right]^2 : . \end{aligned} \quad (3.5)$$

It is possible to show that

$$\psi^2(x; \phi_0) \equiv: \frac{\hbar}{i^2} \frac{\delta^2 W_r}{\delta J(x)^2} \Big|_{J=0} \quad (3.6)$$

goes to zero in the limit that \hbar goes to zero by the following point-splitting argument. Any perturbative expansion for the equal-time operator product $\psi(x + \epsilon, t) \psi(x - \epsilon, t)$ must decompose into two sets of graphs: the set which has a line (or lines) connecting the points $x + \epsilon$ and $x - \epsilon$, and the set which does not. By definition, the connected set is given by

$$\begin{aligned} &:\psi(x + \epsilon, t) \psi(x - \epsilon, t) :_{\text{connected}} : \\ &= : \frac{\hbar^2}{i^2} \frac{\delta^2 W_r}{\delta J(x + \epsilon, t) \delta J(x - \epsilon, t)} \Big|_{J=0} : . \end{aligned} \quad (3.7)$$

In the limit the ϵ tends to zero the connecting line (or lines) becomes a loop (or loops). Hence, in this limit the connected set is proportional at least to \hbar and must vanish when \hbar is zero. Noting that

$$\begin{aligned} \psi^2(x; \phi_0) &\equiv: \frac{\hbar^2}{i^2} \frac{\delta^2 W_r}{\delta J(x)^2} \Big|_{J=0} : \\ &= \lim_{\epsilon \rightarrow 0} : \frac{\hbar^2}{i^2} \frac{\delta^2 W_r}{\delta J(x + \epsilon, t) \delta J(x - \epsilon, t)} \Big|_{J=0} : \end{aligned} \quad (3.8)$$

the proof is complete.

The statement generalizes to arbitrary powers of the functional derivative, so that

$$\lim_{\hbar \rightarrow 0} \frac{\hbar^n}{i^n} \frac{\delta^n W_r}{\delta J(x)^n} \Big|_{J=0} = 0 \quad \forall n > 1. \quad (3.9)$$

The proof breaks down when n equals one since in that case the connected and disconnected sets coincide, so (3.5) becomes

$$\begin{aligned} \psi_0(x; \phi_0) &\equiv \lim_{\hbar \rightarrow 0} \psi(x; \phi_0) \\ &= \phi_0(x) + \lambda \int dy \Delta(x-y) : [\psi_0(x; \phi_0)]^2 : . \end{aligned} \quad (3.10)$$

Relation (3.10) generalizes to interactions involving arbitrary powers and to functions which have well-defined power series expansions.

Form (3.1) is then seen as necessary to ensure that the

right-hand side of (2.1) does not vanish when \hbar is zero and thus to maintain a classical limit for the quantum theory. Form (3.1) is then the path integral representation of the usual Yang–Feldman equation and satisfies the quantized nonlinear equation of motion

$$\lambda (\partial_x) \psi(x; \phi_0) = F[\psi](x; \phi_0) \quad (3.11)$$

and the asymptotic limit (2.2).

B. The boson theorem and vacuum behavior

A form of the boson theorem⁴ may be proved for the simple case being discussed here. Suppose that ϕ_0 appearing in the functional W_r is shifted by the c -number function $f(x)$ which satisfies

$$\lambda (\partial_x) f(x) = 0. \quad (3.12)$$

Then the field operator defined by

$$\psi^f(x; \phi_0) \equiv: \frac{\hbar}{i} \frac{\delta W_r[J, \phi_0 + f]}{\delta J(x)} \Big|_{J=0} : \quad (3.13)$$

satisfies the equation

$$\psi^f(x; \phi_0)$$

$$= \phi_0(x) + f(x) - \int dy \Delta(x-y) : Z_r^{-1}[J, \phi_0 + f] \times \mathcal{L}'_{\text{int}} \left[\frac{\hbar}{i} \frac{\delta}{\delta J(y)} \right] Z_r[J, \phi_0 + f] \Big|_{J=0} : \quad (3.14)$$

Making the obvious identification

$$F^f[\psi](x; \phi_0) = - : Z_r^{-1}[J, \phi_0 + f] \mathcal{L}'_{\text{int}} \left[\frac{\hbar}{i} \frac{\delta}{\delta J(x)} \right] \times Z_r[J, \phi_0 + f] \Big|_{J=0} : \quad (3.15)$$

shows that

$$\lambda (\partial_x) \psi^f(x; \phi_0) = F^f[\psi](x; \phi_0), \quad (3.16)$$

so that ψ^f satisfies the same nonlinear equation of motion as ψ , which is the usual boson theorem.

Result (3.16) gives a covariant method for determining the vacuum behavior of the theory under examination. Due to the normal ordering present in the dynamical map, the possible vacuum expectation values of the field operator and its retarded products are given by the formula

$$\langle 0 | G^f[\psi] | 0 \rangle = : G \left[\frac{\hbar}{i} \frac{\delta}{\delta J} \right] W_r[J, \phi_0 + f] \Big|_{J=\phi_0=0} : \quad (3.17)$$

where the only constraint on f is that it satisfy (3.12). Using (3.17) and the generalization of (3.10) shows that the c -number function

$$\phi^f(x) \equiv \langle 0 | \psi_0^f(x; \phi_0) | 0 \rangle \quad (3.18)$$

satisfies the classical differential equation

$$\lambda (\partial_x) \phi^f(x) = - \mathcal{L}'_{\text{int}}[\phi^f(x)], \quad (3.19)$$

so that the correspondence principle is satisfied.

C. Invariances of the dynamical map

In the event that the path integral is invariant under some internal symmetry operation, certain constraints are

placed upon the form of the dynamical map. This is best illustrated by a case where the form for $W_r[J, \phi_0]$ can be correctly generalized to one where several fields are present. Such a case is given by the Goldstone model¹⁸ where the classical equation of motion is given by

$$(\partial^2 - m^2)\phi = \lambda \phi^* \phi^2. \quad (3.20)$$

This equation is invariant under the phase transformation

$$\phi' = e^{i\theta} \phi, \quad \theta \in \mathbb{R}, \quad (3.21)$$

where θ is some arbitrary constant.

The generalization of (2.20) for this case is

$$\begin{aligned} Z_r[J, J^*, \phi_0, \phi_0^*] \\ = N \int [d\phi][d\phi^*] \exp \frac{i}{\hbar} \int dx \\ \times [\phi^* \lambda (\partial_x) \phi - \frac{1}{2} \lambda (\phi + \phi_0)^2 (\phi^* + \phi_0^*)^2 \\ + J^*(\phi + \phi_0) + J(\phi^* + \phi_0^*)]. \end{aligned} \quad (3.22)$$

Inserting the simultaneous phase transformations

$$J' = e^{i\theta} J, \quad \phi' = e^{i\theta} \phi, \quad \phi_0' = e^{i\theta} \phi_0 \quad (3.23)$$

into (3.22) shows that

$$W_r[J', J'^*, \phi_0', \phi_0'^*] = W_r[J, J^*, \phi_0, \phi_0^*]. \quad (3.24)$$

For θ infinitesimal this yields

$$\begin{aligned} \int dx \left[J(x) \frac{\delta W_r}{\delta J(x)} - J^*(x) \frac{\delta W_r}{\delta J^*(x)} \right. \\ \left. + : \phi_0(x) \frac{\delta W_r}{\delta \phi_0(x)} : - : \phi_0^*(x) \frac{\delta W_r}{\delta \phi_0^*(x)} : \right] = 0 \end{aligned} \quad (3.25)$$

which holds for arbitrary $J(x)$.

Relation (3.25) serves as a generator for an infinite number of relations between retarded products of the field operators. Of course, these relations follow from the Ward–Takahashi identities for the Green’s functions of the theory which originally appear in (2.4). For instance, differentiating (3.25) once and setting $J = 0$ yields

$$\psi(x; \phi_0, \phi_0^*) = \int dy \left[: \phi_0(y) \frac{\delta \psi(x)}{\delta \phi_0(y)} : - : \phi_0^*(y) \frac{\delta \psi(x)}{\delta \phi_0^*(y)} : \right]. \quad (3.26)$$

Further identities may be had from (3.26) by iterating the equation an arbitrary number of times.

It is easy to see that the transformation

$$\phi_0' = e^{i\theta} \phi_0 \quad (3.27)$$

on the in-field generates the transformation

$$\psi' = e^{i\theta} \psi, \quad (3.28)$$

i.e.,

$$\psi'(x; \phi_0', \phi_0'^*) = e^{i\theta} \psi(x; \phi_0, \phi_0^*). \quad (3.29)$$

The proof of (3.29) follows from the form (3.22) coupled with the invariance (3.24). It follows that

$$\begin{aligned} \psi'(x; \phi_0', \phi_0'^*) \\ = : \frac{\hbar}{i} \frac{\delta}{\delta J^*(x)} W_r[J', J'^*, \phi_0, \phi_0^*] \Big|_{J=0} : \\ = e^{i\theta} : \frac{\hbar}{i} \frac{\delta}{\delta J^*(x)} W_r[J, J^*, \phi_0, \phi_0^*] \Big|_{J=0} : \end{aligned}$$

$$= e^{i\theta} \psi(x; \phi_0, \phi_0^*) \quad (3.30)$$

It is apparent that the transformation (3.27) is also an invariance of the free field equation.

D. Broken symmetry solutions

Broken symmetries are accommodated in the path integral form (2.20) by shifting the in-field(s) by a constant and applying the self-consistency method to determine the mass associated with the asymptotic particle states. This method is illustrated in the following example of a hermitian scalar field.

The initial Lagrangian appearing in (2.20) takes the form

$$\mathcal{L}(\phi) = \frac{1}{2} \phi (\partial^2 - m_0^2) \phi + \mathcal{L}_{\text{int}}(\phi). \quad (3.31)$$

The quadratic mass term is treated as an interaction, so that Z_r becomes

$$Z_r[J, \phi_0] = N \int [d\phi] \exp i \int dx \left[\frac{1}{2} \phi \partial^2 \phi - m_0^2 (\phi + \phi_0)^2 + \mathcal{L}_{\text{int}}(\phi + \phi_0) + J(\phi + \phi_0) \right]. \quad (3.32)$$

The asymptotic field is then shifted by a constant ν and the terms linear in $(\phi + \phi_0)$ are removed by making the identification

$$m_0^2 \nu - \left. \frac{\partial^2 \mathcal{L}_{\text{int}}(\phi + \nu)}{\partial \phi^2} \right|_{\phi=0} = 0. \quad (3.33)$$

The coefficient of the terms quadratic in $(\phi + \phi_0)$ is given by

$$\frac{1}{2} \left[m_0^2 - \left. \frac{\partial^2 \mathcal{L}_{\text{int}}(\phi + \nu)}{\partial \phi^2} \right|_{\phi=0} \right] \equiv \frac{1}{2} m^2 \quad (3.34)$$

So that the effective mass of the interacting field for the broken symmetry is m . The reader will notice that these conditions are identical to the usual tree approximation used in functional methods.¹⁹

As an example, suppose that (3.20) has a nonzero ν solution. It then follows that the generating functional (3.22) is invariant under the following simultaneous operators.

$$J'' = e^{i\theta} J, \quad \phi'' = e^{i\theta} \phi, \quad \phi_0'' = e^{i\theta} \phi_0 + \nu(e^{i\theta} - 1), \quad (3.35)$$

the double prime being used to distinguish this transformation from (3.23). It then follows that

$$\psi'(x; \phi_0'', \phi_0''^*, \nu) = e^{i\theta} \psi(x; \phi_0, \phi_0^*, \nu), \quad (3.36)$$

so that the transformations of (3.35) generate the usual phase change of the interpolating field when a broken symmetry solution is selected. This result has been obtained before by other means.⁹ Of course, it is obvious that in order for the free field equation of motion to remain invariant under transformation (3.35) for nonzero ν , ϕ_0 must correspond to a massless particle, the usual Goldstone boson associated with the broken continuous symmetry.

As a final note, it is necessary to examine the boson theorem in the case of a broken symmetry. It happens that two forms of the Yang–Feldman equation (3.1) can be written for the case of broken symmetry, the two forms being equivalent.

The first form is derived from the generating functional

(considering a single Hermitian scalar field)

$$\begin{aligned} Z_r[J, \phi_0 + \nu] &= \exp \frac{i}{\hbar} \int dx \left\{ \mathcal{L}_{\text{int}} \left[\frac{\hbar}{i} \frac{\delta}{\delta J(x)} \right] - \frac{1}{2} \delta m^2 \hbar^2 \frac{\delta^2}{\delta J(x)^2} \right\} \\ &\times \int [d\phi] \exp i \int dy \left[\frac{1}{2} \phi (\partial^2 - m^2) \phi + J(\phi + \phi_0 + \nu) \right], \end{aligned} \quad (3.37)$$

where

$$\delta m^2 = m_0^2 - m^2. \quad (3.38)$$

Expression (3.37) gives rise to the Yang–Feldman equation

$$\begin{aligned} \psi(x; \phi_0) &= \phi_0(x) + \nu - \int dy \Delta(x-y) : Z_r^{-1}[J, \phi_0 + \nu] \\ &\times \mathcal{L}'_{\text{int}} \left[\frac{\hbar}{i} \frac{\delta}{\delta J(y)} \right] Z_r[J, \phi_0 + \nu] \Big|_{J=0} : \\ &+ \delta m^2 \int dy \Delta(x-y) \psi(y; \phi_0), \end{aligned} \quad (3.39)$$

where

$$(\partial_x^2 - m^2) \Delta(x-y) = \delta(x-y). \quad (3.40)$$

The second form comes from rewriting the generating functional as

$$\begin{aligned} Z_r[J, \phi_0 + \nu] &= \exp \frac{i}{\hbar} \int dy Q_{\text{int}} \left[\frac{\hbar}{i} \frac{\delta}{\delta k(y)} \right] \\ &\times \int [d\phi] \exp \frac{i}{\hbar} \int dx \left[\frac{1}{2} \phi (\partial^2 - m^2) \phi \right. \\ &\left. + J(\phi + \phi_0 + \nu) + k(\phi + \phi_0) \right] \Big|_{k=0}, \end{aligned} \quad (3.41)$$

where Q_{int} represents the cubic and higher order terms derived by expanding the shifted interaction appearing in (3.32) and cancelling the linear and quadratic terms. Expression (3.41) gives

$$\begin{aligned} \psi(x; \phi_0) &= \phi_0(x) + \nu + \int dy \Delta(x-y) : Z_r^{-1}[J, \phi_0 + \nu] \\ &\times Q'_{\text{int}} \left[\frac{\hbar}{i} \frac{\delta}{\delta J(y)} - \nu \right] Z_r[J, \phi_0 + \nu] \Big|_{J=0} :. \end{aligned} \quad (3.42)$$

The two forms for ψ , (3.39), and (3.42), must yield identical operators and therefore both must satisfy, in the limit given by (3.18), the equation

$$(\partial^2 - m_0^2) \phi^f(x) = - \mathcal{L}'_{\text{int}}[\phi^f(x)] \quad (3.43)$$

where ϕ_0 is shifted by the c -number function f .

IV. APPLICATIONS

In this section several simple applications of the results derived in Secs. II and III will be given.

A. The free field

As an illustration of these techniques, it is formally possible to expand a free field of mass m in terms of free field operators associated with mass m_0 . Of course, the field is then no longer free since it is defined over the incorrect Fock space. However, for the sake of example, such a procedure

can be realized by using an interaction term of the form

$$\mathcal{L}_{\text{int}} = \frac{1}{2}(m_0^2 - m^2)\phi^2. \quad (4.1)$$

The path integration may be performed exactly to obtain

$$\begin{aligned} W_r[J, \phi_0] &= \frac{i}{\hbar} \int dx dy \{ (m^2 - m_0^2)\phi_0(x)\Delta(x-y)J(y) \\ &\quad - \frac{1}{2}J(x)\Delta(x-y)J(y) \} + \frac{i}{\hbar} \int dx \phi_0(x)J(x), \end{aligned} \quad (4.2)$$

where terms independent of J have been dropped and

$$(\partial_x^2 - m^2)\Delta(x-y) = \delta(x-y). \quad (4.3)$$

It follows that

$$\psi(x; \phi_0) = \phi_0(x) + (m^2 - m_0^2) \int dy \Delta(x-y)\phi_0(y), \quad (4.4)$$

which clearly satisfies

$$(\partial_x^2 - m^2)\psi(x; \phi_0) = 0. \quad (4.5)$$

It is also clear that

$$R\{\psi(x)\psi(y)\}_{\text{connected}} = i\hbar\Delta(x-y), \quad (4.6)$$

all other connected Green's functions vanishing.

The boson theorem is illustrated in this case by shifting ϕ_0 by the function ce^{ikx} , where c is arbitrary, such that

$$k^2 = m_0^2. \quad (4.7)$$

Then, by (3.17),

$$\begin{aligned} \langle 0 | \psi^f(x) | 0 \rangle &= ce^{ikx} + c(m_0^2 - m^2) \int dy \Delta(x-y)e^{iky} = 0, \end{aligned} \quad (4.8)$$

which is a trivial solution to (4.5). However, if the static solution (in one spatial dimension)

$$f(x) = ce^{-m_0|x|} \quad (4.9)$$

is used, it follows that

$$\langle 0 | \psi^f(x) | 0 \rangle = \frac{m_0}{m} ce^{-m|x|}, \quad (4.10)$$

which is a static solution to (4.5).

B. Interacting case

The next case to be examined is the nontrivial example given in (3.4). In addition, a negative mass term will be se-

lected, allowing the nonzero ν solution to (3.33) of

$$\nu = \frac{m_0^2}{\lambda} \quad (4.11)$$

and the effective mass from (3.34) of m_0^2 . Then, in the $\hbar = 0$ limit, the field operator satisfies (3.42), which can be written

$$\psi_0(x; \phi_0) = \phi_0 + \nu + \lambda \int dy \Delta(x-y) [\psi_0(y; \phi_0) - \nu]^2. \quad (4.12)$$

Assuming the form for ψ_0

$$\psi_0(x; \phi_0) = \nu + \sum_{n=1}^{\infty} A_n : \rho^n(x) :, \quad (4.13)$$

where n refers to the power of the in-field operator, leads to the recurrence relation, for $n > 1$,

$$A_n : \rho^n(x) : = \lambda \int dy \Delta(x-y) \sum_{i+j=n} A_i A_j : \rho^{i+j}(x) :. \quad (4.14)$$

Shifting the field operator by the static solution

$$f(x) = ce^{-m_0 x} \quad (4.15)$$

to the free field equation and taking the vacuum expectation value of (4.13) gives the recurrence relation

$$A_n = \frac{\lambda}{(n^2 - 1)m_0^2} \sum_{i+j=n} A_i A_j, \quad (\forall_n > 1). \quad (4.16)$$

After selecting

$$c = -\frac{6m_0^2}{\lambda} e^{m_0 a} \equiv A_1, \quad (4.17)$$

where a is arbitrary, it follows that

$$\phi^f(x) = \frac{m_0^2}{\lambda} + \frac{6m_0^2}{\lambda} \sum_{n=1}^{\infty} (-1)^n n e^{-m_0 n(x-a)}, \quad (4.18)$$

which readily sums to

$$\phi^f(x) = \frac{m_0^2}{\lambda} - \frac{3}{2} \frac{m_0^2}{\lambda} \text{sech}^2 \frac{1}{2} m_0(x-a). \quad (4.19)$$

This is a static solution to the classical equation

$$(\partial^2 + m_0^2)\phi^f(x) = \lambda [\phi^f(x)]^2, \quad (4.20)$$

in agreement with result (3.43). If the positive sign is chosen for c in (4.17) it follows that

$$\phi^f(x) = \frac{m_0^2}{\lambda} + \frac{3}{2} \frac{m_0^2}{\lambda} \text{csch}^2 \frac{1}{2} m_0(x-a), \quad (4.21)$$

which is also a solution of (4.20), although irregular at $x = a$.

Higher order corrections to (4.20) may be obtained by performing the functional differentiation indicated in the Yang-Feldman equation (3.1). This repeated process will generate a double power series in λ and \hbar . For this purpose it is more convenient to use form (3.42) for ψ . After the first differentiation, it follows that

$$\begin{aligned} \psi(x; \phi_0) &= \nu + \phi_0(x) + \lambda \int dy \Delta(x-y) \\ &\quad \times : \mathcal{Z}_r^{-1}[J, \phi_0 + \nu] \left\{ \phi_0(y) + \lambda \int dz \Delta(y-z) \left[\frac{\hbar}{1} \frac{\delta}{\delta J(z)} - \nu \right]^2 \right\}^2 \mathcal{Z}_r[J, \phi_0 + \nu] \Big|_{J=0} : \\ &\quad + i\hbar 2\lambda^2 \int dy dz \Delta(x-y)\Delta^2(y-z) [\psi(x; \phi_0) - \nu] - i\hbar \frac{\lambda}{m_0^2} \Delta(0). \end{aligned} \quad (4.22)$$

Since the first three terms on the right-hand side of (4.22) do not vanish in the event that $\hbar = 0$, they must contain the tree

approximation. It is then assumed that the vacuum expectation value of ψ may be expanded in a double power series in λ and \hbar in the manner

$$\langle 0 | \psi^f(x) | 0 \rangle = \phi^f(x) + \lambda \hbar \phi_1^f(x) + O(\hbar^2) + \dots \quad (4.23)$$

Expression (4.22) may be renormalized consistently by making the demand that

$$\psi(x; \phi_0 = 0) = \nu, \quad (4.24)$$

and introducing counterterms to make (4.4) hold. By inserting (4.23) into the right-hand side of (4.22) and keeping terms in first order it follows that

$$\phi_1^f(x) = i3m_0^2 \int dy dz \Delta(x-y) \Delta^2(y-z) \text{sech}^2 \frac{1}{2} m_0(z-a). \quad (4.25)$$

Correction of higher order in λ and \hbar may be generated by continuing the expression (4.23), applying condition (4.24), and using the results derived at each of the previous stages of the expression.

The form for the connected two-point operator product is

$$R \{ \psi(x) \psi(y) \} = \left[Z_r^{-1} [J, \phi_0 + \nu] \frac{\hbar^2 \delta^2 Z_r [J, \phi_0 + \nu]}{i^2 \delta J(x) \delta J(y)} - Z_r^{-2} [J, \phi_0 + \nu] \frac{\hbar}{i} \frac{\delta Z_r [J, \phi_0 + \nu]}{\delta J(x)} \frac{\hbar}{i} \frac{\delta Z_r [J, \phi_0 + \nu]}{\delta J(y)} \right] \Big|_{J=0}. \quad (4.26)$$

For the interaction given by (3.4), expression (4.26) reduces to

$$R \{ \psi(x) \psi(y) \} = i\hbar \Delta(x-y) + i\hbar 2\lambda \int dz \Delta(x-z) \Delta(y-z) [\psi(z; \phi_0) - \nu] + \lambda^2 \int dr dz \Delta(x-r) \Delta(y-z) \{ W_{rrzz} + 2W_{rrz}(W_z - \nu) + 2W_{zrz}(W_r - \nu) + 2W_{rz}^2 + 4(W_r - \nu)W_{rz}(W_z - \nu) \}, \quad (4.27)$$

where the notation

$$W_z \equiv \frac{\hbar}{i} \frac{\delta W}{\delta J(z)} \Big|_{J=0}, \quad W_{rz} \equiv \frac{\hbar^2}{i^2} \frac{\delta^2 W}{\delta J(r) \delta J(z)} \Big|_{J=0}, \quad \text{etc.}, \quad (4.28)$$

has been introduced, and use has been made of

$$Z[J, \phi_0] = \exp W[J, \phi_0]. \quad (4.29)$$

It follows that (4.27) is at least of order \hbar , so that the vacuum expectation value has an expansion in terms of \hbar of the form

$$\langle 0 | R \{ \psi(x) \psi(y) \} | 0 \rangle = \hbar g_1(x, y) + \hbar^2 g_2(x, y) + \dots \quad (4.30)$$

Inserting (4.30) into (4.27) and dropping all terms of order \hbar^2 or higher leaves

$$g_1^f(x, y) \equiv \lim_{\hbar \rightarrow 0} [\hbar^{-1} \langle 0 | R^f \{ \psi(x) \psi(y) \} | 0 \rangle] = i\Delta(x-y) + i2\lambda \int dz \Delta(x-z) \Delta(y-z) [\phi^f(z) - \nu] + 4\lambda^2 \int dr dz \Delta(x-r) \Delta(y-z) [\phi^f(r) - \nu] [\phi^f(z) - \nu] g_1^f(r, z) \quad (4.31)$$

as the equation which $g_1(x, y)$ must satisfy when constructed on the soliton vacuum. Iteration of (4.31) shows that

$$g_1^f(x, y) = i \sum_{n=0}^{\infty} (2\lambda)^n \int dz_1 \dots dz_n \Delta(x-z_1) [\phi^f(z_1) - \nu] \Delta(z_1 - z_2) \dots [\phi^f(z_n) - \nu] \Delta(z_n - y). \quad (4.32)$$

Using the result (4.21) or (4.19) for $\phi^f(x)$ shows that the factor $(2\lambda)^n$ is cancelled order by order, leaving $g_1^f(x, y)$ independent of λ and expressed as a power series in m_0^2 . Thus, in the limit that λ goes to zero the free field is not obtained. The two-point function so constructed is then disjoint from the Fock representation normally employed to perform perturbation theory. Evaluation of (4.32) will be presented elsewhere.

V. CONCLUSIONS

In Secs. II and III the path integral form for the dynamical map was developed and its properties analyzed. An alternate version of the boson theorem was proved and symmetry behavior was discussed in the context of the results. Section IV illustrated this technique for both a trivial and an interacting case. In the latter case the method for generating quantum corrections to the classical limit was developed and employed. In addition, it was shown that arbitrary operator products could be examined using the path integral generator and the boson theorem.

Of course, the work presented here is contingent upon two major assumptions. The first is that the path integral has

a physically and mathematically meaningful interpretation. The second is that the boson transformation generates unitarily inequivalent representations of the canonical commutation relations for the *same* theory. Both of these assumptions are in dire need of clarification before the results obtained by this and other authors can be fully understood. The answers to these questions are as important as the extension of this work to multisoliton states, statistical mechanics for solitons and gauge theories.

ACKNOWLEDGMENTS

The author is indebted to Professor Y. Takahashi and Professor A. Z. Capri for valuable discussions regarding this work, and to Dr. H. Umezawa and Dr. H. Matsumoto for providing preprints of their work.

- ¹H. Matsumoto, N. J. Papastamatiou, and H. Umezawa, Nucl. Phys. **B 97**, 90 (1975); M. Wadati, H. Matsumoto, Y. Takahashi, and H. Umezawa, Phys. Lett. **A 62**, 255, 258 (1977); M. Wadati, H. Matsumoto, and H. Umezawa, Phys. Rev. **D 18**, 520, (1978); H. Matsumoto, P. Sodano, and H. Umezawa, University of Alberta preprint (1978), to be published (Phys. Rev. D).
²H. Umezawa, Nuovo Cimento **40**, 450 (1965); L. Leplae, R. N. Sen, and H. Umezawa, Prog. Theor. Phys. Suppl. **151** (1965).

- ³H. Umezawa, in *Renormalization and Invariance in Quantum Field Theory*, edited by E.R. Caianello (Plenum, New York, 1974), p. 275.
⁴L. Leplae, H. Umezawa, and F. Mancini, Phys. Rep. **C 10**, 151 (1974); H. Matsumoto, N. J. Papastamatiou, and H. Umezawa, Nucl. Phys. **B 82**, 45 (1974).
⁵J.R. Klauder, J. Math. Phys. **11**, 609 (1971); J. R. Klauder, Commun. Math. Phys. **18**, 307 (1970).
⁶B. DeFacio and C.L. Hammer, J. Math. Phys. **17**, 267 (1976); **18**, 1216 (1977).
⁷H. Matsumoto, P. Sodano, and H. Umezawa, in Ref. 1.
⁸R. P. Feynman, Rev. Mod. Phys. **20**, 367 (1947); Phys. Rev. **74**, 1430 (1948); E. S. Abers and B. W. Lee, Phys. Rep. **C 9**, 1 (1974).
⁹H. Matsumoto, N. J. Papastamatiou, and H. Umezawa, Nucl. Phys. **B 82**, 45 (1974).
¹⁰H. Lehmann, K. Symanzik, and W. Zimmermann, Nuovo Cimento **1**, 205 (1955).
¹¹see, for example, J. D. Bjorken and S. D. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965), p. 178f.
¹²The logarithm guarantees that only the connected part of (2.7) is kept.
¹³See. E. S. Abers and B. W. Lee, in Ref. 8.
¹⁴See, for example, G. Barton, *Introduction to Quantum Field Theory* (Wiley, New York, 1969), p. 348.
¹⁵These boundary conditions can be seen more explicitly in a coherent state formulation of path integrals; see in this respect C. Hammer, J. Schrauner, and B. DeFacio, Phys. Rev. **D 18**, 373 (1978); **D 19**, 667 (1979).
¹⁶E. S. Abers and B. W. Lee, in Ref. 8.
¹⁷C. N. Yang and D. Feldman, Phys. Rev. **79**, 772 (1950).
¹⁸J. Goldstone, Nuovo Cimento **19**, 154 (1961).
¹⁹S. Coleman and E. Weinberg, Phys. Rev. **D 7**, 1888 (1973).

Bipolar expansion of tensor fields: Recurrence relations and analytic construction of form factors

Henry H. K. Tang^{a)}

Niels Bohr Institute, DK 2100, Copenhagen, Denmark

Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Jan S. Vaagen

Department of Physics, University of Bergen, Bergen, Norway

(Received 31 July 1980; accepted for publication 17 October 1980)

Bipolar expansions of tensor fields under transformations of arbitrary translation and scaling of coordinates are obtained. The judicious choice of a coupled representation renders the expression compact and transparent. The radial coefficient functions (form factors) in this expansion can be given as integral transforms. They are shown to satisfy four-term recurrence relations, which together with an appropriate exploitation of the underlying symmetry result in a significant simplification of the original problem. A computational scheme is also worked out with which the radial form factors can be constructed analytically in terms of Gauss hypergeometric functions.

PACS numbers: 03.70. + k, 02.30.Mv, 02.30.Qy

I. INTRODUCTION

It is well known that physical quantities are classified according to their properties under coordinate transformations. Such symmetry considerations not only deepen our understanding of nature at a fundamental level, but may often lead to practical solutions to some apparently complex problems.

In both structure and reaction theories of atomic and nuclear many-body physics, one frequently encounters the overlaps of functions which are defined with respect to different reference frames. This often necessitates expansions of the relevant physical quantities about relatively translated systems. From a pragmatic standpoint, the efficiency with which such expansions can be performed can often help to circumvent what may otherwise be an intractable mathematical procedure.

In this paper, we discuss the bipolar expansion of tensor fields under general coordinate transformations of arbitrary translation and scaling.¹ Our discussions are mainly concerned with two aspects. We derive and analyze the radial coefficient functions (form factors) in the expansion. The salient features of these form factors can be succinctly displayed in four-term recurrence relations. With due exploitation of the underlying symmetries, we are able to construct analytically a subclass of form factors, from which all other form factors in the expansion can be generated recursively.

The more restricted cases of bipolar expansions of scalar functions have been studied by various authors.² Special algorithms have been developed to calculate the coefficient functions in such expansions. Though they are designed to cater for the specific problems in mind, these methods are often cumbersome. Such complexities may sometimes be due to the fact that the symmetry properties inherent in the problems have not been fully used.

A tensor field, of which a scalar field is but a special

case, is much richer in mathematical structure. But it is shown here that, by a judicious choice of a coupled representation which fully exploits the symmetry in the angular parts, the resultant bipolar expansion turns out to be simple, and, from a computational point of view, manageable.

Since a tensor field is usually defined by its properties under rotation,³ spherical tensor fields are natural candidates for our studies. In practice, they are realized in the form of single-particle wave functions and coupled single-particle wave functions relative to some core. In Sec. II, the bipolar expansions of tensor fields are derived by means of a Fourier transform method. The radial form factors in the expansions emerge as integral transforms, from which, four-term recurrence relations can be obtained. A simple diagrammatic representation is then introduced to help the enumeration of these form factors and illustrates, in a transparent manner, the simplification resulting from symmetry considerations. In Sec. III, these form factors are explicitly constructed. They can be expressed analytically in terms of simple polynomials and Gauss hypergeometric functions. Finally, Sec. IV summarizes the main findings and briefly discusses the prospects of practical applications of this method.

II. BIPOLAR EXPANSION AND RECURRENCE RELATIONS

A. Single-particle orbitals: Definition of form factors

$F_{\lambda\Lambda}^L(r, R)$

Consider the case of a spherical tensor field of rank L , of which a single-particle wavefunction with good angular momentum quantum numbers is a typical example,

$$\Psi_L^M(\mathbf{r}) = f_L(r) Y_L^M(\hat{r}). \quad (1)$$

Here, we adopt the usual convention that the argument \hat{r} of the spherical harmonic Y_L^M denotes the angular coordinates defined by the unit vector \mathbf{r}/r .

For most problems of physical interest, it is sufficient to consider square-integrable tensor fields, though the results obtained here are also valid under more general conditions.⁴

^{a)}This work is supported in part through funds provided by NSF grant PHY73-01164.

Since a general tensor field can be decomposed into components of type (1),³ attention is focused on spherical tensor fields.

It is well known that the Fourier transform of a spherical tensor is a tensor of the same rank in \mathbf{k} -space,

$$\begin{aligned} F[\Psi_L^M] &\equiv (2\pi)^{-3/2} \int d^3r e^{-i\mathbf{k}\cdot\mathbf{r}} \Psi_L^M(\mathbf{r}) \\ &= (-i)^L \tilde{f}_L(k) Y_L^M(\hat{\mathbf{k}}) \\ &= \tilde{\Psi}_L^M(\mathbf{k}). \end{aligned} \quad (2)$$

To arrive at Eq. (2), the usual partial wave expansion of the plane wave has been invoked. The radial function \tilde{f}_L is given by the Hankel transform

$$\tilde{f}_L(k) = (2/\pi)^{1/2} \int_0^\infty dr r^2 j_L(kr) f_L(r), \quad (3)$$

where j_L is the spherical Bessel function of order L .

The original tensor in \mathbf{r} -space can be recovered by means of the standard inverse transformation

$$\Psi_L^M(\mathbf{r}) = (2\pi)^{-3/2} \int d^3k e^{i\mathbf{k}\cdot\mathbf{r}} \tilde{\Psi}_L^M(\mathbf{k}), \quad (4)$$

which provides the starting point of the present analysis.

From Eq. (4), an expression can be obtained for $\Psi_L^M(\mathbf{r} + \mathbf{R})$ in terms of tensors in coordinates \mathbf{r} and \mathbf{R} ,

$$\Psi_L^M(\mathbf{r} + \mathbf{R}) = (2\pi)^{-3/2} \int d^3k e^{i\mathbf{k}\cdot(\mathbf{r} + \mathbf{R})} \tilde{\Psi}_L^M(\mathbf{k}). \quad (5)$$

After applying the plane-wave expansion for $\exp(i\mathbf{k}\cdot\mathbf{r})$ and $\exp(i\mathbf{k}\cdot\mathbf{R})$ on the right-hand side of Eq. (5), considerable reductions can be performed by means of standard tensor algebra and angular momentum recouplings. After the angular integration in the \mathbf{k} -space, the final expression is obtained,

$$\Psi_L^M(\mathbf{r} + \mathbf{R}) = \sum_{\lambda\Lambda} S_{\lambda\Lambda}^L F_{\lambda\Lambda}^L(r, R) [Y_\lambda(\hat{\mathbf{r}}) \otimes Y_\Lambda(\hat{\mathbf{R}})]_L^M. \quad (6)$$

Here, we have used conventional notations of angular momentum algebra.³ The indices λ and Λ in Eq. (6) satisfy conditions due to angular momentum addition,

$$|\lambda - \Lambda| \leq L \leq \lambda + \Lambda, \quad (7)$$

as well as a parity selection rule,

$$\lambda + \Lambda = \text{even}. \quad (8)$$

The origin of (8) can be traced back to the geometric factor

$$S_{\lambda\Lambda}^L = 2^{3/2} i^{\lambda + \Lambda + L} \hat{\lambda} \hat{\Lambda} \begin{pmatrix} \lambda & \Lambda & L \\ 0 & 0 & 0 \end{pmatrix}, \quad (9)$$

where

$$\hat{J} \equiv (2J + 1)^{1/2}. \quad (10)$$

The double radial functions $F_{\lambda\Lambda}^L$, hereafter called form factors, are given as integral transforms,

$$F_{\lambda\Lambda}^L(r, R) \equiv \int_0^\infty dk k^2 j_\lambda(kr) j_\Lambda(kR) \tilde{f}_L(k). \quad (11)$$

Steps leading to Eqs. (6)–(11) are outlined in Appendix A.

The judicious choice of a coupled representation in Eq. (6) provides a particularly economic expression for the bipolar expansion. The global symmetries of the tensor field

(which are general properties) are contained, term by term, in the angular parts and the geometric factors. The full burden of describing further details of the field (which are specific of the problem in question) is then carried by the radial form factors which have to be calculated for each case.

In principle, the expansion in Eq. (6) involves an infinite number of terms. It is clear that questions such as the rate of convergence of the series are not susceptible to simple analyses without specifying $f_L(r)$. We shall bypass such general considerations and proceed to study the properties of the form factor themselves.

An exploitation of the well-known recurrence relation of the spherical Bessel functions

$$j_{l+1}(z) + j_{l-1}(z) = [(2l+1)/z] j_l(z) \quad (12)$$

and repeated applications of definition (11) result in a four-term recurrence relation for the form factors (see Appendix B),

$$\begin{aligned} F_{\lambda, \Lambda+1}^L(r, R) + F_{\lambda+2, \Lambda+1}^L(r, R) \\ = \frac{2\lambda+3}{2\Lambda+3} \left(\frac{R}{r}\right) [F_{\lambda+1, \Lambda}^L(r, R) + F_{\lambda+1, \Lambda+2}^L(r, R)]. \end{aligned} \quad (13)$$

This recurrence relation leads to a significant simplification as can be illustrated in the following way. Each form factor in the bipolar expansion (6) can be associated with a lattice point on a (λ, Λ) -plane (see Fig. 1). Clearly, only the region $\lambda, \Lambda \geq 0$ is of interest. The inequalities (7) imposed by angular momentum addition imply that the form factors which constitute the bipolar expansion correspond to the lattice-points confined to a semi-infinite rectangle, bounded by and including the lines $\lambda + \Lambda = L$, $\lambda - \Lambda = L$ and $\Lambda - \lambda = L$. Furthermore, it is easy to see that, from the parity selection rule (8), only a subset of the points on the rectangle contributes to the expansion (see Fig. 1). Finally, the recurrence relation (13) provides a substantial reduction. The form factors associated with any two perpendicular lines, say, $\lambda + \Lambda = L$ and $\lambda - \Lambda = L$ (or $\Lambda - \lambda = L$), can be regarded as the basic elements of the expansion which require initial computations [see Figs. 1(a) and 1(b)]. All other form factors can be easily generated recursively from the basic set.

By way of simple illustration, consider a practical problem in which it is necessary to truncate the expansion series. A natural scheme would be to set

$$\lambda_{\max} = \Lambda_{\max} = L + N - 1 \quad (14)$$

for some N [see Fig. 1(b)]. Since the global symmetry of the tensor field is preserved, term by term, in the angular parts, the adequacy of the truncated series depends only on how rapidly the form factors vanish at higher λ and Λ . The advantage of the recurrence relations is that we can go to larger values of N (and hence λ_{\max} and Λ_{\max}) without any appreciable increase of computational efforts. As can be seen in Fig. 1(b), with given values of L and N , there are $(L+1)N$ form factors in the truncated series, of which, only $(L+N)$ functions need to be actually computed. This is a particularly substantial reduction for cases with large L values. Table I presents some of these reduction ratios $(L+N)/[(L+1)N]$. Even for $L=1$, this reduction ratio is 6/10 for $N=5$. For a

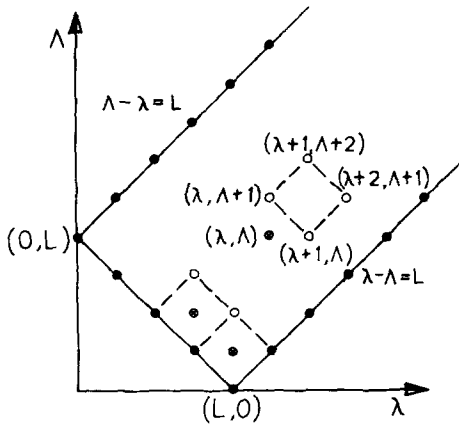


FIG. 1a. Diagrammatic representation of form factors $F_{\lambda\Lambda}^L(r, R)$ and their four-term recurrence relation. The dotted lines join the form factors which are related through the recursion relation. The crosses correspond to form factors which do not appear in the expansion due to the parity selection rule.

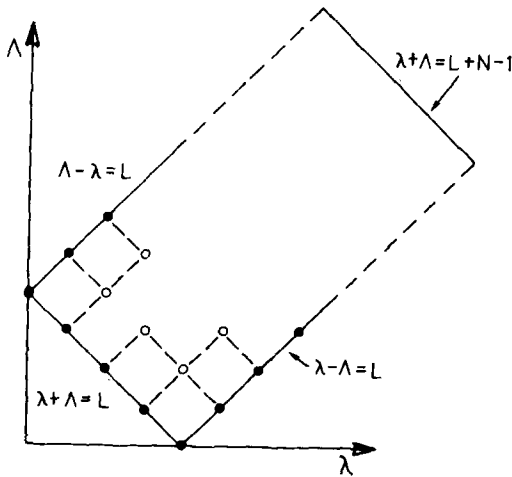


FIG. 1b. Building block form factors $T_{\lambda\Lambda}^L(r, R; r')$ and truncation scheme. Here, the basic form factors are represented by points denoted by solid circles along the lines $\lambda + \Lambda = L$, $\lambda - \Lambda = L$ and $\Lambda - \lambda = L$. The points denoted by open circles represent the form factors which can be obtained recursively from the basic form factors.

moderate value of $L = 5$, the reduction is $1/3$ and it considerably decreases further as we approach larger values of N .

B. Two coupled single-particle orbitals

The basic result in the previous subsection can be readily extended to a slightly more general case,

$$\Psi_{i_1, i_2}^M(\mathbf{r}_1, \mathbf{r}_2) = [f_{i_1}(\mathbf{r}_1) Y_{i_1}(\hat{\mathbf{r}}_1) \otimes f_{i_2}(\mathbf{r}_2) Y_{i_2}(\hat{\mathbf{r}}_2)]_L^M. \quad (15)$$

Such tensors frequently occur in many-body problems, for example, in the configuration of two single-particle states coupled to a total angular momentum L .

A repeated application of the basic formula (6) and the subsequent reduction by angular momentum algebra results in the following expansion:

$$\begin{aligned} \Psi_{i_1, i_2}^M(\mathbf{r}_1 + \mathbf{R}_1, \mathbf{r}_2 + \mathbf{R}_2) &= \sum_{\mathcal{L}'_1} \sum_{\lambda_1, \Lambda_1} \sum_{\mathcal{L}'_2} \sum_{\lambda_2, \Lambda_2} \hat{i}_1 \hat{i}_2 \hat{\mathcal{L}} \hat{\mathcal{L}}' \begin{Bmatrix} \lambda_1 & \lambda_2 & \mathcal{L} \\ \Lambda_1 & \Lambda_2 & \mathcal{L}' \\ l_1 & l_2 & L \end{Bmatrix} S_{\lambda_1, \Lambda_1}^{i_1} S_{\lambda_2, \Lambda_2}^{i_2} \\ &\times F_{\lambda_1, \Lambda_1}^{i_1}(\mathbf{r}_1, \mathbf{R}_1) F_{\lambda_2, \Lambda_2}^{i_2}(\mathbf{r}_2, \mathbf{R}_2) \\ &\times \{ [Y_{\lambda_1}(\hat{\mathbf{r}}_1) \otimes Y_{\lambda_2}(\hat{\mathbf{r}}_2)]_{\mathcal{L}} \otimes [Y_{\Lambda_1}(\hat{\mathbf{R}}_1) \otimes Y_{\Lambda_2}(\hat{\mathbf{R}}_2)]_{\mathcal{L}'} \}_{\mathcal{L}}^M. \end{aligned} \quad (16)$$

The conventions in Eq. (16) are the same as those in Eq. (6). The $9j$ -symbol arises from angular momentum recouplings and the form factors $F_{\lambda_1, \Lambda_1}^{i_1}$ and $F_{\lambda_2, \Lambda_2}^{i_2}$ are derived from f_{i_1} and f_{i_2} , respectively.

C. Bipolar expansions under general coordinate transformations

Consider the more general coordinate transformation for the single spherical tensor field discussed in subsection A,

$$\mathbf{r}_i \rightarrow \alpha_i \mathbf{R}_1 + \alpha_2 \mathbf{R}_2. \quad (17)$$

Equation (6) is then generalized to the following form:

$$\begin{aligned} \Psi_L^M(\alpha_1 \mathbf{R}_1 + \alpha_2 \mathbf{R}_2) &= \sum_{\lambda\Lambda} (-1)^{P_{\lambda}(\alpha_1) + P_{\Lambda}(\alpha_2)} S_{\lambda\Lambda}^L F_{\lambda\Lambda}^L(|\alpha_1| \mathbf{R}_1, |\alpha_2| \mathbf{R}_2) \\ &\times [Y_{\lambda}(\hat{\mathbf{R}}_1) \otimes Y_{\Lambda}(\hat{\mathbf{R}}_2)]_L^M. \end{aligned} \quad (18)$$

In arriving at Eq. (18), the parity property of spherical harmonics has been invoked,

$$Y_l^m(\alpha \hat{\mathbf{r}}) = \begin{cases} Y_l^m(\hat{\mathbf{r}}) & \text{if } \text{sgn} \alpha > 0 \\ (-1)^l Y_l^m(\hat{\mathbf{r}}) & \text{if } \text{sgn} \alpha < 0 \end{cases} \quad (19)$$

TABLE I. Ratio of the number of basic building block form factors to the total number of form factors in a given truncation.

N					
L	5	10	50	100	150
0	5/5	10/10	50/50	100/100	150/150
1	6/10	11/20	51/100	101/200	151/300
2	7/15	12/30	52/150	102/300	152/450
3	8/20	13/40	53/200	103/400	153/600
4	9/25	14/50	54/250	104/500	154/750
5	10/30	15/60	55/300	105/600	155/900
⋮	⋮	⋮	⋮	⋮	⋮
10	15/55	20/110	60/550	110/1100	160/1650

and phase factors $P_\lambda(\alpha)$ have been introduced,

$$P_\lambda(\alpha) \equiv \begin{cases} 0 & \text{if } \text{sgn}\alpha > 0 \\ \lambda & \text{if } \text{sgn}\alpha < 0 \end{cases} \quad (20)$$

Equipped with the results in Eqs. (16) and (18), we can now consider the expansion of a coupled tensor field $\Psi_{l_1 l_2 L}^M$ under the general (real) linear coordinate transformation

$$\begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{pmatrix} \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{pmatrix}, \quad (21)$$

where $\alpha_1, \alpha_2, \beta_1,$ and β_2 are real. The generalized result is

$$\begin{aligned} & \Psi_{l_1 l_2 L}^M(\alpha_1 \mathbf{R}_1 + \alpha_2 \mathbf{R}_2, \beta_1 \mathbf{R}_1 + \beta_2 \mathbf{R}_2) \\ &= \frac{1}{4\pi} \sum_{\mathcal{L}, \mathcal{L}'} \sum_{\lambda_1, \lambda_2} \sum_{\lambda_1, \lambda_2} \\ & \quad \times (-1)^{\mathcal{L} + \mathcal{L}' + P_{\lambda_1}(\alpha_1) + P_{\lambda_2}(\alpha_2) + P_{\lambda_1}(\beta_1) + P_{\lambda_2}(\beta_2)} \\ & \quad \times \hat{l}_1 \hat{l}_2 \hat{\lambda}_1 \hat{\lambda}_2 \hat{A}_1 \hat{A}_2 \hat{\mathcal{L}} \hat{\mathcal{L}}' \begin{pmatrix} \lambda_1 & \lambda_2 & \mathcal{L} \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 & \lambda_2 & \mathcal{L}' \\ 0 & 0 & 0 \end{pmatrix} \\ & \quad \times \begin{Bmatrix} \lambda_1 & \lambda_2 & \mathcal{L} \\ \lambda_1 & \lambda_2 & \mathcal{L}' \\ l_1 & l_2 & L \end{Bmatrix} S_{\lambda_1 \lambda_1}^{l_1} S_{\lambda_2 \lambda_2}^{l_2} F_{\lambda_1 \lambda_1}^{l_1} (|\alpha_1| R_1, |\alpha_2| R_2) \\ & \quad \times F_{\lambda_2 \lambda_2}^{l_2} (|\beta_1| R_1, |\beta_2| R_2) [Y_{\mathcal{L}}(\hat{R}_1) \otimes Y_{\mathcal{L}'}(\hat{R}_2)]_L^M. \end{aligned} \quad (22)$$

The recurrence relation (13) is now generalized to the following form:

$$\begin{aligned} & F_{\lambda, \lambda+1}^L(ar, bR) + F_{\lambda+2, \lambda+1}^L(ar, bR) \\ &= \frac{2\lambda+3}{2\lambda+3} \cdot \frac{bR}{ar} [F_{\lambda+1, \lambda}^L(ar, bR) + F_{\lambda+1, \lambda+2}^L(ar, bR)] \end{aligned} \quad (23)$$

for scaling factors $a, b > 0$.

III. ANALYTIC CONSTRUCTION OF FORM FACTORS

A. Definition of form factors $T_{\lambda\lambda}^L(r, R; r')$

The previous section concerns the recurrence relations and the global symmetries of the double radial form factors in the bipolar expansion. In a given problem, the practicality of such an expansion depends, to a considerable extent, on the ease with which these radial functions can be calculated. We now show that they can in fact be constructed as simple transforms involving only known functions.

Equations (3) and (11) can be combined to give the following:

$$F_{\lambda\lambda}^L(r, R) = \left(\frac{2}{\pi}\right)^{1/2} \int_0^\infty dr' r' f_L(r') T_{\lambda\lambda}^L(r, R; r'), \quad (24)$$

where a family of auxiliary form factors $\{T_{\lambda\lambda}^L\}$ has been introduced,

$$T_{\lambda\lambda}^L(r, R; r') \equiv \int_0^\infty dk k^2 j_\lambda(kr) j_\lambda(kR) j_L(kr'). \quad (25)$$

Obviously, these form factors also satisfy the recurrence relation (23).

It is easy to see the convergence of the integral in Eq. (25). From the analytic properties of spherical Bessel functions, it is obvious that the integral is well-defined over an interval $[0, k_{\max}]$ for any large but finite k_{\max} . On the other hand, the spherical Bessel functions also have well-known asymptotic behavior,

$$\begin{aligned} & k^2 j_\lambda(kr) j_\lambda(kR) j_L(kr') \\ & \rightarrow_{k \rightarrow \infty} k^2 \frac{\sin(kr - \frac{1}{2}\lambda\pi)}{kr} \frac{\sin(kR - \frac{1}{2}\lambda\pi)}{kR} \frac{\sin(kr' - \frac{1}{2}L\pi)}{kr'}. \end{aligned} \quad (26)$$

Hence, the oscillating integrand in Eq. (25) is absolutely bounded by $(kRr')^{-1}$ over $[k_{\max}, \infty]$ and the integral can be easily shown to converge over this semi-infinite interval.

A more rigorous proof for the convergence of integral (25) is as follows: It is sufficient to demonstrate the convergence of

$$\int_{k_{\max}}^\infty dk k^2 j_\lambda(kr) j_\lambda(kR) j_L(kr')$$

for any nonzero k_{\max} . Since we have the decomposition formula (30), it is clear that it is sufficient to demonstrate the convergence of integrals of the type:

$$\int_{x_0}^\infty dx \begin{pmatrix} \sin(x+\alpha) \\ \cos(x+\alpha) \end{pmatrix} x^{-\beta} \quad \begin{pmatrix} \beta > 0 \\ x_0 > 0 \end{pmatrix}.$$

Formally, this is

$$\begin{aligned} & \int_{x_0}^\infty dx \frac{d}{dx} \left[\begin{pmatrix} -\cos(x+\alpha) \\ \sin(x+\alpha) \end{pmatrix} \right] x^{-\beta} \\ &= \left[\begin{pmatrix} -\cos(x+\alpha) \\ \sin(x+\alpha) \end{pmatrix} x^{-\beta} \right]_{x_0}^\infty \\ & \quad + \beta \int_{x_0}^\infty dx \begin{pmatrix} -\cos(x+\alpha) \\ \sin(x+\alpha) \end{pmatrix} x^{-(\beta+1)}. \end{aligned}$$

The above manipulation is formal. But the first term is finite and the second integral is absolutely bounded by $x^{-(1+\beta)}$ and converges. Hence the original integral converges.

There are obvious advantages for Eqs. (24) and (25). In Eq. (11), $F_{\lambda\lambda}^L$ are given as transforms of $\tilde{f}_L(k)$. But in Eq. (24), they are directly expressed as transforms of the original radial function $f_L(r)$. From the standpoint of numerical implementations, the use of Eq. (24) will bypass any loss of accuracy that results from calculating $\tilde{f}_L(k)$. Moreover, the form factors $T_{\lambda\lambda}^L$ are independent of the tensor field in question. They can be computed once for all and can be saved for later steps of the computations. This is a nontrivial saving of effort, particularly in large scale calculations. Hence, such considerations warrant an emphasis on $T_{\lambda\lambda}^L$ as the basic "building blocks" for the bipolar expansion.

It may seem that there is still a drawback for $T_{\lambda\lambda}^L$. As functions of three variables, their structure may be complicated. From a numerical point of view, their computation may require a three-dimensional storage. However, such shortcoming is only superficial. It can be readily verified from definition (25) that these form factors have the scaling property,

$$T_{\lambda\lambda}^L(\alpha r, \beta R; \gamma r') = (\alpha\beta\gamma)^{-3} T_{\lambda\lambda}^L\left(\frac{r}{\beta\gamma}, \frac{R}{\gamma\alpha}; \frac{r'}{\alpha\beta}\right) \quad (27)$$

for $\alpha, \beta, \gamma > 0$. Hence, the only relevant elements which determine the structure of $T_{\lambda\lambda}^L$ are the ratios between any two pairs of radii. In an actual numerical calculation, we need only to calculate the values of the function $T_{\lambda\lambda}^L$ at a selected set of points which require a two-dimensional storage; the rest can be generated by appropriate scalings. This feature

will be further illustrated later from the explicit construction of these form factors.

We now develop a computational scheme by which the form factors $T_{\lambda A}^L$ can be constructed analytically. This can be accomplished by exploiting some well-known properties of Bessel functions.

The functions $T_{\lambda A}^L$ can be rewritten in terms of Bessel functions of the first kind,

$$T_{\lambda A}^L(r, R; r') = \left(\frac{\pi}{2}\right)^{3/2} \frac{1}{(Rr r')^{1/2}} \int_0^\infty dk k^{1/2} J_{\lambda+1/2}(kr) \times J_{\lambda+1/2}(kR) J_{L+1/2}(kr'). \quad (28)$$

Equation (28) is directly obtained from Eq. (25) after applying the relation

$$j_l(z) = \left(\frac{\pi}{2z}\right)^{1/2} J_{l+1/2}(z). \quad (29)$$

Furthermore, for integral order l , $J_{l+1/2}$ has a simple representation,⁵

$$J_{l+1/2}(z) = \left(\frac{2}{\pi z}\right)^{1/2} \{ \sin(z - \frac{1}{2}l\pi) \Pi_1(l; z) + \cos(z - \frac{1}{2}l\pi) \Pi_2(l; z) \} \quad (30)$$

for $z \neq 0$, where Π_1 and Π_2 are finite polynomials in inverse powers of z . That is,

$$\Pi_1(l; z) \equiv \sum_{n=0}^{[l/2]} \frac{(-1)^n (l+2n)!}{(2n)!(l-2n)!} \frac{1}{(2z)^{2n}} \quad (31)$$

and

$$\Pi_2(l; z) \equiv \sum_{n=0}^{[(l-1)/2]} \frac{(-1)^n (l+2n+1)!}{(2n+1)!(l-2n-1)!} \frac{1}{(2z)^{2n+1}}, \quad (32)$$

where we have adopted the notation

$$[\kappa] \equiv \begin{cases} 0 & \text{if } \kappa < 0 \\ \text{integral part of } \kappa & \text{if } \kappa \geq 0 \end{cases} \quad (33)$$

$$\begin{aligned} J_{\lambda+1/2}(kr) J_{\lambda+1/2}(kR) &= \frac{1}{\pi k} \frac{1}{(rR)^{1/2}} \left\{ \sin(k(r+R)) \left[\cos\left(\frac{(\lambda+A)\pi}{2}\right) (\Pi_1(\lambda; kr) \Pi_2(\lambda; kR) + \Pi_2(\lambda; kr) \Pi_1(\lambda; kR)) \right. \right. \\ &+ \left. \sin\left(\frac{(\lambda+A)\pi}{2}\right) (-\Pi_1(\lambda; kr) \Pi_1(\lambda; kR) + \Pi_2(\lambda; kr) \Pi_2(\lambda; kR)) \right] \\ &+ \cos(k(r+R)) \left[-\sin\left(\frac{(\lambda+A)\pi}{2}\right) (\Pi_1(\lambda; kr) \Pi_2(\lambda; kR) + \Pi_2(\lambda; kr) \Pi_1(\lambda; kR)) \right. \\ &+ \left. \cos\left(\frac{(\lambda+A)\pi}{2}\right) (-\Pi_1(\lambda; kr) \Pi_1(\lambda; kR) + \Pi_2(\lambda; kr) \Pi_2(\lambda; kR)) \right] \\ &+ \sin(k(r-R)) \left[\cos\left(\frac{(\lambda-A)\pi}{2}\right) (\Pi_1(\lambda; kr) \Pi_2(\lambda; kR) - \Pi_2(\lambda; kr) \Pi_1(\lambda; kR)) \right. \\ &+ \left. \sin\left(\frac{(\lambda-A)\pi}{2}\right) (\Pi_1(\lambda; kr) \Pi_1(\lambda; kR) + \Pi_2(\lambda; kr) \Pi_2(\lambda; kR)) \right] \\ &+ \cos(k(r-R)) \left[-\sin\left(\frac{(\lambda-A)\pi}{2}\right) (\Pi_1(\lambda; kr) \Pi_2(\lambda; kR) - \Pi_2(\lambda; kr) \Pi_1(\lambda; kR)) \right. \\ &+ \left. \left. \cos\left(\frac{(\lambda-A)\pi}{2}\right) (\Pi_1(\lambda; kr) \Pi_1(\lambda; kR) + \Pi_2(\lambda; kr) \Pi_2(\lambda; kR)) \right] \right\}. \quad (37) \end{aligned}$$

The expression in Eq. (37) encompasses two cases, namely, $\lambda \pm A = \text{even}$ and $\lambda \pm A = \text{odd}$. In each case, half of the terms in Eq. (37) vanish.

$$\Pi_1(\lambda; kr) \Pi_1(\lambda; kR) = \sum_{N=0}^{[\lambda/2] + [A/2]} \frac{(-1)^N}{2^{2N}} C_N^{(1,1)}(\lambda, r, A, R) k^{-2N}, \quad (38)$$

If the representation (30) is substituted into Eq. (28) for, say, $J_{\lambda+1/2}$ and $J_{\lambda+1/2}$, the form factor $T_{\lambda A}^L$ can be formally expressed as a sum of terms with the form

$$S_{\alpha\beta}^{[s]} = \int_0^\infty dk \left\{ \frac{\sin[k(r \pm R)]}{\cos[k(r \pm R)]} \right\} k^{-1/2} \Pi_\alpha(\lambda; kr) \times \Pi_\beta(\lambda; kR) J_{L+1/2}(kr'), \quad (34)$$

where $\alpha, \beta = 1, 2$. After some straightforward manipulations, to be discussed below, the product of polynomials can be expressed in the form

$$k^{-1/2} \Pi_\alpha \Pi_\beta = \sum_n C_n(r, R) k^{-2n-m}, \quad (35)$$

where C_n 's are k -independent algebraic functions of r and R , and where m 's are specified below.

Provided that the integrand is well behaved at $k=0$ and vanishes at $k \rightarrow \infty$, $S_{\alpha\beta}$ can be decomposed as

$$S_{\alpha\beta}^{[s]} = \sum_n C_n(r, R) \int_0^\infty dk \left\{ \frac{\sin[k(r \pm R)]}{\cos[k(r \pm R)]} \right\} \times k^{-2n-m} J_{L+1/2}(kr'). \quad (36)$$

If these integrals are convergent, they can be expressed in terms of Gauss hypergeometric functions.

Before making digression into the mathematical details of such decomposition, it should be emphasized that in the present bipolar expansion, it is always possible to select an optimal set of form factors which can be constructed from terms like (36), for which the finite sum is term-by-term convergent. The rest of the form factors can then be generated exactly from the four-term recurrence relation.

B. Product of Bessel functions

From Eq. (30), it is straightforward to obtain the following expression for a product of two Bessel functions of half-integral orders,

Considered as polynomials of k , the terms $\Pi_\alpha \Pi_\beta$ ($\alpha, \beta = 1, 2$) fall into three types.

(1) The first type is

where we have introduced coefficient functions $C_N^{(1,1)}$ which are finite polynomials in inverse powers of r and R only. They can be given in different representations

$$C_N^{(1,1)}(\lambda, r, A, R) \equiv \sum_{j=\max(0, N - \lfloor A/2 \rfloor)}^{\min(N, \lfloor \lambda/2 \rfloor)} \binom{\lambda}{2j} \binom{A}{2(N-j)} (\lambda+1)_{2j} (A+1)_{2(N-j)} \frac{1}{r^{2j} R^{2(N-j)}} \\ \equiv \sum_{j=\max(0, N - \lfloor \lambda/2 \rfloor)}^{\min(N, \lfloor A/2 \rfloor)} \binom{A}{2j} \binom{\lambda}{2(N-j)} (A+1)_{2j} (\lambda+1)_{2(N-j)} \frac{1}{R^{2j} r^{2(N-j)}}, \quad (39)$$

where we have used the standard notation

$$\binom{n}{m} = \frac{n!}{m!(n-m)!} \quad (40a)$$

for the binomial and furthermore,

$$(x)_n = x(x+1)\cdots(x+n-1). \quad (40b)$$

Equation (39) can be given in a more compact form after the introduction of the following notations:

$$r_{>} \equiv \max(r, R), \quad (41a)$$

$$r_{<} \equiv \min(r, R), \quad (41b)$$

$$\lambda_{>} \equiv \begin{cases} \lambda & \text{if } r > R \\ A & \text{if } r < R \end{cases}, \quad (41c)$$

and

$$\lambda_{<} \equiv \begin{cases} \lambda & \text{if } r < R \\ A & \text{if } r > R \end{cases}. \quad (41d)$$

Equation (39) can then be expressed as

$$C_N^{(1,1)}(\lambda, r, A, R) = \frac{1}{r_{>}^{2N}} \sum_{j=\max(0, N - \lfloor \lambda_{<}/2 \rfloor)}^{\min(N, \lfloor \lambda_{>}/2 \rfloor)} \binom{\lambda_{>}}{2j} \binom{\lambda_{<}}{2(N-j)} (\lambda_{>}+1)_{2j} (\lambda_{<}+1)_{2(N-j)} \left(\frac{r_{<}}{r_{>}}\right)^{2j}, \quad (42)$$

where the summation is a polynomial in the dimensionless ratio $r_{<}/r_{>} (\leq 1)$.

(2) By a similar analysis the second type products are

$$II_2(\lambda; kr) II_2(A; kR) = \sum_{N=0}^{[(\lambda-1)/2] + [(A-1)/2]} \frac{(-1)^N}{2^{2(N+1)}} C_N^{(2,2)}(\lambda, r, A, R) k^{-2(N+1)} \quad (43)$$

for which the coefficient functions $C_N^{(2,2)}$ are defined by

$$C_N^{(2,2)}(\lambda, r, A, R) \equiv \frac{1}{r_{>}^{2N+1}} \sum_{j=\max(0, N - \lfloor (\lambda_{<} - 1)/2 \rfloor)}^{\min(N, \lfloor (\lambda_{>} - 1)/2 \rfloor)} \binom{\lambda_{>}}{2j+1} \binom{\lambda_{<}}{2(N-j)+1} (\lambda_{>}+1)_{2j+1} (\lambda_{<}+1)_{2(N-j)+1} \left(\frac{r_{<}}{r_{>}}\right)^{2j}. \quad (44)$$

(3) The third type includes:

$$II_1(\lambda; kr) II_2(A; kR) = \sum_{N=0}^{[\lambda/2] + [(A-1)/2]} \frac{(-1)^N}{2^{2N+1}} C_N^{(1,2)}(\lambda, r, A, R) k^{-(2N+1)}, \quad (45)$$

where

$$C_N^{(1,2)}(\lambda, r, A, R) \equiv \begin{cases} \frac{1}{r_{<}^{2N+1}} \sum_{j=\max(0, N - \lfloor \lambda_{<} - 1 \rfloor / 2)}^{\min(N, \lfloor \lambda_{>} / 2 \rfloor)} \binom{\lambda_{>}}{2j} \binom{\lambda_{<}}{2(N-j)+1} (\lambda_{>}+1)_{2j} (\lambda_{<}+1)_{2(N-j)+1} \left(\frac{r_{<}}{r_{>}}\right)^{2j} & \text{(if } r > R) \\ \frac{1}{r_{>}^{2N}} \sum_{j=\max(0, N - \lfloor \lambda_{<} / 2 \rfloor)}^{\min(N, \lfloor (\lambda_{>} - 1)/2 \rfloor)} \binom{\lambda_{>}}{2j+1} \binom{\lambda_{<}}{2(N-j)} (\lambda_{>}+1)_{2j+1} (\lambda_{<}+1)_{2(N-j)} \left(\frac{r_{<}}{r_{>}}\right)^{2j} & \text{(if } r < R) \end{cases} \quad (46)$$

Similarly,

$$II_2(\lambda; kr) II_1(A; kR) = \sum_{N=0}^{[(\lambda-1)/2] + [A/2]} \frac{(-1)^N}{2^{2N+1}} C_N^{(2,1)}(\lambda, r, A, R) k^{-(2N+1)}, \quad (47)$$

where

$$C_N^{(2,1)}(\lambda, r, A, R) \equiv \begin{cases} \frac{1}{r_{>} r_{<}^{2N}} \sum_{j=\max(0, N - \lfloor \lambda_{<} / 2 \rfloor)}^{\min(N, \lfloor (\lambda_{>} - 1)/2 \rfloor)} \binom{\lambda_{>}}{2j+1} \binom{\lambda_{<}}{2(N-j)} (\lambda_{>}+1)_{2j+1} (\lambda_{<}+1)_{2(N-j)} \left(\frac{r_{<}}{r_{>}}\right)^{2j} & \text{(if } r > R) \\ \frac{1}{r_{<}^{2N+1}} \sum_{j=\max(0, N - \lfloor \lambda_{<} - 1 \rfloor / 2)}^{\min(N, \lfloor \lambda_{>} / 2 \rfloor)} \binom{\lambda_{>}}{2j} \binom{\lambda_{<}}{2(N-j)+1} (\lambda_{>}+1)_{2j} (\lambda_{<}+1)_{2(N-j)+1} \left(\frac{r_{<}}{r_{>}}\right)^{2j} & \text{(if } r < R) \end{cases} \quad (48)$$

C. Building block functions

The decomposition of the form factors suggested by Eqs. (30) and (36) implies that it is necessary to construct a class of functions which may be regarded as the ultimate building blocks of this method. From Eq. (36), it is clear that these functions can be defined categorically by the transform:

$$I_{\mu}^{\nu}(y, x) \equiv \int_0^{\infty} dt t^{\mu} J_{\nu}(xt) e^{iyt}, \quad (49)$$

where $x, y > 0$ and, in the present problem, ν is a half integral order.

Equation (49) is only a formal definition. It is necessary

to find the criteria under which the integral is well defined. By an argument similar to that for the integral (25), it is clear that the sufficient conditions for the convergence of Eq. (49) are that the integrand is finite at $t = 0$ and vanishes at $t \rightarrow \infty$.

From the known properties of the Bessel functions, the integrand is finite at $t = 0$ if and only if

$$\mu + \nu \geq 0. \quad (50)$$

It vanishes at $t \rightarrow \infty$ if and only if

$$\mu - \frac{1}{2} < 0. \quad (51)$$

Under the conditions (50) and (51), the integrand in Eq. (49) can be evaluated analytically.⁵ For $0 < y < x$,

$$I_{\mu}^{\nu}(y, x) = 2^{\mu} x^{-(1+\mu)} \left[\frac{\Gamma((1+\mu+\nu)/2)}{\Gamma((\nu-\mu+1)/2)} F\left(\frac{1+\mu+\nu}{2}, \frac{1+\mu-\nu}{2}; \frac{1}{2}; \left(\frac{y}{x}\right)^2\right) + i \left(\frac{2y}{x}\right) \frac{\Gamma((2+\mu+\nu)/2)}{\Gamma((\nu-\mu)/2)} F\left(\frac{2+\mu+\nu}{2}, \frac{2+\mu-\nu}{2}; \frac{3}{2}; \left(\frac{y}{x}\right)^2\right) \right], \quad (52a)$$

and for $0 < x < y$,

$$I_{\mu}^{\nu}(y, x) = \left(\frac{x}{2}\right)^{\nu} y^{-(\mu+\nu+1)} \frac{\Gamma(1+\mu+\nu)}{\Gamma(\nu+1)} \times \left\{ \cos\left[\frac{\pi}{2}(1+\mu+\nu)\right] F\left(\frac{1+\mu+\nu}{2}, \frac{2+\mu+\nu}{2}; \nu+1; \left(\frac{x}{y}\right)^2\right) + i \sin\left[\frac{\pi}{2}(1+\mu+\nu)\right] F\left(\frac{2+\mu+\nu}{2}, \frac{1+\mu+\nu}{2}; \nu+1; \left(\frac{x}{y}\right)^2\right) \right\}, \quad (52b)$$

where $\Gamma(z)$ are gamma functions and $F(a, b; c; z)$ are Gauss hypergeometric functions.

D. Analytic expressions of form factors

With the expressions for $\Pi_{\alpha} \Pi_{\beta}$ available and the functions I_{μ}^{ν} evaluated, the form factors $T_{\lambda A}^L$ can be calculated explicitly.

For the sake of definiteness, consider a polynomial expansion for $J_{\lambda+1/2}$ and $J_{A+1/2}$ in Eq. (28). In order that a component like (36) be well defined, it is necessary that the polynomial expansion is term-by-term convergent. This implies that the integrands in Eq. (36) must be well behaved at $k \rightarrow 0$. From Eqs. (37), (38), (43), (45), and (47), the most singular terms are $O(k^{-2([\lambda/2] + [A/2])})$ from $\Pi_{\alpha} \Pi_{\beta}$. Then, in the integrals,

$$k^{-2([\lambda/2] + [A/2] + 1)} J_{L+1/2}(kr') \xrightarrow{k \rightarrow 0} k^{-2([\lambda/2] + [A/2] + 1)} \frac{(kr')^{L+1}}{2^{L+1} \Gamma(L+3/2)}. \quad (53)$$

Hence, all integrands are well-behaved at $k \rightarrow 0$ if and only if

$$L - 2([\lambda/2] + [A/2]) \geq 0. \quad (54)$$

Moreover, in the asymptotic region, even the least convergent term goes as $k^{-1/2}$ as $k \rightarrow \infty$. From the (λ, A) -diagram in Fig. 1a, it is clear that the form factors which satisfy condition (54) correspond to points lying on the $\lambda + A = L$ line.

By similar arguments, if we choose $J_{\lambda+1/2} J_{L+1/2}$ for the polynomial expansion, then, the convergence condition implies that we have to compute the form factors along the $\lambda - A = L$ line. If the polynomial expansion is performed for $J_{L+1/2} J_{\lambda+1/2}$ then we must calculate form factors along the $A - \lambda = L$ line.

A previous argument shows that if the form factors as-

sociated with $\lambda + A = L$ and $\lambda - A = L$ (or $A - \lambda = L$) are known, all others can be generated recursively. This completes the demonstration that our decomposition scheme is valid for the bipolar expansion.

We can now work out the detailed expressions for $T_{\lambda A}^L$. For simplicity, we consider the form factors associated with the $\lambda + A = L$ line. With appropriate exchanges of indices and arguments, the form factors associated with the lines $\lambda - A = L$ and $A - \lambda = L$ can be obtained from the same formulas.

For the sake of clarity, we consider the cases of $\lambda \pm A = \text{even}$ and $\lambda \pm A = \text{odd}$ separately.

(1) If $\lambda + A = L = \text{even}$,

$$\begin{aligned}
T_{\lambda\lambda}^L(r, R; r') &= \frac{1}{2} \left(\frac{\pi}{2r'} \right)^{\frac{1}{2}} \frac{1}{Rr} (-)^{(\lambda+A)/2} \left\{ \sum_{N=0}^{\lfloor \frac{\lambda}{2} \rfloor + \lfloor (A-1)/2 \rfloor} \frac{(-)^N}{2^{2N+1}} C_N^{(1,2)}(\lambda, r; A, R) \right. \\
&\quad \times \mathcal{I} \mathfrak{m} [I_{-(2N+3/2)}^{L+\frac{1}{2}}(r+R, r') + (-)^{\lambda} \text{sgn}(r-R) I_{-(2N+3/2)}^{L+\frac{1}{2}}(|r-R|, r')] \\
&\quad + \sum_{N=0}^{\lfloor (\lambda-1)/2 \rfloor + \lfloor A/2 \rfloor} \frac{(-)^N}{2^{2N+1}} C_N^{(2,1)}(\lambda, r; A, R) \\
&\quad \times \mathcal{I} \mathfrak{m} [I_{-(2N+3/2)}^{L+\frac{1}{2}}(r+R, r') - (-)^{\lambda} \text{sgn}(r-R) I_{-(2N+3/2)}^{L+\frac{1}{2}}(|r-R|, r')] \\
&\quad + \sum_{N=0}^{\lfloor \lambda/2 \rfloor + \lfloor A/2 \rfloor} \frac{(-)^N}{2^{2N}} C_N^{(1,1)}(\lambda, r; A, R) \\
&\quad \times \mathcal{R} \mathfrak{e} [-I_{-(2N+\frac{1}{2})}^{L+\frac{1}{2}}(r+R, r') + (-)^{\lambda} I_{-(2N+\frac{1}{2})}^{L+\frac{1}{2}}(|r-R|, r')] + \sum_{N=0}^{\lfloor (\lambda-1)/2 \rfloor + \lfloor (A-1)/2 \rfloor} \frac{(-)^N}{2^{2(N+1)}} C_N^{(2,2)}(\lambda, r; A, R) \\
&\quad \times \mathcal{R} \mathfrak{e} [I_{-(2N+5/2)}^{L+\frac{1}{2}}(r+R, r') + (-)^{\lambda} I_{-(2N+5/2)}^{L+\frac{1}{2}}(|r-R|, r')] \Big\}. \tag{55a}
\end{aligned}$$

(2) If $\lambda + A = L = \text{odd}$,

$$\begin{aligned}
T_{\lambda\lambda}^L(r, R; r') &= \frac{1}{2} \left(\frac{\pi}{2r'} \right)^{1/2} \frac{1}{Rr} (-1)^{(\lambda+A-1)/2} \left\{ - \sum_{N=0}^{\lfloor \frac{\lambda}{2} \rfloor + \lfloor (A-1)/2 \rfloor} \frac{(-)^N}{2^{2N+1}} C_N^{(1,2)}(\lambda, r; A, R) \right. \\
&\quad \times \mathcal{R} \mathfrak{e} [I_{-(2N+3/2)}^{L+\frac{1}{2}}(r+R, r') + (-)^{\lambda} I_{-(2N+3/2)}^{L+\frac{1}{2}}(|r-R|, r')] - \sum_{N=0}^{\lfloor (\lambda-1)/2 \rfloor + \lfloor A/2 \rfloor} \frac{(-)^N}{2^{2N+1}} C_N^{(2,1)}(\lambda, r; A, R) \\
&\quad \times \mathcal{R} \mathfrak{e} [I_{-(2N+3/2)}^{L+\frac{1}{2}}(r+R, r') - (-)^{\lambda} I_{-(2N+3/2)}^{L+\frac{1}{2}}(|r-R|, r')] + \sum_{N=0}^{\lfloor \lambda/2 \rfloor + \lfloor A/2 \rfloor} \frac{(-)^N}{2^{2N}} C_N^{(1,1)}(\lambda, r; A, R) \\
&\quad \times \mathcal{I} \mathfrak{m} [-I_{-(2N+\frac{1}{2})}^{L+\frac{1}{2}}(r+R, r') + (-)^{\lambda} \text{sgn}(r-R) I_{-(2N+\frac{1}{2})}^{L+\frac{1}{2}}(|r-R|, r')] \\
&\quad + \sum_{N=0}^{\lfloor (\lambda-1)/2 \rfloor + \lfloor (A-1)/2 \rfloor} \frac{(-)^N}{2^{2(N+1)}} C_N^{(2,2)}(\lambda, r; A, R) \\
&\quad \times \mathcal{I} \mathfrak{m} [I_{-(2N+5/2)}^{L+\frac{1}{2}}(r+R, r') + (-)^{\lambda} \text{sgn}(r-R) I_{-(2N+5/2)}^{L+\frac{1}{2}}(|r-R|, r')] \Big\}. \tag{55b}
\end{aligned}$$

Hence, in Eqs. (55a) and (55b), the basic form factors $T_{\lambda\lambda}^L$ are expressed as finite sums of Gauss hypergeometric functions I_{μ}^{ν} which have been given in Eqs. (52a) and (52b). The coefficients $C_N^{(1,1)}$, $C_N^{(2,2)}$, $C_N^{(1,2)}$ and $C_N^{(2,1)}$ have also been evaluated explicitly in Eqs. (42), (44), (46), and (48).

In passing, it is also noted that the Gauss hypergeometric functions also satisfy well-known recurrence relations among themselves. It is then clear that in the calculations of any form factor by the present decomposition method, all the hypergeometric functions involved can also be generated from a small, basic set of functions through appropriate application of these recurrence relations.

IV. CONCLUSIONS

We have presented the bipolar expansions of tensor fields in a simple but general manner. The simplicity and generality of our results stem from the fact that full advantage has been taken of the symmetry through the angular parts in the expansion. It has been demonstrated that, irrespective of the tensor field in question, the radial coefficient functions are simple integral transform of a class of "universal" form factors $\{T_{\lambda\lambda}^L\}$. Our key observations are characterized by two complementary aspects. Firstly these form factors satisfy recurrence relations, therefore they can be generated recursively from appropriately chosen subclasses

of members. In addition we have developed a computational scheme from which such a basic set of form factors can be constructed analytically.

Accordingly, our results not only exhibit the basic structure of such expansions, but they also provide a feasible computational scheme which can be easily implemented in practical physical problems.

ACKNOWLEDGMENTS

The authors want to express their appreciation of support from WNSL Yale University, Brookhaven National Laboratory and the Niels Bohr Institute at various stages in the development of this work.

APPENDIX A

We now derive the basic formula for bipolar expansion. The starting point is Eq. (5). First consider the plane-wave expansion,

$$e^{ik \cdot r} = 4\pi \sum_{\lambda} (-i)^{\lambda} \hat{\lambda} j_{\lambda}(kr) [Y_{\lambda}(\hat{k}) \otimes Y_{\lambda}(\hat{r})]_0^0, \tag{A1}$$

where the angular parts are expressed in a coupled representation. Then,

$$e^{ik \cdot (r+R)} = (4\pi)^2 \sum_{\lambda\lambda'} (-i)^{\lambda+\lambda'} \hat{\lambda} \hat{\lambda}' j_{\lambda}(kr) j_{\lambda'}(kR)$$

$$\begin{aligned} & \times [Y_\lambda(\hat{k}) \otimes Y_\lambda(\hat{r})]_0^0 [Y_\lambda(\hat{k}) \otimes Y_\lambda(\hat{R})]_0^0 \\ & = (4\pi)^2 \sum_{\lambda\lambda} (-i)^{\lambda+A} \hat{\lambda} \hat{\lambda} j_\lambda(kr) j_\lambda(kR) \\ & \quad \times \sum_{\mathcal{L}} (2\mathcal{L} + 1) \begin{Bmatrix} \lambda & A & \mathcal{L} \\ \lambda & A & \mathcal{L} \\ 0 & 0 & 0 \end{Bmatrix} \\ & \quad \times \{ [Y_\lambda(\hat{k}) \otimes Y_\lambda(\hat{k})]_{\mathcal{L}} \otimes [Y_\lambda(\hat{r}) \otimes Y_\lambda(\hat{R})]_{\mathcal{L}} \}_0^0. \quad (\text{A2}) \end{aligned}$$

To arrive at the last expression in Eq. (A2), some standard angular momentum coupling has been invoked. It can be further simplified by the reduction formulas,

$$\begin{Bmatrix} \lambda & A & \mathcal{L} \\ \lambda & A & \mathcal{L} \\ 0 & 0 & 0 \end{Bmatrix} = (\hat{\lambda} \hat{A} \hat{\mathcal{L}})^{-1} \quad (\text{A3})$$

and

$$[Y_\lambda(\hat{k}) \otimes Y_\lambda(\hat{k})]_{\mathcal{L}} = (-)^{\mathcal{L}} \frac{\hat{\lambda} \hat{A}}{(4\pi)^{\mathcal{L}}} \begin{Bmatrix} \lambda & A & \mathcal{L} \\ \lambda & A & \mathcal{L} \\ 0 & 0 & 0 \end{Bmatrix} Y_{\mathcal{L}}(\hat{k}). \quad (\text{A4})$$

Equation (A2) is then reduced to the following:

$$\begin{aligned} & e^{i\mathbf{k} \cdot (\mathbf{r} + \mathbf{R})} \\ & = (4\pi)^{3/2} \sum_{\lambda\lambda\mathcal{L}} i^{\lambda+A} \hat{\lambda} \hat{A} \hat{\mathcal{L}} j_\lambda(kr) j_\lambda(kR) \\ & \quad \times \begin{Bmatrix} \lambda & A & \mathcal{L} \\ \lambda & A & \mathcal{L} \\ 0 & 0 & 0 \end{Bmatrix} \{ Y_{\mathcal{L}}(\hat{k}) \otimes [Y_\lambda(\hat{r}) \otimes Y_\lambda(\hat{R})]_{\mathcal{L}} \}_0^0 \\ & = (4\pi)^{3/2} \sum_{\lambda\lambda} i^{\lambda+A} \hat{\lambda} \hat{\lambda} j_\lambda(kr) j_\lambda(kR) \\ & \quad \times \sum_{\mathcal{L}} (-1)^{\mathcal{L}} \begin{Bmatrix} \lambda & A & \mathcal{L} \\ \lambda & A & \mathcal{L} \\ 0 & 0 & 0 \end{Bmatrix} (Y_{\mathcal{L}}(\hat{k}) \cdot [Y_\lambda(\hat{r}) \otimes Y_\lambda(\hat{R})]_{\mathcal{L}}). \quad (\text{A5}) \end{aligned}$$

A substitution of Eq. (A5) into Eq. (5) and an integration over the angular parts in the \mathbf{k} -space leads to

$$\begin{aligned} \Psi_L^M(\mathbf{r} + \mathbf{R}) & = 2^{3/2} \sum_{\lambda\lambda} i^{\lambda+A+L} \hat{\lambda} \hat{\lambda} \begin{Bmatrix} \lambda & A & L \\ \lambda & A & L \\ 0 & 0 & 0 \end{Bmatrix} \\ & \quad \times \int_0^\infty dk k^2 j_\lambda(kr) j_\lambda(kR) \tilde{f}_L(k) [Y_\lambda(\hat{r}) \otimes Y_\lambda(\hat{R})]_L^M \quad (\text{A6}) \end{aligned}$$

which is the expressions in Eqs. (6), (9) and (11). It is clear that the indices λ and A are restricted by angular momentum coupling and hence the conditions (7). Also, because of the parity rule for $\begin{Bmatrix} \lambda & A & L \\ \lambda & A & L \\ 0 & 0 & 0 \end{Bmatrix}$, the only contributions to expansion (A6) come from the terms for which $\lambda + A + L = \text{even}$.

APPENDIX B

We now derive the basic recurrence relation. From the definition (11) of the form factors,

$$\begin{aligned} & F_{\lambda, \lambda+1}^L(r, R) + F_{\lambda+2, \lambda-1}^L(r, R) \\ & = \int_0^\infty dk k^2 [j_\lambda(kr) + j_{\lambda+2}(kr)] j_{\lambda+1}(kR) \tilde{f}_L(k) \\ & = \int_0^\infty dk k^2 \frac{(2\lambda+3)}{kr} j_{\lambda+1}(kr) j_{\lambda+1}(kR) \tilde{f}_L(k) \\ & = \int_0^\infty dk k^2 (2\lambda+3) \left(\frac{R}{r}\right) j_{\lambda+1}(kr) \frac{j_{\lambda+1}(kR)}{kR} \tilde{f}_L(k) \\ & = \int_0^\infty dk k^2 \left(\frac{2\lambda+3}{2\lambda+3}\right) \left(\frac{R}{r}\right) j_{\lambda+1}(kr) \\ & \quad \times [j_\lambda(kR) + j_{\lambda+2}(kR)] \tilde{f}_L(k) \\ & = \frac{2\lambda+3}{2\lambda+3} \left(\frac{R}{r}\right) [F_{\lambda+1, \lambda}^L(r, R) + F_{\lambda+1, \lambda+2}^L(r, R)], \quad (\text{B1}) \end{aligned}$$

where the expressions in the second and fourth equalities have been obtained from the recurrence relation Eq. (12) of spherical Bessel functions.

¹Some preliminary results of the present work had been reported by the authors: H. H. K. Tang and J. S. Vaagen, *Bull. Am. Phys. Soc.* **23**, 586 (1978).

²Bipolar expansions for special scalar functions have been discussed by various authors. The following list, by no means exhaustive, presents some typical examples: R. Nozawa, *J. Math. Phys.* **7**, 1841 (1966); A. R. Sourour and A. A. Ashour, *J. Math. Phys.* **11**, 56 (1970); A. K. Rafiqullah, *J. Math. Phys.* **12**, 549 (1971).

³A. de-Shalit and I. Talmi, *Nuclear Shell Theory* (Academic, New York, 1963).

⁴The choice of square-integrable functions renders possible the simple derivation by the Fourier transform method. Strictly speaking, the results in this paper can be valid for tensor fields satisfying weaker conditions. For example, if the radial function $f_L(r)$ grows asymptotically as a polynomial of r of finite order, our derivation still applies for a modified radial function $e^{-\alpha r^2} f_L(r)$. The final results can then be obtained by taking the limit $\alpha^2 \rightarrow 0$.

⁵M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, (Dover, New York, 1965); I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 4th ed., translated by A. Jeffrey (Academic, New York 1965).

4-space formulation of field equations for multicomponent eigenfunctions

John R. Fanchi

2709 West Greeley Street, Broken Arrow, Oklahoma 74012

(Received 30 November 1979; accepted for publication 20 March 1980)

Beginning with the assumptions employed in the development of the 4-space formulation (FSF) for spinless particles, a formalism for multicomponent eigenfunctions is constructed. The primary result is a general expression for the field equations of multicomponent eigenfunctions. The "Relativistic Dynamics" of Horwitz, Piron, and Reuse for spin-0 and spin- $\frac{1}{2}$ particles is shown to be consistent with the FSF. Expectation values are defined and briefly discussed in the appendix.

PACS numbers: 03.70. + k, 03.65.Ca, 11.10.Qr

INTRODUCTION

A probabilistic foundation for the quantum mechanical description of relativistic spinless particles was recently provided.¹⁻³ This theory, referred to as the 4-space formulation (FSF), is a consistent single-particle theory of relativistic spinless particles in the presence of an arbitrary 4-vector potential. As mentioned in Ref. 1, the FSF is capable of describing particles with nonzero spin (such as the "Relativistic Dynamics" of Horwitz, Piron, and Reuse⁴⁻⁷) and nonelectromagnetic interactions. The purpose of this paper is to present the extension of the FSF to include particles with spin and nonelectromagnetic interactions. Given the assumptions of Refs. 1-3, this extension becomes primarily an exercise in mathematics. The results reported here should be useful in many ways. Three of the most important are the following.

First, an extension of the FSF will show the consistency of "Relativistic Dynamics" with the FSF. This is particularly significant for spin- $\frac{1}{2}$ particles because the spin- $\frac{1}{2}$ formalism of Horwitz, Piron, and Reuse⁵⁻⁷ differs from an earlier formalism suggested by Nambu.⁸

A second reason for extending the FSF is to determine the general form of field equations which can be considered candidates for multicomponent eigenfunctions. An expression of this kind will help guide the construction of acceptable Lagrangian densities for use in the generalized quantum field theory (GQFT) presented recently.⁹

Finally, this extension will aid in devising experiments using particles with spin and undergoing interactions other than the electromagnetic interaction. One such example, relativistic scattering from a step potential using leptons (Klein's paradox with leptons), may be particularly useful for testing the validity of the FSF.

The extension of the FSF parallels the derivation of the spinless particle formulation which in turn, parallels Collins's¹⁰ nonrelativistic formulation. Beginning with the same assumptions, a multicomponent formalism is constructed. From this formalism a general expression for the field equations of multicomponent eigenfunctions is derived. Two simple examples show that the work of Refs. 1-7 and 11 are all consistent with this formulation. A discussion of expectation values is appended.

FORMALISM

Define a conditional probability density $\rho(x|\tau)$ satisfying the following positive-definite, normalization, and conservation equations, respectively:

$$\rho(x|\tau) \geq 0, \quad (1a)$$

$$\int \rho(x|\tau) d^4x = 1, \quad d^4x \equiv dx^0 dx^1 dx^2 dx^3, \quad (1b)$$

and

$$\partial_\tau \rho(x|\tau) + \partial_\mu [\rho(x|\tau) V^\mu] = 0. \quad (1c)$$

Here x signifies the space-time coordinates (\vec{x}, ct) , τ is an independent scalar parameter identified as proper time, the domain of integration of the μ th coordinate is $|x^\mu| \leq \infty$, the nonzero metric elements are

$$g_{00} = 1 = -g_{11} = -g_{22} = -g_{33}, \quad (2)$$

and the operators $\partial_\tau, \partial_\mu$ signify differentiation with respect to τ and x^μ , respectively. The 4-vector V^μ is related to the τ rate of change of the 4-position expectation value $\langle x^\mu \rangle$. It is discussed in the Appendix.

Assume that the conditional probability $\rho(x|\tau) d^4x$ can be written in terms of a sum over L independent parameters:

$$\rho(x|\tau) d^4x = \sum_{l_1} \sum_{l_2} \dots \sum_{l_L} \rho(x, l_1, l_2, \dots, l_L | \tau) d^4x. \quad (3)$$

The sum is over the entire range of allowed values of (l_1, \dots, l_L) . The parameter labels $\{l_i\}$ represent "hidden" variables in the sense that Eq. (1c) is satisfied without explicit reference to $\{l_i\}$. Experiments such as the Stern-Gerlach experiment indicate that the probabilities of quantum mechanics must include discrete labels. These labels can be expressed as in Eq. (3). Physically, these parameters could represent spin, isospin, hypercharge, etc. Their physical interpretation will not be examined here. Rather, let us examine what mathematics says about $\rho(x|\tau) d^4x$.

Equation (3) can be written as

$$\rho(x|\tau) d^4x = \sum_{l_1, \dots, l_L} P(l_1|\tau) P(l_2|\tau) \dots P(l_L|\tau) \times \rho(x|l_1, \dots, l_L, \tau) d^4x, \quad (4)$$

where the sums $\Sigma_{l_1}, \Sigma_{l_2}, \dots, \Sigma_{l_L}$ are denoted by Σ_{l_1, \dots, l_L} and $P(l_i|\tau)$ is a conditional probability satisfying

$$\sum_{l_i} P(l_i|\tau) = 1. \quad (5)$$

Notice that Eq. (4) is possible *only if* the parameters $\{l_i\}$ characterize mutually independent events. Furthermore, Eqs. (1b) and (5) imply the normalization condition

$$\int \rho(x|l_1, \dots, l_L, \tau) d^4x = 1, \quad (6)$$

as expected. Mathematically, the ensuing derivation does not need to assume $\{l_i\}$ are all independent events. It is only necessary to assume $\rho(x|\tau)$ can be written as in Eqs. (17) and (18) below. The independence assumption here is motivated by the physical observation that the discrete variables $\{l_i\}$ label states of commuting observables. The notation is simplified by defining

$$\rho \equiv \rho(x|\tau), \quad (7a)$$

$$\rho_s \equiv \rho(x|l_1, \dots, l_L, \tau), \quad (7b)$$

and

$$P \equiv P(l_1|\tau)P(l_2|\tau)\dots P(l_L|\tau). \quad (7c)$$

Observe that P depends on τ but not $\{x^\mu\}$.

The Hilbert-space formalism is now developed by invoking the Born representation; thus write

$$\rho_s \equiv \psi^*(x, \tau, l_1, \dots, l_L) \psi(x, \tau, l_1, \dots, l_L) \geq 0, \quad (8)$$

or

$$\psi(x, \tau, l_1, \dots, l_L) = \rho_s^{1/2} e^{i\phi(x, \tau)}, \quad (9)$$

where ψ and ψ^* are Lorentz-invariant scalars and ϕ is an as yet undetermined real scalar function which is independent of $\{l_i\}$. It is straightforward to derive from Eq. (9) the following useful identity:

$$\partial^\mu \phi = (-i/2\rho_s)(\psi^* \partial^\mu \psi - \psi \partial^\mu \psi^*). \quad (10)$$

As in the FSF,¹⁻³ let us express V^μ in the form

$$V^\mu = \epsilon_1 \partial^\mu \phi + \epsilon_2 A^\mu, \quad (11)$$

where ϵ_1 and ϵ_2 are c -numbers and it is assumed that A^μ , hence V^μ , are independent of $\{l_i\}$. Equations (8), (10), and (11) are now combined to obtain the relation

$$\begin{aligned} \partial_\mu(\rho_s V^\mu) &= \epsilon_2 \psi^* (\frac{1}{2} A^\mu \partial_\mu \psi + \frac{1}{2} \partial_\mu A^\mu \psi) \\ &\quad + \epsilon_2 \psi (\frac{1}{2} A^\mu \partial_\mu \psi^* + \frac{1}{2} \partial_\mu A^\mu \psi^*) \\ &\quad - \frac{1}{2} \psi^* i \epsilon_1 \partial_\mu \partial^\mu \psi + \frac{1}{2} \psi i \epsilon_1 \partial_\mu \partial^\mu \psi^*. \end{aligned} \quad (12)$$

This relation will be useful in expanding the probability conservation equation [Eq. (1c)].

Employing Eqs. (4) and (7) in (1c) yields

$$\partial_\tau \rho + \sum_{l_1, \dots, l_L} P \partial_\mu(\rho_s V^\mu) = 0. \quad (13)$$

Substituting Eq. (12) into (13) gives

$$\begin{aligned} \sum_{l_1, \dots, l_L} \{ \partial_\tau(P\rho_s) + P [\epsilon_2 \psi^* (\frac{1}{2} A^\mu \partial_\mu \psi + \frac{1}{2} \partial_\mu A^\mu \psi) \\ + \epsilon_2 \psi (\frac{1}{2} A^\mu \partial_\mu \psi^* + \frac{1}{2} \partial_\mu A^\mu \psi^*) \\ - \frac{1}{2} \psi^* i \epsilon_1 \partial_\mu \partial^\mu \psi + \frac{1}{2} \psi i \epsilon_1 \partial_\mu \partial^\mu \psi^*] \} = 0. \end{aligned} \quad (14)$$

Further physically interesting restructuring of Eq. (14) requires some additional definitions.

The probability $P(l_i|\tau)$ can be written as

$$P(l_i|\tau) = u^*(l_i, \tau) u(l_i, \tau) \equiv u_i^* u_i, \quad (15)$$

where u_i is a scalar that depends only on τ and l_i . Denote the product of $\{u_i\}$ by u , i.e.,

$$u \equiv u(l_1, \dots, l_L, \tau) = u(l_1, \tau) u(l_2, \tau) \dots u(l_L, \tau). \quad (16)$$

This is true also for u^* . Then Eq. (4) now has the form

$$\rho(x|\tau) = \sum_{l_1, \dots, l_L} \psi_i^* \psi_i, \quad (17)$$

where

$$\psi_i = u(l_1, \dots, l_L, \tau) \psi(x, \tau, l_1, \dots, l_L), \quad (18)$$

and a similar expression holds for ψ_i^* .

These definitions are now used in Eq. (14) to yield

$$\begin{aligned} 0 &= \sum_{l_1, \dots, l_L} \{ \psi_i^* \partial_\tau \psi_i + \psi_i \partial_\tau \psi_i^* \\ &\quad + \psi_i^* [-\frac{1}{2} i \epsilon_1 \partial_\mu \partial^\mu \psi_i + \frac{1}{2} \epsilon_2 (A^\mu \partial_\mu \psi_i + \partial_\mu A^\mu \psi_i)] \\ &\quad + \psi_i [\frac{1}{2} i \epsilon_1 \partial_\mu \partial^\mu \psi_i^* + \frac{1}{2} \epsilon_2 (A^\mu \partial_\mu \psi_i^* + \partial_\mu A^\mu \psi_i^*)] \}. \end{aligned} \quad (19)$$

Multiplying by $i\epsilon_3$, where ϵ_3 is a real c -number, and then rearranging gives

$$\begin{aligned} \sum_{l_1, \dots, l_L} \{ \psi_i^* [i \epsilon_3 \partial_\tau \psi_i + \frac{1}{2} \epsilon_1 \epsilon_3 \partial_\mu \partial^\mu \psi_i \\ + \frac{1}{2} i \epsilon_2 \epsilon_3 (A^\mu \partial_\mu \psi_i + \partial_\mu A^\mu \psi_i)] \} \\ = \sum_{l_1, \dots, l_L} \{ -i \epsilon_3 \partial_\tau \psi_i^* + \frac{1}{2} \epsilon_1 \epsilon_3 \partial_\mu \partial^\mu \psi_i^* \\ - \frac{1}{2} i \epsilon_2 \epsilon_3 (A^\mu \partial_\mu \psi_i^* + \partial_\mu A^\mu \psi_i^*) \} \psi_i \}. \end{aligned} \quad (20)$$

Denote the number of allowed values of l_i as L_i . Then the number N of terms in the sum Σ_{l_1, \dots, l_L} is the product of the L_i 's, i.e.,

$$N = \prod_{i=1}^L L_i. \quad (20a)$$

Let us replace the L sums $\Sigma_{l_i=1}^{L_i}$ with one sum over the range $1 \leq n \leq N$ and assign a one-to-one correspondence between n and each term of the sum Σ_{l_1, \dots, l_L} . Equation (20) can be written now as

$$\sum_{n=1}^N \psi_n^* F_n = \sum_{n=1}^N F_n^* \psi_n, \quad (21)$$

where

$$\begin{aligned} F_n &\equiv i \epsilon_3 \partial_\tau \psi_n + \frac{1}{2} \epsilon_1 \epsilon_3 \partial_\mu \partial^\mu \psi_n \\ &\quad + \frac{1}{2} i \epsilon_2 \epsilon_3 (A^\mu \partial_\mu \psi_n + \partial_\mu A^\mu \psi_n). \end{aligned} \quad (22)$$

Equation (21) has the form

$$\Psi^* F = F^* \Psi, \quad (23)$$

where $*$ means conjugate transpose and Ψ is the N -column vector

$$\Psi = \begin{bmatrix} \psi_1 \\ \vdots \\ \psi_N \end{bmatrix}. \quad (24)$$

The n th element of the N -vector F is given by Eq. (22). It

should be noted that, although a metric was specified in Eq. (2), the above derivation does not depend on a particular choice of metric.

By decoupling Eq. (23) it is possible to obtain field equations which satisfy the probabilistic constraints imposed at the outset of this derivation. Only these field equations will be considered candidates for describing physically realizable systems.

FIELD EQUATIONS

The decoupling of Eq. (23) begins by observing that Ψ^*F must be real because

$$\Psi^*F = F^*\Psi = (\Psi^*F)^* \quad (25)$$

The decoupling technique will be demonstrated by two examples that make use of the reality of Ψ^*F . These examples provide a basis for the most common field equations.

Scalar decoupling

To assure the reality of Ψ^*F , let us assume F is an N -vector with the n -th element having the form $U\psi_n$ where U is a real scalar. This assumption yields

$$F = U\Psi \quad (26)$$

and

$$\Psi^*F = \Psi^*U\Psi = U\Psi^*\Psi = U\rho. \quad (27)$$

The product $U\rho$ is real since U and ρ are real. An example of this decoupling is obtained by defining

$$\begin{aligned} \epsilon_1 &= \hbar/\bar{m}, \\ \epsilon_2 &= -e/\bar{m}c, \\ \epsilon_3 &= \hbar, \end{aligned} \quad (28)$$

and

$$U = e^2 A^\mu A_\mu / 2\bar{m}c^2, \quad (29)$$

with the result,

$$i\hbar\partial_\tau \Psi = (1/2\bar{m})p^\mu p_\mu \Psi, \quad (30)$$

where

$$p^\mu \equiv (\hbar/i)\partial^\mu - (e/c)A^\mu. \quad (31)$$

Equation (30) is just the generalized Schrödinger equation of Refs. 1-3, 4, and 11, applied now to a multicomponent eigenfunction. This result illustrates two important results.

First, Eq. (30) was obtained by defining U as in Eq. (29). This definition is based on the requirements that U should be both Lorentz- and gauge-invariant. The gauge-invariance is necessary to preserve the values of V^μ and ρ under gauge transformations. Lorentz-invariance is imposed in order to preserve the form of the equations when viewed by observers in different reference frames. Both of these invariance properties are physically meaningful, but neither the gauge-invariance [necessitated by the assumptions of Eqs. (8) and (11)] nor the covariant formulation of the equations is required by probability theory. Probability theory does not provide all of the information needed to completely define the specific form of quantum mechanics. Additional requirements, such as Lorentz-invariance, must also be imposed,

but in such a way that the probabilistic basis of the theory is retained.

The second notable result is that the analysis of the Klein paradox in Refs. 1 and 3 should apply, to some level of approximation, to particles with spin. This conclusion is based on the fact that multicomponent eigenfunctions satisfy Eq. (30). Thus, leptons should behave, at least to some approximate degree, as described in Refs. 1 and 3. This significantly lowers the magnitude of the step potential used in the scattering problem discussed in Refs. 1 and 3. The subsequent step potential may be physically realizable. If so, then the scattering of a lepton from a step potential provides an experimental test of the FSF.

Matrix Decoupling

Defining F as in Eq. (26) is the simplest means of assuring the reality of Ψ^*F . The reality constraint is more generally satisfied as follows. Suppose

$$F = V\Psi, \quad (32)$$

where V is an $N \times N$ matrix. Then Eq. (25) implies that

$$\Psi^*F = \Psi^*V\Psi = F^*\Psi. \quad (33)$$

But the relation

$$F^* = (V\Psi)^* = \Psi^*V^*, \quad (34)$$

must also be true. Comparing Eqs. (33) and (34) yields the result

$$V^* = V, \quad (35)$$

i.e., the square matrix V must be self-adjoint (or Hermitian). The general form of the field equations represented by Eq. (32) is

$$\begin{aligned} [i\epsilon_3\partial_\tau + \frac{1}{2}\epsilon_1\epsilon_3\partial_\mu\partial^\mu + \frac{1}{2}i\epsilon_2\epsilon_3(A^\mu\partial_\mu + \partial_\mu A^\mu)]\Psi - V\Psi \\ = 0. \end{aligned} \quad (36)$$

It should be noted here that Lagrangian densities which yield field equations having the above form are considered acceptable candidates for use in generalized quantum field theory.⁹

The simplest example of V is

$$V = UI, \quad (37)$$

where I is the identity matrix and U is the real scalar defined in the previous subsection. Another example is given by Piron and Reuse.^{6,7} The derivation of Eqs. (30) and (36) from a probabilistic basis shows that the "Relativistic Dynamics" of Horwitz, Piron, and Reuse⁴⁻⁷ for both spin-0 and spin- $\frac{1}{2}$ particles is consistent with the FSF.

ACKNOWLEDGMENT

The author would like to thank Professor R. E. Collins for suggesting the decomposition employed in Eq. (3).

APPENDIX: EXPECTATION VALUES

The last topic to be considered here is the definition of expectation values. As usual, the expectation value of position is¹⁻⁶

$$\langle x^\mu \rangle = \int x^\mu \rho d^4x = \int \Psi^* x^\mu \Psi d^4x. \quad (\text{A1})$$

The τ -derivative of $\langle x^\mu \rangle$ is

$$\frac{d \langle x^\mu \rangle}{d\tau} = \int x^\mu \frac{\partial \rho}{\partial \tau} d^4x = - \int x^\mu \frac{\partial \rho V^\nu}{\partial x^\nu} d^4x, \quad (\text{A2})$$

by Eq. (1c) and noting x^μ is independent of τ . The expression $x^\mu \partial_\nu (\rho V^\nu)$ can be written as

$$x^\mu \partial_\nu (\rho V^\nu) = \partial_\nu (\rho x^\mu V^\nu) - V^\nu \rho \partial_\nu x^\mu \\ = \partial_\nu (\rho x^\mu V^\nu) - V^\mu \rho. \quad (\text{A3})$$

Equation (A3) is valid whenever the coordinate x^μ is independent of x^ν for $\nu \neq \mu$, i.e., if

$$\partial_\nu x^\mu = \delta_\nu^\mu, \quad (\text{A3a})$$

then Eq. (A3) is satisfied, where δ_ν^μ is the Kronecker delta. Substituting this result into Eq. (A2), applying the divergence theorem, and then assuming that ρ vanishes as $|x^\mu| \rightarrow \infty$, yields the result

$$\frac{d \langle x^\mu \rangle}{d\tau} = \int \rho V^\mu d^4x = \langle V^\mu \rangle. \quad (\text{A4})$$

Thus the expectation value of V^μ gives the τ -rate of change of $\langle x^\mu \rangle$. A more familiar quantum mechanical expression for $\langle V^\mu \rangle$ is obtained by employing Eqs. (4), (8), (10), and (11) to find

$$\frac{d \langle x^\mu \rangle}{d\tau} = \sum_{i, -i} u^* u \int \left[-\frac{1}{2} i \epsilon_1 (\psi^* \partial^\mu \psi - \psi \partial^\mu \psi^*) \right. \\ \left. + \epsilon_2 \psi^* A^\mu \psi \right] d^4x. \quad (\text{A5})$$

Substituting the relation

$$\partial^\mu \psi^* \psi = \psi^* \partial^\mu \psi + \psi \partial^\mu \psi^* \quad (\text{A6})$$

into Eq. (A5) yields

$$\frac{d \langle x^\mu \rangle}{d\tau} = \sum_{i, -i} \left\{ u^* u \int \left[\frac{\epsilon_1}{i} \psi^* \partial^\mu \psi + \epsilon_2 \psi^* A^\mu \psi \right] d^4x \right. \\ \left. + u^* u \int \left[-\frac{\epsilon_1}{2i} \partial^\mu \psi^* \psi \right] d^4x \right\}. \quad (\text{A7})$$

Applying the divergence theorem to the integral of $\partial^\mu \psi^* \psi$ and evaluating the subsequent surface integral at $|x^\mu| \rightarrow \infty$ (where $\psi^* \psi \rightarrow 0$) gives the result

$$\frac{d \langle x^\mu \rangle}{d\tau} = \sum_{i, -i} u^* u \int \psi^* \left[\frac{\epsilon_1}{i} \partial^\mu + \epsilon_2 A^\mu \right] \psi d^4x, \quad (\text{A8})$$

since the integral of $\partial^\mu \psi^* \psi$ vanishes. Equation (A5) can be rewritten as

$$\langle V^\mu \rangle = \sum_n \int \psi_n^* \left[\frac{\epsilon_1}{i} \partial^\mu + \epsilon_2 A^\mu \right] \psi_n d^4x. \quad (\text{A9})$$

If Eqs. (28) are assumed here, Eq. (A9) becomes the familiar result

$$\bar{m} \frac{d \langle x^\mu \rangle}{d\tau} = \sum_n \int \psi_n^* \left[\frac{\hbar}{i} \partial^\mu - \frac{e}{c} A^\mu \right] \psi_n d^4x, \quad (\text{A10})$$

or

$$\langle p^\mu \rangle = \int \Psi^* \left[\frac{\hbar}{i} \partial^\mu - \frac{e}{c} A^\mu \right] \Psi d^4x. \quad (\text{A11})$$

Equations (A1) and (A11) suggest that the usual definition of the expectation values of an observable Ω be adopted; thus

$$\langle \Omega \rangle \equiv \int \Psi^* \Omega \Psi d^4x. \quad (\text{A12})$$

This definition is employed by Piron and Reuse⁶ for their spin- $\frac{1}{2}$ formalism. By also imposing a superselection rule they demonstrated that their formulation, which is a special case of the FSF, yields the usual results of Dirac's spin- $\frac{1}{2}$ theory.

¹J. R. Fanchi and R. E. Collins, *Found. Phys.* **8**, 851 (1978).

²R. E. Collins and J. R. Fanchi, *Nuovo Ciminto A* **48**, 314 (1978).

³J. R. Fanchi, Ph.D. Dissertation, Univ. of Houston, 1977 (Univ. of Michigan Microfilm).

⁴L. P. Horwitz and C. Piron, *Helv. Phys. Acta* **46**, 316 (1973).

⁵L. P. Horwitz, C. Piron, and F. Reuse, *Helv. Phys. Acta* **48**, 546 (1975).

⁶C. Piron and F. Reuse, *Helv. Phys. Acta* **51**, 146 (1978).

⁷F. Reuse, *Helv. Phys. Acta* **51**, 157 (1978).

⁸Y. Nambu, *Prog. Theor. Phys.* **5**, 82 (1950).

⁹J. R. Fanchi, *Phys. Rev. D* **20**, 3108 (1979).

¹⁰R. E. Collins, *Found. Phys.* **7**, 475 (1977); *Lett. Nuovo Ciminto* **18**, 581 (1977). An alternative approach to the probabilistic formulation of nonrelativistic quantum mechanics is given by P. Huguenin, *Z. Naturforsch. A* **28**, 1090 (1973) and also J.-P. Amiet and P. Huguenin, *Helv. Phys. Acta.* (to be published).

¹¹J. H. Cooke, *Phys. Rev.* **166**, 1293 (1968).

Interior dynamics in the unified connected reaction theory of Polyzou and Redish

Wayne Polyzou

Theoretical Division, Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545

(Received 26 September 1980; accepted for publication 21 November 1980)

We discuss the unified reaction theory of Polyzou and Redish [Ann. Phys. (N.Y.) **119**, 1(1979)] with regard to the freedom to perturb the approximate Hamiltonian by a connected operator. A proper treatment of this freedom is crucial for a good treatment of the dynamics. We show how to construct an effective N -body force so the approximate Hamiltonian is a projection of the exact Hamiltonian on an appropriate infinite dimensional subspace. Because this operator is connected it should not alter the unitarity considerations in the above paper. The methods used in constructing this effective interaction can also be used to reformulate the scattering integral equations as equivalent equations with noncompact contractive kernels. This reduces the scattering problem to uniformly convergent perturbation theory.

PACS numbers: 03.80. + r, 24.10.Dp

I. INTRODUCTION

In a previous paper,¹ which we denote by I, E. F. Redish and I developed a unified connected theory of few-body reaction mechanisms. The goal of this work is to combine the best features of conventional reaction theories with new developments in formal N -body scattering theory to construct systematic approximation schemes that go beyond the conventional ones. The theory proposed in I is an extension of the two-potential formalism of Gell-Mann and Goldberger.² In the two-potential approach the N -body Hamiltonian is expressed as a sum of two terms:

$$H = H_I + H_{II}, \quad (1.1)$$

where H_I can be treated exactly and H_{II} is a more complicated, but less important perturbation. In the standard theory H_I is usually a simple two-body effective Hamiltonian that accounts for most of the elastic scattering. The extension proposed in I allows H_I to be a few-body operator that may include breakup and rearrangement degrees of freedom in addition to elastic scattering. Extensions of this type allow us to treat a larger part of the problem explicitly, and are commensurate with our extended calculational capabilities.

In I the choice of H_I is partially dictated by unitarity considerations. What this means is that the part of the dynamics determined by H_I requires that all of the scattered flux comes out in a chosen set of asymptotic channels. Thus, in the lowest order approximation, the coupling to all of the residual channels is turned off. This unitarity constraint does not uniquely determine the operator H_I . In I it is argued that the unitarity constraint on the decomposition (1.1) is unchanged if we perturb the decomposition by an effective Hermitian N -body potential. By an N -body potential we mean one that falls off sufficiently fast as *any* of the interparticle coordinates are asymptotically separated. The main purpose of this paper is to provide a prescription for dealing with this additional degree of freedom in the decomposition given in I. We refer to this as the interior dynamics problem.

Our solution to the interior dynamics problem has sev-

eral desirable features that confront some of the criticisms levied at truncated few-body theories. What we find is that by solving an appropriate nonorthogonality problem we can construct an effective Hermitian N -body potential U that leads to a decomposition satisfying

$$H_I \rightarrow H_I + U = IIH_I, \quad (1.2)$$

for an appropriate orthogonal projector II . This suggests, at least in the context of I, that (i) one can express truncated few-body theories as projected Hamiltonian theories, (ii) there are clearly no overcounting problems with the decomposition associated with (1.2), and (iii) one does not avoid the nonorthogonality problem by starting with a well posed N -body equation and truncating.

These comments should clarify some aspects of the underlying structure of the approximations presented in I. The II constructed in this paper depends on the choice of dominant reaction mechanism.¹ The important asymptotic channels live on different nonorthogonal subspaces and in the approximation (1.2) are coupled through the overlap between these subspaces and the range of II . On the other hand, the range of II is constrained by the condition that (1.2) only allows scattering in the set of asymptotic channels associated with the dominant reaction mechanism. It is our hope that the construction presented here emphasizes the most important aspects of these couplings consistent with the limitations imposed by the optical theorem.

Our method of dealing with the nonorthogonality problem suggests alternate formulations of the scattering integral equations with contractive kernels. Because this may have use beyond the applications suggested in this paper we give a discussion of these techniques in Sec. III.

This paper is divided into seven sections. In Sec. II we introduce our notation. In Sec. III we introduce the notion of the Moore–Penrose generalized inverse for operators.^{3–5} This will be one of our main tools in the paper. In Sec. IV we show how to formally construct the connected potential in (1.2). This construction uses the Moore–Penrose generalized inverse. In Sec. V we discuss the construction of this effective

potential from a practical point of view. We give a convergent iterative method for the construction of the operators used in this effective potential. Section VI is a discussion of how the Moore–Penrose techniques required to solve the nonorthogonality problem may be applied to the scattering problem. The result is that one may replace the scattering integral equation by an equivalent equation with a contractive kernel. (Assuming the scattering equation can be formulated as a compact kernel equation and has a unique solution). The last section contains a summary and conclusion.

II. NOTATION

The notation utilized in this paper conforms with the notation used in Ref. 6. N -body operators are indexed by partitions, a, b, c, \dots of the N particles into n_a, n_b, n_c, \dots disjoint groups or clusters. We reserve the symbol $\mathbf{1}$ to denote the unique 1-cluster partition. We use upper case Latin letters A, B, C, \dots to denote N -body operators. For an operator A we let A_a denote the operator obtained from A by turning off all interactions between the particles in different clusters of a . Residual operators, A^a , are defined by $A^a = A - A_a$.

Additional classifications of the N -body operators utilize the lattice structure^{6–8} on the set of partitions. We say $b \supseteq a$ if a can be obtained by breaking up some of the clusters of b . The Zeta and Möbius function for the partition lattice are defined by Ref. 9,

$$\delta_{a \supseteq b} = \begin{cases} 1, & \text{if } a \supseteq b, \\ 0, & \text{otherwise,} \end{cases} \quad (2.1)$$

and

$$\delta_{a \supseteq b}^{-1} : \sum_c \delta_{a \supseteq c}^{-1} \delta_{c \supseteq b} = \delta_{a, b}, \quad (2.2)$$

respectively. The Möbius function, $\delta_{a \supseteq b}^{-1}$, has been calculated elsewhere.⁸

In graph theoretical language the a -connected part, $[A]_a$, of an operator A is the sum of all graphs in perturbation theory with interactions between every particle in the same cluster of a and no interaction between particles in different clusters of a . We may construct $[A]_a$ in terms of the A_b 's using the Möbius function:

$$[A]_a = \sum_b \delta_{a \supseteq b}^{-1} A_b. \quad (2.3)$$

The origin of this expansion is easy to see if we invert the Möbius function in (2.3). This leads to an expression for A_b as a simple sum of those $[A]_a$ with connectivity $b \supseteq a$. The operators under consideration admit cluster expansions of the form

$$A = \sum_a [A]_a. \quad (2.4)$$

If we insert this in (2.3) we obtain the expansion

$$A = \sum_a \mathcal{C}_a A_a + [A]_1, \quad (2.5)$$

where $[A]_1$ is the completely connected part of A and Refs. 6, 8,

$$\mathcal{C}_a \equiv \sum_b \delta_{b \supseteq a}^{-1} = (-)^{n_a} (n_a - 1)!. \quad (2.6)$$

The prime on the sum indicates the unique 1-cluster partition is eliminated from the sum. We use this notation in all that follows.

This partition labelling convention is used throughout. The operators of interest are the full N -body Hamiltonian, H ; the resolvent of H , $G(z)$; the interaction potential, V ; and the kinetic energy operator, K . We also introduce projection operators $P_a^+(\mathcal{A})$ which are the orthogonal projectors on the invariant subspace of H_a associated with the scattering channels (\mathcal{A}) and outgoing wave boundary conditions. The notation $P_a^+(\mathcal{A})$ also conforms with the above partition labelling conventions. $P_a^+(\mathcal{A})$ is what one obtains from the corresponding projector on the invariant subspace of the full Hamiltonian by turning off the appropriate interactions.¹⁰

In terms of the operators discussed above the general form of the Hamiltonian decomposition introduced in I for a choice of channels \mathcal{A} is

$$H_I = \sum_a \mathcal{C}_a H_a P_a^+(\mathcal{A}) + U, \quad (2.7)$$

$$H_{II} = \sum_a \mathcal{C}_a H_a (\mathbf{1} - P_a^+(\mathcal{A})) - U + V_N,$$

where U is an arbitrary connected Hermitian operator and V_N is a possible N -body force. The only connected terms in this decomposition are U and V_N .

The construction of dynamical equations for the decomposition (2.7) has been thoroughly discussed in I. Our interest in Secs. IV and V is in how to choose and construct a suitable U that best represents the dynamics of the approximation.

III. THE MOORE–PENROSE GENERALIZED INVERSE

In this section we want to introduce a tool that will be utilized in the next three sections. This tool, the Moore–Penrose generalized inverse, is useful in solving the various kinds of nonorthogonality problems that arise in scattering theory.¹¹

For finite dimensional matrix the Moore–Penrose generalized inverse is the unique solution, X , to the four Penrose equations¹²:

$$\begin{aligned} X &= XAX, & A &= AXA, \\ XA &= (XA)^\dagger & AX &= (AX)^\dagger. \end{aligned} \quad (3.1)$$

If the matrix A has an inverse, A^{-1} , $X = A^{-1}$ clearly satisfies (3.1) and is necessarily the unique solution of this system. If A does not have an inverse, the system (3.1) still has a unique solution which we denote by $A^\#$ and call the Moore–Penrose generalized inverse of A . It is this case that is really interesting for the applications to be discussed. The property of interest is that $A^\# A$ is the orthogonal projector on the orthogonal complement of the null space of A .¹³

To proceed further we must extend (3.1) from finite dimensional matrices to linear operators on Hilbert space. For densely defined, closed linear operators the generalization is that there is a unique solution, $X = A^\#$, to the system:

$$\begin{aligned}
XAX &= X \quad \text{on } \mathcal{D}(X), \\
AXA &= A \quad \text{on } \mathcal{D}(A), \\
AX &= P_{\overline{\mathcal{R}(A)}} \quad \text{on } \mathcal{D}(X), \\
XA &= P_{\mathcal{N}(A)^\perp} \quad \text{on } \mathcal{D}(A),
\end{aligned}
\tag{3.2}$$

where $\mathcal{D}(B)$ is the domain of the operator B , $\mathcal{R}(B)$ is the closure of the range of the operator B and $\mathcal{N}(B)^\perp$ is the orthogonal complement of the null space of B .¹³ The operators $P_{\overline{\mathcal{R}(A)}}$ and $P_{\mathcal{N}(A)^\perp}$ are orthogonal projectors on the subspaces $\mathcal{R}(A)$ and $\mathcal{N}(A)^\perp$ respectively. The second of these equations is redundant as it follows from the last.¹³

In the applications to the interior dynamics problem, the Π in equation (1.2) will arise as the orthogonal projector on the orthogonal complement of the null space of a given bounded operator. From the above considerations we see that the problem is reduced to computing $A^\#A$ for the appropriate operator A . For bounded A , $\mathcal{D}(A)$ is the full Hilbert space, \mathcal{H} , so there are no problems with domains. By the closed graph theorem¹⁴ bounded operators, A , with $\mathcal{D}(A) = \mathcal{H}$ are closed so the conditions of (3.2) are automatically met.

The only question remaining is whether we can actually construct this projector. An immediate corollary of Theorem 10 of Ref. 3 (pp. 354–5) is that if we define the iteration:

$$\begin{aligned}
X_0 &= \alpha A^\dagger A, \\
X_1 &= (1 - \alpha A^\dagger A) X_0 \\
X_N &= \alpha A^\dagger A + (1 - \alpha A^\dagger A) X_{N-1},
\end{aligned}
\tag{3.3}$$

where α is a real constant satisfying

$$0 < \alpha < 2/\|A\|^2 \tag{3.4}$$

then X_N converges strongly to $P_{\mathcal{N}(A)^\perp}$. This corollary follows from the quoted theorem if we take $y = Ax$ and observe that this means $y \in \text{Range}(A)$. If in addition A has closed range this convergence is uniform.¹⁵

From (3.3) we determine that the exact $P_{\mathcal{N}(A)^\perp}$ has the representation

$$P_{\mathcal{N}(A)^\perp} = \sum_{k=0}^{\infty} (1 - \alpha A^\dagger A)^k \alpha A^\dagger A, \tag{3.5}$$

which converges strongly. This gives a constructive algorithm for computing $P_{\mathcal{N}(A)^\perp}$ given a bounded operator A .

IV. H_1 AS A PROJECTED HAMILTONIAN

The strategy for choosing an important set of channels \mathcal{A} in the decomposition (2.7) is that the channels \mathcal{A} are strongly coupled to one another, and not strongly coupled to the residual channels. Physically two systems approach each other, collide, and come out with a substantial probability of being found in one of the channels \mathcal{A} . The collision occurs when all N -particles of the system are close together, and the time evolution of the collision is governed by the full Hamiltonian. This suggests that we use the freedom in choosing the connected effective N -body interaction, U , so that the time evolution of the system during the collision is generated by the full Hamiltonian for states strongly coupled to the asymptotic channels. The unitarity requirement, i.e., that U be

connected and Hermitian, poses some restrictions on how this can be done.

One choice of U that reflects these properties to a reasonable extent arises naturally in the formalism of I. We define the operator:

$$Y(\mathcal{A}) \equiv \sum_a \mathcal{C}_a P_a^+(\mathcal{A}). \tag{4.1}$$

This operator is clearly

bounded, since it is a finite sum of orthogonal projectors, and hence is closed.¹⁴ Thus the machinery concerning the Moore–Penrose operator inverse discussed in the previous section is applicable. In particular $Y^\#(\mathcal{A})$ exists as a possible unbounded operator and

$$\Pi(\mathcal{A}) = Y^\#(\mathcal{A}) Y(\mathcal{A}), \tag{4.2}$$

is the orthogonal projector on the orthogonal complement of the null space of $Y(\mathcal{A})$. In particular, it projects on the space of all vectors of the form $Y(\mathcal{A})|x\rangle$. Since $Y(\mathcal{A})$ has a piece of each $P_a^+(\mathcal{A})$ in its expansion, it is hoped that a large part of the range of the various P_a^+ 's will be contained in the range of $Y(\mathcal{A})$. The fact that the coefficients \mathcal{C}_a come in is to ensure the connectivity of U , which is required for unitarity.¹⁶

The projector $\Pi(\mathcal{A})$ has been constructed to have some very important properties regarding its connectivity structure. These properties follow from the connectivity structure of the $P_a^+(\mathcal{A})$'s. The relevant property is that if the interaction between the clusters of b are turned off then $P_a^+(\mathcal{A})$ becomes $P_{a \cap b}^+(\mathcal{A})$. Where $a \cap b$ is the partition obtained from a by separating the particles in the same cluster of a that become mutually noninteracting when the interactions between the clusters of b are turned off. This property follows from the integral representation of $P_a^+(\mathcal{A})$.¹⁰

From (4.1) it follows that as the clusters of b are separated beyond the range of the interactions $Y(\mathcal{A})$ becomes

$$Y(\mathcal{A}) \rightarrow \sum_a \mathcal{C}_a P_{a \cap b}^+(A). \tag{4.3}$$

We may evaluate this sum using (2.6) and the result

$$\delta_{a \cap b \supseteq c} = \delta_{a \supseteq c} \delta_{b \supseteq c}, \tag{4.4}$$

to obtain

$$\begin{aligned}
Y(\mathcal{A}) &\rightarrow \sum_{a,c,d,e} \delta_{e \supseteq a}^{-1} \delta_{a \supseteq c} \delta_{b \supseteq c} \delta_{c \supseteq d}^{-1} P_d^+(\mathcal{A}) \\
&= P_b^+(\mathcal{A}).
\end{aligned}
\tag{4.5}$$

If we use $P_a^+(\mathcal{A}) = P_a^+(\mathcal{A})^\dagger$; $Y(\mathcal{A}) = Y(\mathcal{A})^\dagger$ and insert the limiting form (4.5) in the representation (3.5) for $\Pi(\mathcal{A})$ we obtain for sufficiently small α

$$\begin{aligned}
\Pi(\mathcal{A}) &\rightarrow \sum_{k=0}^{\infty} (1 - \alpha P_b^+(\mathcal{A}))^k \alpha P_b^+(\mathcal{A}) \\
&= P_b^+(\mathcal{A}) \cdot \alpha \sum_{k=0}^{\infty} (1 - \alpha)^k = P_b^+(\mathcal{A}).
\end{aligned}
\tag{4.6}$$

This means, in the notation of Sec. II, that

$$\Pi_a(\mathcal{A}) = P_a^+(\mathcal{A}). \quad (4.7)$$

If we consider the projected Hamiltonian

$$H_1 = \Pi(\mathcal{A}) H \Pi(\mathcal{A}), \quad (4.8)$$

and use the internal-external decompositions

$$H = H_a + V^a, \\ \Pi(\mathcal{A}) = P_a^+(\mathcal{A}) + \Pi^a(\mathcal{A}),$$

it follows that H_1 has the structure

$$H_1 = P_a^+(\mathcal{A}) H_a P_a^+(\mathcal{A}) + (\text{terms with connectivities external to } a). \quad (4.9)$$

This means that

$$H_{1a} = P_a^+(\mathcal{A}) H_a P_a^+(\mathcal{A}). \quad (4.10)$$

Since $P_a^+(\mathcal{A})$ projects on an invariant subspace of H_a , it follows that

$$[H_a, P_a^+(\mathcal{A})]_- = 0,$$

and $P_a^+(\mathcal{A}) H_a P_a^+(\mathcal{A}) = H_a P_a^+(\mathcal{A})$. Equation (4.10) may be combined with (2.5) to yield the main result of this paper:

$$H_1 = \Pi(\mathcal{A}) H \Pi(\mathcal{A}) = \sum_a' \mathcal{C}_a H_a P_a^+(\mathcal{A}) + [H_1]_1, \quad (4.11)$$

which is the result promised in (2.7) if we identify the connected part of H_1 , $[H_1]_1$, with the N -body force $U_1[H_1]_1$, is the most simply expressed as

$$[H_1]_1 \equiv \Pi(\mathcal{A}) H \Pi(\mathcal{A}) - \sum_a' \mathcal{C}_a H_a P_a^+(\mathcal{A}), \quad (4.12)$$

and is clearly connected by (2.5). Equations (4.11) and (4.12) exhibit a proper choice of connected potential that puts the theory proposed in I in a projected Hamiltonian form.

We remark that if \mathcal{A} includes all scattering channels each $P_a^+(\mathcal{A})$ becomes the identity operator, $Y(\mathcal{A})$ becomes the identity because of the relation $\Sigma' \mathcal{C}_a = 1$,^{6,8} $\Pi(\mathcal{A})$ becomes the identity by uniqueness of the Moore-Penrose generalized inverse, and H_1 becomes full Hamiltonian by (4.8). Thus the full theory emerges if we include all channels in H_1 . We remark that if Q is a finite rank projector satisfying

$$\Pi(\mathcal{A}) Q = 0, \\ (\Pi(\mathcal{A}) + Q)^2 = (\Pi(\mathcal{A}) + Q), \\ Q = Q^\dagger \quad (4.13)$$

then

$$(Q + \Pi(\mathcal{A})) H (Q + \Pi(\mathcal{A})), \quad (4.14)$$

defines a projected Hamiltonian approximation that also satisfies (2.7). Because Q is finite rank, the corresponding U is still connected and the unitarity considerations of I remain unchanged. As Q begins to fill out the null space of $\Pi(\mathcal{A})$ one begins to build up singularities associated with the eliminated channels. This will inevitably lead to continuum poles in the resolvent of H_1 . Since these poles contain some effects of the eliminated channels they are not necessarily undesirable. The existence of continuum poles associated with (4.11) cannot be ruled out either.

V. CONSTRUCTION OF THE CONNECTED POTENTIAL

All of the formalism introduced in the previous section is of little value if we cannot construct U or equivalently $\Pi(\mathcal{A})$ in practical applications. Fortunately there are methods for constructing the Moore-Penrose generalized inverse in practice. The obvious approach is to use the strongly convergent iteration associated with the series

$$\Pi(\mathcal{A}) = \sum_{k=0}^{\infty} (1 - \alpha Y(\mathcal{A})^2)^k \cdot \alpha Y(\mathcal{A})^2, \quad (5.1)$$

where we have used $Y(\mathcal{A}) = Y(\mathcal{A})^\dagger$ in (3.5). To utilize this expansion one must be able to find a suitable α . Since in any numerical calculation the operator $Y(\mathcal{A})^2$ will be put on a mesh and (5.1) will be reduced to a matrix equation it turns out that by the Gershgorin theorem¹⁹

$$||[Y(\mathcal{A})^2]^2|| \leq \max_{i=1, N} \sum_{j=1}^N |Y(\mathcal{A})_{ij}^2| \equiv \gamma, \quad (5.2)$$

where N is the dimension of the matrix $Y(\mathcal{A})$. In numerical applications it is enough to choose α so that

$$0 < \alpha < 2/\gamma. \quad (5.3)$$

The best choice of α for these iterations has been considered elsewhere²⁰ and is

$$\alpha_{\text{optimal}} = \frac{2}{\lambda_{\text{max}} + \lambda_{\text{min}}} \quad (5.4)$$

where λ_{max} and λ_{min} are the largest and smallest nonzero eigenvalues of the matrix $Y^2(\mathcal{A})$. A lower bound on the number of iterations required for convergence has been estimated to be²¹

$$N_{\text{min}} = 2 \log_2 (\lambda_{\text{max}}/\lambda_{\text{min}}). \quad (5.5)$$

Other higher order iterative methods have been discussed elsewhere.³

The size of the matrix that must be inverted will clearly depend on the number of continuous degrees of freedom in the operators $P_a^+(\mathcal{A})$. Since this is the same constraint that fixes the size of the matrices for the scattering integral equations, we expect that the construction of $\Pi(\mathcal{A})$ will be numerically tractable in exactly the same situations that the approximate scattering equations discussed in I are tractable. In particular, this will be when (\mathcal{A}) contains only two- and three-body asymptotic channels.

In general the iterative methods discussed above may not be the most efficient way to construct $\Pi(\mathcal{A})$. Once $Y(\mathcal{A})$ has been put on a mesh there exist many methods for computing its Moore-Penrose generalized inverse. These methods are discussed extensively in Refs. 3,4 and the references cited therein.

VI. APPLICATION OF MOORE-PENROSE TECHNIQUES TO DYNAMICS

The techniques used to solve our nonorthogonality problem in the last two sections have several desirable features. Because of this it is natural to ask if the scattering problem itself can be better handled in this framework. It turns out that for Faddeev-like formalisms the situation is actually better than discussed previously. This is because $(1 - K)$ has closed range for compact operators.²²

To consider the applications of the Moore–Penrose techniques we assume that we have a scattering integral equation of the form

$$X = D + KX, \quad (6.1)$$

where X is unknown and K is compact on the N -body Hilbert space. The solution is formally given by

$$X = (1 - K)^{-1}D. \quad (6.2)$$

We will make the assumption that $(1 - K)^{-1}$ exists in all that follows so that $(1 - K)^{-1}$ coincides with the unique Moore–Penrose generalized inverse of $(1 - K)$. The problem is then reduced to finding this generalized inverse. The solution is given by the iteration²³

$$\begin{aligned} W_0 &= \alpha(1 - K^\dagger), \\ W_N &= \alpha(1 - K^\dagger) + (1 - \alpha(1 - K^\dagger)(1 - K))W_{N-1}, \end{aligned} \quad (6.3)$$

for

$$0 < \alpha < 2/|(1 - K)|^2. \quad (6.4)$$

By a theorem of Petryshyn¹⁵ this iteration converges uniformly in the closed range case. Since we have also assumed $(1 - K)$ is invertible, its range is the entire Hilbert space. In this case we may consider the integral equation

$$X = \alpha(1 - K^\dagger)(D + 1 - \alpha(1 - K^\dagger)(1 - K))X, \quad (6.5)$$

which is derived from the iteration (6.4). This has a unique solution because

$$(1 - 1 + \alpha(1 - K^\dagger)(1 - K)), \quad (6.6)$$

has the inverse

$$(1/\alpha)(1 - K)^{-1}(1 - K^\dagger)^{-1}. \quad (6.7)$$

The unique solution to (6.5) is clearly $(1 - K)^{-1}D$. By reformulating the scattering problem in this way we have replaced a compact noncontractive kernel equation by a noncompact contractive kernel equation. In addition, the new kernel is Hermitian while the kernel of (6.1) is not. Whether (6.5) is better starting point than (6.1) for treating the scattering problem is an open question.

VII. CONCLUSION AND DISCUSSION

In this paper we have given a more complete discussion of the theory presented in I. Our attention focused primarily on the short ranged degrees of freedom involved in making the decomposition (2.7). We have shown that it is possible to construct an effective N -body force, U , that puts the H_I of (2.7) in the form (1.2). The operator $\Pi(\mathcal{A})$ is the orthogonal projector on the orthogonal complement to the null space of the operator $Y(\mathcal{A})$ defined by (4.1). From the structure of $Y(\mathcal{A})$ we hope that for an appropriate choice of \mathcal{A} this subspace will be sufficiently large to treat all of the important couplings. As discussed in Sec. IV this subspace can be enlarged by finite dimensional extensions of $\Pi(\mathcal{A})$. The appearance of the \mathcal{C}_a coefficients in (4.1) turns out to be crucial for maintaining the unitarity constraint for arbitrary \mathcal{A} 's. This result of Sec. IV shows quite conclusively that the approximation discussed in I, if treated properly, have no overcounting. One may have anticipated some overcounting by

the appearance of the coefficient $\mathcal{C}_a = (-)^a(n_a - 1)!$ in (2.7). The reason that there is no overcounting is easily seen in equation (2.4) and (2.5) from which it follows that (2.7) is an alternate way of writing a cluster expansion. The proposed treatment of the interior region requires the solution of a nonorthogonality equation to obtain the projected form (4.11). If one chooses U by another method one cannot be assured that there is no overcounting in the interior region.

The appearance of a nonorthogonality problem in our truncated Hamiltonian indicates that one cannot necessarily avoid these problems by truncating Faddeev-type theories, even though these problems do not arise in the untruncated forms.

The methods used to deal with the nonorthogonality problem in this paper involve new techniques which have not been previously used in the few-body physics. These methods lead to alternate formulations of the few-body integral equations where compact kernels are replaced by noncompact contractive kernels. This allows us to treat the few-body problem by perturbation theory, even for the case of strong potentials.

¹W. N. Polyzou and E. F. Redish, *Ann. Phys. (N.Y.)* **119**, 1 (1979).

²M. Gell-Mann and M. L. Goldberger, *Phys. Rev.* **91**, 398 (1953).

³A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications* (Wiley, New York 1974).

⁴*Generalized Inverses and Applications*, edited by M. Z. Nashed (Academic, New York, 1976).

⁵C. W. Groetsch, *Generalized Inverses of Linear Operators, Representations and Approximation* (Dekker, New York, 1977).

⁶W. N. Polyzou, *Math. Phys.* **21**, 506 (1980).

⁷C. Birkoff, 1980 *Lattice Theory* (AMS Colloquium Publications, 1949), Vol. 25.

⁸K. L. Kowalski, W. N. Polyzou, and E. F. Redish preprint.

⁹J. E. Graves and M. E. Watkins, *Combinatorics with Emphasis on the Theory of Graphs* (Springer-Verlag, Berlin, 1977).

¹⁰W. N. Polyzou, Ph.D. thesis, University of Maryland, TR# 80-018, ORO 5126-82.

¹¹In this paper we use the term Moore–Penrose generalized inverse to denote the orthogonal operator generalized inverse defined on p. 62 of Ref. 4.

¹²R. Penrose, *Proc. Cambridge Phil. Soc.* **51**, 406 (1955).

¹³Ref. 4, Theorem (5.7), p. 62.

¹⁴M. Reed and B. Simon, *Functional Analysis* (Academic, New York, 1972).

¹⁵W. V. Petryshyn, *J. Math. Anal. Appl.* **18**, 417 (1967).

¹⁶The discussion in Ref. 10 actually corresponds to spectral representations of partition Hamiltonians. An analogous argument applies to the projectors.

¹⁷This result is discussed in Ref. 8 and is an elementary consequence of the lattice properties; i.e. $a \supseteq c$ and $b \supseteq c \rightarrow a \cap b \supseteq c$; also $a \cap b \supseteq c \rightarrow a \supseteq c$ and $b \supseteq c$, as one naively expects.

¹⁸The restriction on the a sum ($n_a \equiv \neq 1$) is automatically implied by the $\delta_{b \supseteq c}^{-1}$.

¹⁹A. S. Householder, *The Theory of Matrices in Numerical Analysis* (Blaisdell, Waltham, 1964).

²⁰A. Ben-Israel and D. Cohen, *SIAM J. Num. Anal.* **3**, 410 (1966).

²¹T. Söderström and G. W. Stewart, *SIAM J. Num. Anal.* **11**, 61 (1974).

²²Ref. 3, p. 326.

²³Ref. 3, Theorem 10, pp. 354–5.

Linkages in general relativity ^{a)}

Robert Geroch

University of Chicago, 933 E. 56th Street, Chicago, Illinois 60637

Jeffrey Winicour

University of Pittsburgh, Pittsburgh, Pennsylvania 15260

(Received 15 July 1980; accepted for publication 11 November 1980)

For an asymptotically flat space-time in general relativity there exist certain integrals, called linkages, over cross sections of null infinity, which represent the energy, momentum, or angular momentum of the system. A new formulation of the linkages is introduced and applied. It is shown that there exists a flux, representing the contribution of gravitational and matter radiation to the linkage. A uniqueness conjecture for the linkages is formulated. The ambiguities due to the possible presence of supertranslations in asymptotic rotations are studied using the behavior of the linkages under first-order perturbations in the metric. While in certain situations these ambiguities disappear in the first-order treatment, an example is given which suggests that they are an essential feature of general relativity and its asymptotic structure.

PACS numbers: 04.20. — q

1. INTRODUCTION

There is available, for space-times in general relativity, a definition of asymptotic flatness at null infinity.¹⁻⁴ Physically, asymptotically flat space-times represent isolated systems. This definition has turned out to be a particularly fruitful one. For example, it leads naturally to the peeling property of asymptotic fields¹ and to the Bondi-Metzner-Sachs (BMS) asymptotic symmetry group^{5,6}; and it underlies much of the discussion of such topics as black holes,⁷ quantum fields in curved space-times,⁸ and cosmic censorship.⁹

In flat space-time, symmetries give rise to conserved integrals. One might expect, therefore, that, in asymptotically flat space-times, the asymptotic symmetries would give rise to similar integrals. This turns out to be the case. For the case of asymptotic translations, these integrals, taken over cross sections of null infinity, yield the Bondi energy-momentum.^{5,10,11} For a stationary space-time, the Bondi energy reduces to the usual energy. Furthermore, there is a formula giving the change in the Bondi energy-momentum for two cross sections in terms of the total gravitational energy-momentum radiated during the intervening time. For other asymptotic symmetries the situation is somewhat more complicated. There does exist a formula for an asymptotic integral, the linkage, over any cross section for any asymptotic symmetry in the asymptotically flat space-time.^{12,13} These linkages reduce to the Bondi energy-momentum for translational symmetries, and to the usual angular momentum for axi-symmetric space-times (in which we know what "angular momentum" means). Finding any reasonable generalization of the Bondi energy-momentum is a delicate business, for one must achieve two distinct types of gauge-invariance: one under conformal transformations, and one under alternative representations of a single asymptotic symmetry. The complications inherent in achieving the second type of

gauge invariance, in particular, make it difficult to settle such questions as the existence of a flux, an expression, in terms of the asymptotic gravitational radiation, which gives the change in the linkage from one cross section to the next. Further, even aside from these technical difficulties, there arise problems of physical interpretation.^{6,14,15} The cause of these is the presence of supertranslations: an infinite-dimensional collection of asymptotic symmetries, "distorted translations," which are extraneous to—and so have no physical interpretation in—special relativity. Since "a translation without any supertranslation" makes sense, the supertranslations can be ignored in dealing with Bondi energy-momentum. But "a rotation without any supertranslation" does not make sense. The problem, then, is to identify the "pure rotational symmetries" for which the linkage is to be computed. The severity of this problem is illustrated by the following fact: even in Minkowski space-time, the linkage of certain supertranslations is nonzero. It is not even clear whether these various difficulties—technical and interpretive—are inherent in the structure of general relativity itself, or are merely features of the particular linkage integral which has been written down. There is, for example, an alternative generalization—but just to the supertranslations—of Bondi energy-momentum, which avoids the problem of a nonzero linkage in Minkowski space-time.³ Might there be an alternative, and more easily interpreted, generalization to all the asymptotic symmetries?

We here obtain a number of properties of the linkages, and discuss their impact on the issues above.

In Sec. 2, we introduce an alternative formulation of the linkage integral. This new version achieves gauge-invariance in a simpler and more natural way. It is well-suited to the treatment of such issues as the dependence of the linkage on the cross section and its behavior under perturbations. The price paid is that now the integrand itself, as opposed to the numerical value of the integral, becomes gauge-dependent.

A striking open question involving the linkage is whether or not there exists a flux, which describes the rate of

^{a)}Supported in part by the NSF, under Contract Nos. PHY 78-24275 and PHY 78-13926-A01, respectively.

change of the linkage integral due to radiation. In Sec. 3 we answer this question: Such a flux exists. This flux is then generalized to include the presence of matter, such as electromagnetic fields, which can radiate to infinity. It is found, further, that there is no single, natural decomposition of the total flux, in the presence of matter, into a “matter part” and a “gravitational part”; the two seem unavoidably mixed. This feature reflects the well-known “factor of two anomaly.”¹⁵ The conserved integrals over the matter for symmetries in flat space–time and the Komar¹⁶ gravitational integrals for symmetries in curved space–time agree for rotations, but differ by a factor of two for time translations. (This anomaly is not due to some convention in a choice of factor, for there are no “pure rotations” in the Poincaré group.) Finally, it is suggested in Sec. 3 that the linkages may be the only asymptotic integrals satisfying a list of physical requirements.

Section 4 deals with linear perturbations, which preserve asymptotic flatness, of the gravitational fields.¹⁷ The purpose is to understand better the ambiguities, inherent in the linkages, associated with hiding an extra supertranslation in a rotational symmetry. It turns out that one can shift this ambiguity between the gauge in which the perturbation is expressed, the choice of asymptotic symmetry, and the choice of cross section over which the integral is performed. Under certain restrictions—on the background space–time, the perturbation, the asymptotic symmetry, or the cross-section—these ambiguities can be made to disappear. Whenever the answer is unambiguous, the linkage seems to yield what one would expect physically. Finally, an example is given which suggests that the ambiguities associated with supertranslations are an essential feature of general relativity and its asymptotic structure.

2. LINKAGES

Let \tilde{M} , \tilde{g}_{ab} be a space–time which is asymptotically flat at null infinity. That is, we have a manifold with boundary, M , consisting of \tilde{M} with boundary attached, a function Ω (the conformal factor) on M with $\Omega = 0$ on the boundary, and a metric $g_{ab} = \Omega^2 \tilde{g}_{ab}$ on M (called the unphysical metric, as distinguished from the physical metric \tilde{g}_{ab} on \tilde{M}). Throughout, let I denote the boundary at either future or past null infinity. Asymptotic flatness requires that this I have topology $S^2 \times \mathbb{R}$, with the \mathbb{R} 's the null geodesics tangent to the vector field $n^a = g^{am} \nabla_m \Omega$, and that n^a vanish nowhere on I . Further, let Ω be chosen to be a Bondi conformal factor, i.e., to satisfy $\nabla_a \nabla_b \Omega = 0$ on I . In this section, we shall deal with the exterior vacuum case, for which Einstein's equation takes the form

$$\Omega^2 (R_{ab} - \frac{1}{2} R g_{ab}) = n^m n_m g_{ab} - 2\Omega \nabla_a n_b. \quad (1)$$

Consider next the Bondi–Metzner–Sachs (BMS) asymptotic symmetry group. Its generators are given by vector fields ξ^a on M such that the right side of

$$\Omega^2 \mathcal{L}_\xi \tilde{g}_{ab} = \mathcal{L}_\xi g_{ab} - 2\Omega^{-1} \xi^m n_m g_{ab}, \quad (2)$$

is smooth and vanishes on I . It follows immediately that $\xi^a n_a$ vanishes on I , and so that $\xi^a n_a = \Omega K$ for some smooth function K on M . The vanishing of (2) on I now takes the form

$$\nabla_{(a} \xi_{b)} = K g_{ab} + \Omega X_{ab}, \quad (3)$$

where X_{ab} is some smooth tensor field on M . Since its vanishing is precisely Killing's equation in the physical space–time, we may interpret X_{ab} as a measure of the extent to which the BMS generator fails to arise from a physical symmetry. Note in particular that every Killing field in the physical space–time gives rise to a BMS generator. This X_{ab} is automatically transverse to n^a , in the sense that $X_a = \Omega^{-1} X_{am} n^m$ must remain finite at I . To see this, set $X = X^m{}_m$, and expand the identity $(\mathcal{L}_\xi \mathcal{L}_n - \mathcal{L}_n \mathcal{L}_\xi - \mathcal{L}_{\xi^n}) g_{ab} = 0$, using (1) and (3), to obtain

$$\begin{aligned} -\nabla_a \nabla_b K + 4n_{(a} X_{b)} + 2\Omega \nabla_{(a} X_{b)} - n^m X_m g_{ab} \\ - \frac{1}{2} \mathcal{L}_\xi (R_{ab} - \frac{1}{2} R g_{ab}) - \mathcal{L}_n X_{ab} = 0. \end{aligned} \quad (4)$$

We call two BMS generators *equivalent* if they are equal on I .¹⁸ The BMS generators form a Lie algebra—where the bracket is the Lie bracket—as do the equivalence classes. Those generators which, on I , are multiples of n^a are called supertranslations.

The present description of asymptotic symmetries is subject to two distinct types of gauge freedom.

The first corresponds to the choice between equivalent BMS generators. By definition, the difference between two equivalent generators must vanish on I , i.e., it must be of the form Ωw^a for some vector field w^a on M . But, substituting into (3), we find that Ωw^a is a BMS generator if and only if w^a vanishes on I , i.e., if and only if $\Omega w^a = \Omega^2 u^a$ for some smooth field u^a on M . Thus, the gauge freedom to choose between equivalent generators is given by $\delta \xi^a = \Omega^2 u^a$. Under this change, we have from (3)

$$\begin{aligned} \delta K &= \Omega u^a n_a, \\ \delta X_{ab} &= 2n_{(a} u_{b)} - n^m u_m g_{ab} + \Omega \nabla_{(a} u_{b)}. \end{aligned} \quad (5)$$

The second type of gauge freedom corresponds to a change in the conformal factor. Setting $\Omega' = \omega \Omega$, for ω a smooth function on M , we obtain, since the physical metric must remain fixed, $g'_{ab} = \omega^2 g_{ab}$. Invertibility of g'_{ab} requires that ω vanish nowhere, while the condition that Ω' also be a Bondi conformal factor requires that $n^a \nabla_a \omega = 0$ at points of I . This, then, is the conformal gauge freedom. Now let ξ^a be a BMS generator with respect to g_{ab} . Then, from (3), $\xi'^a = \xi^a$ is again a generator with respect to g'_{ab} . Equation (3) now yields

$$\begin{aligned} K' &= K + \omega^{-1} \xi^m \nabla_m \omega, \\ X'_{ab} &= \omega X_{ab}. \end{aligned} \quad (6)$$

To summarize, we have the generators of asymptotic symmetries, a tensor field X_{ab} , for each generator, which measures the extent to which that generator fails to represent a physical symmetry, and two types of gauge freedom.

We now wish to introduce “asymptotic quantities” associated with these asymptotic symmetries. That is, we seek real-valued, multilinear, gauge-invariant functions on BMS generators, possibly also depending on cross sections of I . The value of the function represents some property of the system: the cross section, the “time” at which that property is determined. Gauge-invariance guarantees that we are

dealing with a physical property, and multilinearity guarantees that this property fits into a representation¹⁹ of the BMS group.

Let us, in order to get an idea of the possibilities, consider first the case of a BMS generator which arises from a Killing field in the physical space-time. For ξ^a a Killing field, the Komar integral, $\int_S \bar{\epsilon}_{abmn} (\bar{\nabla}^m \xi^n) dS^{ab}$, where S is any topological 2-sphere in the physical space-time surrounding the sources, is independent of the choice of sphere. This integral represents the energy, momentum, angular momentum, etc., of the system, depending on what type of symmetry is represented by ξ^a . Rewriting in terms of unphysical fields, with $\xi^a = \bar{\xi}^a$, we have

$$Q = \int_S \epsilon_{abmn} \nabla^m (\Omega^{-2} \bar{\xi}^n) dS^{ab}. \quad (7)$$

This suggests the following generalization of the Komar integral to the case in which $\bar{\xi}^a$ does not arise from a physical symmetry: use (7), where $\bar{\xi}^a$ is any BMS generator, and S any cross section of I .

Unfortunately, there appear to be two difficulties with such a definition. The first is that, because of the presence of inverse powers of Ω , the integrand in (7) is, in general, infinite at I . But this is easily remedied. We claim: for $\bar{\xi}^a$ any BMS generator, the limit of the right side of (7), as S , a 2-sphere in the physical space-time \bar{M} , approaches a fixed cross section of I , always exists, and furthermore is independent of the details of the limit. To see this, first rewrite (7) as the integral over some fixed 2-sphere S_0 in the physical space-time (which integral is always finite) plus the integral of the curl of the integrand over a three-dimensional surface joining S_0 to S . It suffices, then, to show that this curl remains finite at I . From (3) and the identity

$$\nabla^m \nabla_{[a} v_{m]} = R_{am} v^m + \nabla_a (\nabla_m v^m) - \nabla^m \nabla_{[a} v_{m]}, \quad \text{we obtain}$$

$$\nabla^m \nabla_{[a} (\Omega^{-2} \bar{\xi}_{m]}^n) = \Omega^{-1} (-\nabla^m X_{am} + \nabla_a X + 3X_a). \quad (8)$$

But the field in brackets on the right in (8) vanishes on I , as one sees by expanding the term $\mathcal{L}_{\bar{\xi}}(R_{ab} - 1/6 R g_{ab})$ in (4) in terms of $\mathcal{L}_{\bar{\xi}} g_{ab}$ and then using (3)—a straightforward but tedious calculation. So, the curl is indeed finite. We conclude that (7), regarded as a limit at a cross section S of I , always makes sense.

The second—and more serious—difficulty involves gauge dependence. First note that the Q given by (7) is indeed unchanged under further conformal transformations. The problem is with changes in choice of generator within an equivalence class. In fact, under the change $\delta \bar{\xi}^a = \Omega^2 u^a$ in the generator, the integrand in (7) changes by $\epsilon_{abmn} \nabla^m u^n$. Clearly, by a suitable choice of u^a the value of Q itself can change. Thus, gauge invariance does not hold. A naive generalization of the Komar integral, from genuine symmetries to asymptotic symmetries, does not yield physically acceptable asymptotic properties of systems.

A method for avoiding this difficulty—for obtaining gauge-independent functions of BMS generators—would be to use some combination of the following two strategies: (i) Add to the integrand in (7) some additional expression linear in the generator, this expression so chosen that the resulting integral continues to reduce to the Komar integral for a real

symmetry; or (ii) impose on the generator some additional gauge conditions, so chosen that they can always be achieved via (5).

The linkages^{12,13,15} provide one example of the employment of these strategies. The definition, in this language, is the following. Fix a cross section S of I . Let N be the outgoing null surface in M which meets I at S , and let l^a be a null, geodesic vector field which is tangent to N at points of N . We now demand, as the gauge condition, that, at points of N ,

$$(X_{am} - \frac{1}{2} X g_{am}) l^m = 0, \quad (9)$$

or, in terms of physical variables, that

$$(\bar{\nabla}_{[a} \bar{\xi}_{b]}^m - \frac{1}{2} \bar{g}_{ab} \bar{\nabla}_m \bar{\xi}^m) l^b = 0. \quad \text{Let the asymptotic integral be}$$

$$L = \int_S [\epsilon_{abmn} \nabla^m (\Omega^{-2} \bar{\xi}^n) + \Omega^{-1} X \epsilon_{ab}] dS^{ab}, \quad (10)$$

where $\epsilon_{ab} = \epsilon_{abcd} n^c l^d (n_m l^m)^{-1}$ is the induced surface element on the cross section S . In terms of the physical variables, this integrand is $\bar{\epsilon}_{abmn} \bar{\nabla}^m \bar{\xi}^n + \bar{\epsilon}_{ab} (\bar{\nabla}_m \bar{\xi}^m)$. The integral (10) is to be understood in the sense of a limit at the cross section S . We first note that this limit always exists. For the first term on the right in (10), existence of the limit has already been shown. For the second, contract (9) with n^a to obtain

$$X = 2\Omega X_a l^a (n_m l^m)^{-1}. \quad (11)$$

Thus, the integrand for this second term, $\Omega^{-1} X$, remains finite at I . So, the integral (10), under (9), makes sense.

We next consider gauge dependence. It is immediate from (6) that the linkage integral, (10), is invariant under changes in the conformal factor. To show invariance under passage to an equivalent generator is more difficult. We first obtain a preliminary result. Consider a change to an equivalent generator, $\delta \bar{\xi}^a = \Omega^2 u^a$, such that (9) is preserved, i.e., by (5), consider any vector field u^a satisfying

$$[2n_{[a} u_{b]} + \Omega \nabla_{[a} u_{b]} - \frac{1}{2} \Omega g_{ab} \nabla_m u^m] l^b = 0, \quad (12)$$

on N . We show that this implies $u^a = 0$ on N near the cross section S . First, contract (12) with l^a to obtain

$\mathcal{L}_l (\Omega^2 l^b u_b) = 0$. But this equation, together with the fact that $\Omega^2 l^b u_b$ vanishes on S , implies that $l^b u_b$ vanishes on N . Next, antisymmetrize (12) with l_c , using the vanishing of $l^b u_b$ on N , to obtain $\mathcal{L}_l (\Omega^2 l_{[c} u_{a]}) = 0$ on N . But this equation, together with the fact that $\Omega^2 l_{[c} u_{a]}$ vanishes on S , implies that u^a is a multiple of l^a on N . Finally, substituting that $u^a = \alpha l^a$ on N into (12), we obtain $\alpha (2l^m n_m - \Omega \nabla_m l^m) = 0$ on N . It follows, since the field in parentheses is nonzero on N near S , that $\alpha = 0$ on N near S . We conclude, then, that two equivalent generators, both satisfying (9), must coincide on N near S . [We remark that essentially this same argument also shows that the gauge condition (9) can always be achieved.] The proof that linkages are gauge-invariant now proceeds as follows. Since the limit implicit in (10) can be taken in any way we choose, consider that through cross sections of N . For such cross sections the integrand, as one verifies, depends only on the value of the generator $\bar{\xi}^a$ on N and its derivative within N . But since, as we have just seen, the gauge condition (9) fixes $\bar{\xi}^a$ uniquely on N , it fixes these integrals over cross sections of N uniquely, and so fixes the

linkage integral (10) uniquely.

Thus, the linkage, given by the integral (10) under the gauge condition (9), yields, for a given cross section S of I , a gauge-invariant linear function on BMS generators. It reduces, in the case of a physical symmetry, to the Komar integral.

Our basic result is an alternative—and, for many applications, a more convenient—definition of the linkage. Let the gauge condition be, instead of (9), simply

$$X = 0, \quad (13)$$

and let the asymptotic integral be, instead of (10),

$$\hat{L} = \int_S \epsilon_{abmn} \nabla^m (\Omega^{-2} \xi^n) dS^{ab}, \quad (14)$$

i.e., what (10) reduces to with $X = 0$.

We prove the equivalence of (13), (14) and (9), (10). Fix a cross section S of I , a corresponding field l^a , and a BMS generator ξ^a subject to the gauge condition (9). Let X_{ab} refer to this ξ^a . Now consider an equivalent generator, $\xi^a + \Omega^2 u^a$, so chosen to satisfy the gauge condition (13). That is, by (5), u^a must satisfy

$$X = 2u^m n_m - \Omega \nabla_m u^m, \quad (15)$$

everywhere. We shall show that, under these conditions, the linkage integral of ξ^a and the integral (14) of $\xi^a + \Omega^2 u^a$ coincide, i.e., that the right side of

$$\hat{L}(\xi + \Omega^2 u) - L(\xi) = \int_S (\epsilon_{abmn} \nabla^m u^n - \Omega^{-1} X \epsilon_{ab}) dS^{ab}, \quad (16)$$

vanishes. Setting $q_{ab} = g_{ab} - 2(l^m n_m)^{-1} l_a n_b$, we have

$$\begin{aligned} & 2(l^m n_m)^{-1} l_a n_b \nabla^a u^b \\ &= q_{ab} \nabla^a u^b + 2(l^m n_m)^{-1} l_a n_b \nabla^a u^b - \nabla_a u^a \\ &= q_{ab} \nabla^a u^b + (l^m n_m)^{-1} l^a \nabla_a X, \end{aligned} \quad (17)$$

on I —the first equality is an identity, and the second uses (15) and the fact that $\nabla_a n_b = 0$ on I . Now substitute (17) into the first term in the integrand on the right in (16). The term “ $q_{ab} \nabla^a u^b$,” as an intrinsic divergence in S , integrates to zero by Gauss’s law, and so we are left with

$$\begin{aligned} & \hat{L}(\xi + \Omega^2 u) - L(\xi) \\ &= \int_S [(l^m n_m)^{-1} l^a \nabla_a X - \Omega^{-1} X] \epsilon_{ab} dS^{ab}. \end{aligned} \quad (18)$$

But (9) implies $X = 0$ on I , and so that the integrand in (18) vanishes on I .²⁰

We conclude, then, that these two formulations of the linkage integrals are equivalent. In the original version, (9), (10), the gauge condition removes essentially all gauge freedom, once a cross section is selected. The result is that not only the linkage integral, but even the integrand itself, is gauge invariant. Even the “local (on S) asymptotic contribution” to the integral makes physical sense. Furthermore, both the gauge condition and the integral itself are intrinsic to a null surface in the space-time. This version is thus well-suited, e.g., to the characteristic initial-value formulation. But there are also some disadvantages. First, the collection of BMS generators satisfying the gauge condition (9), for a

fixed cross section, does not form a Lie algebra, for the Lie bracket of two need not be another. [Of course, given two generators satisfying (9), there always exists a third, equivalent to their Lie bracket and satisfying (9).] Consequently, it can become necessary to repeatedly solve (9) in passing to equivalent generators. More serious is the delicate dependence of the generator on the choice of cross section. Consider, for example, a slicing of I by cross sections, and a generator ξ^a satisfying the gauge condition (9) for these cross sections. Now choose a new family of cross sections. Then, while there will of course be a BMS generator equivalent to ξ^a and satisfying (9) for these new cross sections, this new generator will in general be different from ξ^a . Thus, this version is not well-suited to such questions as the dependence of the linkage integral on cross section. The present alternative version, (13), (14), has quite different strengths and weaknesses. The gauge condition (13) makes no reference to any cross-section; it fixes, once and for all, a class of allowed generators. These generators do form a Lie algebra. However, this gauge condition does permit some limited gauge freedom. While the integral (14) is invariant under passage to an equivalent generator subject to (13), the integrand itself is not. In this version, there is no physically meaningful “local asymptotic contribution” to the integral.

3. FLUX

Consider any asymptotic integral, $A(S)$, which assigns, for fixed BMS generator, a number to each cross section S . Is there an “asymptotic flux” for this integral, which determines its dependence on cross section? In more detail, does there exist a function F on I , independent of any cross sections, such that the difference, $A(S_2) - A(S_1)$, for any two cross sections, is precisely the integral of F over the region on I between the cross sections? Clearly, such an F is unique if it exists, and F alone determines the asymptotic integral uniquely up to an overall additive constant. But such a flux need not necessarily exist. As a simple example, consider the Euclidean cylinder, $S^1 \times \mathbb{R}$, let “cross sections” be curves which go once around the cylinder, and let the integral be the length of the curve. There exists in this example no function F on the cylinder such that the difference in length of two cross sections is the integral of F between them.

Do the linkages possess an asymptotic flux? From the original formulation, (9) and (10), the answer is by no means clear, for l^a in the gauge-condition (9) and ϵ_{ab} in the integral (10) both involve explicitly the choice of cross section. Consequently, the difference between (10) for two cross sections might appear to depend on those cross sections in a way too delicate to permit an asymptotic flux. In fact, there is a flux in this case.

Proof: Consider the alternative version (13) and (14). Here, the cross section enters only via the region of integration in (14). Indeed, not only must a flux exist, but it must be precisely the curl of the integrand, at I :

$$\begin{aligned} F &= n^a \nabla^m \nabla_{[a} (\Omega^{-2} \xi_{m]}), \\ &= n^a \Omega^{-1} [-\nabla^m X_{am} + \nabla_a X + 3X_a], \\ &= -\nabla^a \nabla^b X_{ab} + \nabla^2 X + 3\nabla^a X_a. \end{aligned} \quad (19)$$

In the first step, we have taken the dual and a contraction with n^a , to obtain a scalar flux; in the second, we have used (8); in the third, we have used the divergence of (8). A more convenient expression for the flux, obtained by virtue of the gauge condition $X = 0$, is

$$F = -\nabla^a \nabla^b X_{ab} + 3\nabla^a X_a + (3/4)\nabla^2 X + (1/24)RX. \quad (20)$$

This, then, is the general formula for the asymptotic flux for the linkages. It of course has the proper weight under changes of Bondi conformal factor. Furthermore, the expression (20), as a consequence of the changes from (19), is completely gauge-invariant, i.e., it has the same value for all equivalent generators, whether or not they satisfy $X = 0$. Note that the flux possesses all the advantages—and none of the disadvantages—of both versions of the linkage integral. It is locally defined, requires no gauge conditions, and makes no reference to cross sections.

As an example, let us consider the case of the supertanslation. On I , the BMS generator must be of the form τn^a , where τ is a scalar field on I satisfying $n^a \nabla_a \tau = 0$ there. Choose any extension of τ to a scalar field on M , and consider the vector field τn^a , defined everywhere. This field is not in general a BMS generator, for it need not satisfy (3). What is a solution of (3) is the combination

$$\xi^a = \tau n^a - \Omega \nabla^a \tau. \quad (21)$$

We compute the flux for this generator. Define σ , which depends on our generator, by $\Omega \sigma = n^a \nabla_a \tau$, and β , which does not, by $\Omega^2 \beta = n^a n_a$. [Finiteness of β , for a Bondi conformal factor, follows from (1).] The field X_{ab} and K [Eq. (3)] for this ξ^a follow immediately:

$$X_{ab} = -\frac{1}{2}\beta \tau g_{ab} + \sigma g_{ab} - \frac{1}{2}\tau(R_{ab} - \frac{1}{6}Rg_{ab}) - \nabla_a \nabla_b \tau, \quad (22)$$

Note that the generator (21) does not in general satisfy the gauge condition $X = 0$. Thus, if we wished to compute the linkage via (14), it would be necessary to add to the right side of (21) a term $\Omega^2 u^a$, so adjusted to achieve this gauge condition. The flux (20), however, can be computed in any gauge. Substituting (22) into (20), we obtain

$$F = \frac{1}{4}\tau^{-1}(\nabla^2 - 2\Omega^{-1}n^a \nabla_a)H - 2\tau^{-1}X^{ab}X_{ab} + \tau^{-1}X^2, \quad (23)$$

where we have set

$$H = \tau \nabla^2 \tau + (\frac{1}{6}R + \beta)\tau^2 - (\nabla^a \tau)(\nabla_a \tau) - 2\sigma\tau. \quad (24)$$

Note that Eq. (23) for the flux makes manifest neither its finiteness nor its linearity in the generator.

We now specialize still further, to the case in which ξ^a is actually a translation. In terms of τ , this means that τ on I is a linear combination of $l = 0$ and $l = 1$ spherical harmonics; in terms of H , it means that H is constant on I . In the case of a translation, X_{ab} reduces to essentially the news $X_{ab} - \frac{1}{2}Xh_{ab} = \tau N_{ab}$, where N_{ab} is the news field and h_{ab} the induced metric on I . Substituting into (23), using the constancy of H on I , we obtain

$$F = -2\tau N_{ab}N^{ab}. \quad (25)$$

This will be recognized as the familiar expression for the flux of the Bondi energy-momentum.

So far, we have dealt entirely with the vacuum case, i.e., that in which the stress-energy vanishes in a neighborhood of I . We now include asymptotic matter. Consider, then, an asymptotically flat space-time, but now demand that $T_{ab} = \Omega^{-2}\tilde{T}_{ab}$ have a smooth extension to I . In essence, we are demanding that the physical stress-energy vanish asymptotically as $1/r^2$, which turns out to be the physically correct condition on the matter.

What would one expect physically to be the matter flux at infinity? One might think that, since $T^a{}_m \xi^m$ would represent the energy-momentum or angular momentum current of the matter when ξ^a is a Killing field, and since n^a is the normal to I , the correct flux, when ξ^a is any BMS generator, should be $T_{ab}n^a \xi^b$. However, there is a complication, involving the numerical factor for this expression. We illustrate it by the following two examples.

Consider first an exact asymptotically flat solution, with axial symmetry given by physical Killing field φ^a . Assign, to each cross section S of I , the Komar integral, $\int_S \tilde{\epsilon}_{abmn} \tilde{\nabla}^m \varphi^n dS^{ab}$. Then the total flux, which determines the difference between this integral for two cross sections, must be just the curl of this integrand. Evaluating this curl on I , we obtain $T_{ab}n^a \varphi^b$. Since the space-time is axisymmetric, we would expect no gravitational flux of angular momentum. So, this total flux should also be the flux of matter alone. In this example, then, we are able to argue physically for a particular expression for the matter flux.

Now consider, as the second example, an exact solution of Einstein's equation which is at all times spherically symmetric. Let the solution, at early times, be that of a static fluid ball. At some point in time, however, let the matter of the ball suddenly turn into a null fluid without pressure, which will then radiate away to infinity. At late epochs, then, the space-time is flat (Fig. 1). Let \tilde{t}^a be a time-translational BMS generator which equals the static Killing field at both early and late times. In the intermediate regime, during the radiation of the null fluid, \tilde{t}^a will not of course be a Killing field. Let S_i and S_f be respective cross sections of I at early and late times, and assign to each its Komar integral using

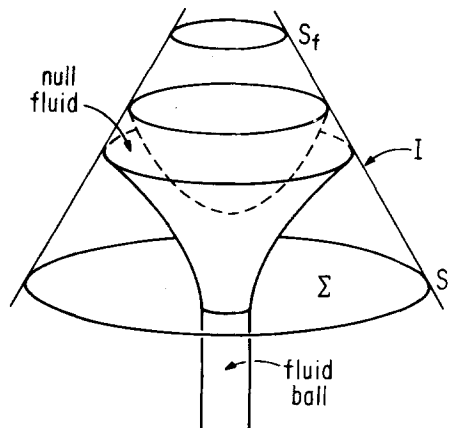


FIG. 1. A spherically symmetric, asymptotically flat space-time representing a fluid ball which becomes a null fluid. The region to the future of the null fluid is flat. The spacelike 3-surface Σ cuts the ball, and meets I at S_f . The surface Σ' joins S_i to S_f along I , and then cuts across the flat space-time.

f^a . Then the latter must be zero, while the former, by Gauss's theorem, must be $m = \int_{\Sigma} (\tilde{T}_{ab} - 1/2 \tilde{T} \tilde{g}_{ab}) f^b d\tilde{S}^a$, where Σ is any three-dimensional spacelike slice, with boundary S_i , in the initial region. Since spherical symmetry should preclude the radiation of gravitational energy, the difference between these two, m , should be just the matter flux at I , integrated between S_i and S_f . To compute this flux, restrict to the weak-field limit, i.e., work to first order in the mass m . The virial theorem then gives

$$m = \frac{1}{2} \int_{\Sigma} \tilde{T}_{ab} f^b d\tilde{S}^a. \quad (26)$$

But, to the present order in m , \tilde{T}_{ab} is conserved and f^a is a Killing field, at all times. So, the integral on the right in (26) can be performed over *any* surface Σ' with boundary S_i . We now choose for Σ' a surface which coincides with I from S_i to S_f , and then cuts across the space-time in the flat region, as shown in the figure. Then the only contribution to the integral in (26) will be from I , and so we obtain

$$m = \frac{1}{2} \int_I T_{ab} t^b dS^a. \text{ Thus, we expect, for the matter flux in this example, } \frac{1}{2} T_{ab} n^a t^b.$$

These examples suggest, then, that for any asymptotic integral which reduces to the Komar integral in the Killing case the matter contribution to the flux should be $T_{ab} n^a \xi^b$ for a rotational BMS generator, and $\frac{1}{2} T_{ab} n^a \xi^b$ for a translational generator. But this conclusion would seem to contradict the linear dependence of the flux on the generator. Two rotations, for example, could differ by a translation.

With this discussion as background, we now turn to the generalization, to the presence of asymptotic matter, of the linkage and its flux. We retain $X = 0$ as the gauge condition on the generator, and (14) as the definition of the linkage integral. Again compute the flux as in (20), but now, instead of using Einstein's vacuum equation (1), use that equation with its source term. There results

$$F = -\nabla^a \nabla^b X_{ab} + 3\nabla^a X_a + \frac{3}{4} \nabla^2 X + \frac{1}{24} R X + T_{ab} n^a \xi^b. \quad (27)$$

Thus, the general formula for the flux of the linkage in the presence of radiating matter is just (20) with an additional matter term, $T_{ab} n^a \xi^b$. Now consider the special case of a supertranslation. Einstein's equation was also used in deriving (23) from (19). Again including the source terms, we obtain, for the flux for a supertranslation, just (23), but with an additional term, $\frac{1}{2} T_{ab} n^a \xi^b$, on the right.

Thus, the anomaly of the numerical factor before the matter term which arose in our example is reflected, in the case of the linkages, by the alternative forms in which the total flux may be expressed. The general formula (20) for the flux vanishes when the generator is a Killing field (i.e., when $X_{ab} = 0$). In this form, the matter contribution is $T_{ab} n^a \xi^b$. The reduced formula (23) for the flux in the case of a supertranslation reduces to the square of the news for a translation. In this form, the matter contribution is $\frac{1}{2} T_{ab} n^a \xi^b$. There is no general expression for the flux of the form "gravitational field contribution plus matter contribution," where the first term vanishes both for a Killing field and for a translation in the absence of news, and the second term is some

multiple of $T_{ab} n^a \xi^b$.

These considerations rather suggest that there may be a theorem to the effect that the linkage integral is the unique physically reasonable candidate for the asymptotic energy-momentum-angular momentum. The linkage integral has all the following properties: (i) it is defined for any BMS generator and any cross section, and is linear in the former; (ii) it is invariant under conformal changes and passage to an equivalent generator; (iii) it is "local" in the sense that the value of the integral depends only on the geometry and generator in a neighborhood of the cross section; (iv) it reduces to the Komar value for a Killing field, and to the Bondi value for a translation; (v) it possesses a flux; and (vi) it applies also in the presence of radiating matter. We conjecture that these properties characterize the linkage uniquely.

Consider, for instance, the following possibility for an alternative candidate. Let the flux be, instead of (20), simply

$$F = -2N_{ab} X^{ab}. \quad (28)$$

This flux expression reduces to the Bondi formula, (25), in the case of a translation ($X_{ab} - \frac{1}{2} X h_{ab} = \tau N_{ab}$), and to zero in the case of a Killing field ($X_{ab} = 0$). Furthermore, this expression has the attractive feature that it always vanishes in Minkowski space-time, since $N_{ab} = 0$ there, whereas, by (23) neither the linkage nor its flux vanish in general in Minkowski space-time. But more must be done in order to have a candidate satisfying the six properties above. Of what local integral, over cross sections, will this be the flux? It seems very likely that there exists none. Further, how is the flux expression (28) to be modified in the presence of radiating matter? Our examples show that the addition to (28) of neither $T_{ab} n^a \xi^b$ nor half this will do.

4. LINEAR PERTURBATIONS

In this section we consider the effects, on generators and their linkages, of first-order perturbations of the gravitational field. The consideration of such perturbations illuminates not only the physical meaning of the linkages, but also the role of supertranslations. For simplicity, we restrict the discussion to the case with all sources—both in the background and for the perturbation—vanishing in a neighborhood of I .

Let $\tilde{M}, \tilde{g}_{ab}$ be an asymptotically flat space-time, with unphysical metric $g_{ab} = \Omega^2 \tilde{g}_{ab}$. Let $\tilde{\gamma}_{ab}$ be a first-order perturbation of the physical metric. We wish to impose on this $\tilde{\gamma}_{ab}$ the condition that "asymptotic flatness be preserved to first order." In more detail, we require that there exist some perturbation, $\delta\Omega$, of the conformal factor such that the various conditions for asymptotic flatness, which of course already hold in the background, continue to hold to first order in the perturbation. The condition that Ω vanish at I requires that $\delta\Omega$ also vanish at I , i.e., that $\delta\Omega = \Omega\omega$ for some smooth function ω on M . Then the perturbation, γ_{ab} , of the unphysical metric becomes

$$\gamma_{ab} = \Omega^2 \tilde{\gamma}_{ab} + 2\omega g_{ab}. \quad (29)$$

The condition that g_{ab} be smooth at I requires that this γ_{ab} also be smooth at I . Finally, the condition that Ω be a Bondi conformal factor requires, from (1), that the combination

$\gamma_{ab}n^an^b - 2n^a\nabla_a\omega$ vanish at I to second order in Ω .

Let $\tilde{\gamma}_{ab}$ be any perturbation of the physical metric which preserves asymptotic flatness to first order, as characterized above. This $\tilde{\gamma}_{ab}$ is of course subject to the freedom of gauge transformations: $\tilde{\gamma}_{ab} \rightarrow \tilde{\gamma}_{ab} + 2\tilde{\nabla}_{(a}\tilde{\lambda}_{b)}$, where $\tilde{\lambda}^a$ is any vector field on M . We exploit this gauge freedom to set γ_{ab} equal to zero at I , and ω equal to zero everywhere in a neighborhood of I . That this may be achieved follows immediately from the fact that, given two asymptotically flat space-times, one can find a diffeomorphism between them, in neighborhoods of I , which preserves Ω everywhere in these neighborhoods and g_{ab} at I . In this gauge, the perturbation of the conformal factor is zero, while the perturbation of the unphysical metric is given by $\gamma_{ab} = 2\Omega Y_{ab}$, for some smooth field Y_{ab} on M with, by the remarks following (29), $Y_{ab}n^an^b$ vanishing on I .

A particularly convenient feature of this gauge is that it yields the following property: under any perturbation in the present gauge, any BMS generator ξ^a remains a generator to first order. This property is immediate, taking the perturbation of (3) and setting $\delta g_{ab} = 2\Omega Y_{ab}$. In fact, it follows from this that, for ξ^a a generator, with K and X_{ab} given by (3), the first order change in K under this perturbation of the metric vanishes, while that of X_{ab} is given by

$$\delta X_{ab} = \mathcal{L}_{\xi} Y_{ab} - KY_{ab}. \quad (30)$$

The present gauge still permits the following, more restricted, gauge freedom: $\tilde{\gamma}_{ab} \rightarrow \tilde{\gamma}_{ab} + 2\tilde{\nabla}_{(a}\tilde{\lambda}_{b)}$, where now $\tilde{\lambda}^a = \lambda^a$ must itself be a BMS generator. That is, the generators of the remaining gauge group are exactly the BMS generators. Under such a gauge transformation, the change in Y_{ab} is particularly simple; $Y_{ab} \rightarrow Y_{ab} + X_{ab}(\lambda)$, where $X_{ab}(\lambda)$ is the field obtained from the generator λ^a via (3).

This remaining gauge freedom, unfortunately, is enough to yield ambiguities in the perturbation of the linkage integrals. Fix an asymptotically flat space-time, with g_{ab} and Ω , a generator ξ^a , and a cross section S . Write $L(\xi, S)$ for the linkage integral over the cross section S for the generator ξ^a . Consider now a first-order perturbation of the metric which is in the present gauge and which itself is pure gauge, i.e., which is given by $Y_{ab} = X_{ab}(\lambda)$ for some BMS generator λ^a . Then, since the linkage integral $L(\xi, S)$ must be unchanged under applying simultaneously to g_{ab} , ξ^a , and S the diffeomorphism generated by λ^a , we have

$$\delta_{\lambda} L(\xi, S) = -\delta_S L(\xi, S) - L([\lambda, \xi], S), \quad (31)$$

where $\delta_{\lambda} L(\xi, S)$ is the change in the linkage, keeping ξ^a and S fixed, under our gauge perturbation in g_{ab} , and $\delta_S L(\xi, S)$ is the change in the linkage integral, keeping ξ^a and g_{ab} fixed, under the displacement of the cross section S along λ^a . This last term is of course expressible in terms of the flux. In short, a gauge-change in the background geometry is equivalent to moving the cross section slightly and changing the generator slightly. In particular, the change in the linkage under a perturbation in the geometry, keeping the generator and cross-section fixed is not gauge-invariant. Note from (31) however that, for the case in which the gauge generator λ^a is equivalent to the zero generator, no linkages are changed to first order. That is, as far as the values of linkages are concerned, the asymptotic gauge group on metric perturbations is the

BMS group.

Under certain special conditions on ξ^a and S the gauge ambiguities of the previous paragraph can be avoided. When they can, "the first order change under a metric perturbation of the ξ -linkage over S " will be physically meaningful. One such set of conditions is ξ^a is a translation generator having zero flux at points of S and such that all translations orthogonal to ξ^a have vanishing linkage integral over S . The vanishing of the flux ensures that the first term on the right in (31) vanishes for all gauge-generators λ^a , while, since the most general generator of the form $[\lambda, \xi]^a$ is a translation orthogonal to ξ^a , the vanishing of the linkages of these ensures that the second term on the right in (31) also vanishes. These conditions are satisfied, for example, for ξ^a , the Killing time-translation generator in the Schwarzschild solution, and S , any cross section. Thus, the first-order change in the mass of the Schwarzschild solution under any perturbation preserving asymptotic flatness makes physical sense. Note, however, that the analogous first-order change in the momentum does not. A second set of conditions is ξ^a is a (rotational) Killing field tangent to S . That ξ^a be a Killing field ensures that its flux vanishes, and so that the first term on the right in (31) always vanishes. That the second term also vanishes follows from rewriting (31) with the roles of ξ^a and λ^a reversed and using the fact that ξ^a is tangent to S . These conditions are satisfied, for example, for ξ^a , any rotational Killing field in the Schwarzschild solution, and the cross section S , an orbit of the spherical symmetry. Thus, the first-order change in the angular momentum of the Schwarzschild solution (measured on a spherically symmetric cross section) under any perturbation preserving asymptotic flatness makes physical sense.

Another way of avoiding these gauge ambiguities is by imposing conditions on the perturbation Y_{ab} , rather than on the generator or the cross section. Consider, as one example, perturbations which are "internally generated" in the space-time, in the following sense: prior to some retarded time, Y_{ab} vanishes in a neighborhood of I . We show that, in most background space-times, the first-order change in any linkage under such an internally generated perturbation is unambiguous. The gauge transformations which preserve the property of being internally generated are those which change Y_{ab} by $X_{ab}(\lambda)$, where λ^a is any BMS generator for which $X_{ab}(\lambda)$ vanishes in a neighborhood of I prior to some retarded time. But this implies that, near I at early times, λ^a is a Killing field of the background space-time. We now demand of the background that it not "spontaneously break symmetries" in the following sense: any Killing field in a neighborhood of I prior to some retarded time must extend to a Killing field in a neighborhood of I for all times. Under this condition, there exists a Killing field $\hat{\lambda}^a$ of the background in a neighborhood of I , such that $\hat{\lambda}^a = \lambda^a$ at early times. By the Killing character of $\hat{\lambda}^a$, $X_{ab}(\lambda) = X_{ab}(\lambda - \hat{\lambda})$, and so we may take as the gauge vector $\lambda^a - \hat{\lambda}^a$ rather than λ^a . By the vanishing of $\lambda^a - \hat{\lambda}^a$ at early times, this BMS generator is equivalent to the zero generator. Thus, all gauge transformations which preserve the property of being internally generated²¹ are via generators equivalent to zero. But,

by (31), such gauge transformations do not change the values of any linkages. The additional condition that the background not spontaneously break symmetries holds in most cases of interest, e.g., for stationary space-times, and for space-times having no Killing fields anywhere. It appears that the conclusion of the argument actually fails without this condition. It seems peculiar that the breaking of an internal symmetry in the background—so small as to produce no physical effects—can nonetheless destroy the gauge-invariance of the first-order change in a linkage under an internally generated perturbation.

In order to discuss the actual changes in the linkages under various perturbations, it is convenient to have the following formula. Consider any asymptotically flat space-time, with g_{ab} , Ω , and consider any perturbation $\tilde{\gamma}_{ab}$ of the physical metric which preserves asymptotic flatness, is in the present gauge, and satisfies the linearized Einstein equation. Contracting the latter with $n^a n^b$, substituting $\tilde{\gamma}_{ab} = 2\Omega^{-1}Y_{ab}$, and evaluating on I , we obtain, after some manipulation,

$$n^a \nabla_a F(Y) = -\Omega^{-1} C_{abcd} n^b n^d Y^{ac}, \quad (32)$$

where $F(Y)$ is the flux expression (20) with Y_{ab} substituted for X_{ab} , and C_{abcd} is the background Weyl tensor. It follows from asymptotic flatness that $\Omega^{-1} C_{abcd}$ remains finite on I , essentially a reflection of the peeling property. So, the right side of (32) is well defined.²² In general, the $F(Y)$ in (32) is not the flux of any linkage. However, this equation holds in particular when it is, i.e., with Y_{ab} pure gauge ($Y_{ab} = X_{ab}(\lambda)$ with λ^a any BMS generator).

The following are examples of the simplifications which arise from restricting consideration to internally generated perturbations. One would expect physically that, to first order, energy cannot be radiated away in a stationary space-time, nor angular momentum in an axisymmetric one. What precise results are available which reflect these expectations?

An appropriate result in the stationary case is easy: for ξ^a , a timelike Killing field in the background, no internally generated perturbation can change the linkage of ξ^a , for any cross section, to first order. This follows immediately from the flux expression for a translation, Eq. (25). Applied to the background, this equation implies that the news vanishes in the background. But now, because of the quadratic dependence on the news, this equation also implies that the first-order change in the flux of ξ^a vanishes. This implies, finally, that the first-order change in any linkage of ξ^a must vanish (since this necessarily holds at early times, when $Y_{ab} = 0$.)

A similar, but apparently somewhat weaker, result is available for other Killing fields. Let ξ^a be a Killing field of the background, so $X_{ab}(\xi) = 0$, while the perturbation of X_{ab} is given by (30). Replacing Y_{ab} by X_{ab} in (32), and taking its first-order perturbation, we obtain

$$n^a \nabla_a \delta F(X) = -\Omega^{-1} C_{abcd} n^b n^d \delta X^{ac}. \quad (33)$$

We now suppose that the background space-time is nonradiative, in the sense that $\Omega^{-1} C_{abcd} n^b n^d$ is a multiple of $n_a n_c$ on I . Then, by transversality of X_{ab} , the right side of (33) vanishes, and so $\delta F(X) = 0$. Thus, the flux of ξ^a is unchanged to first order under the perturbation, and so, therefore, are

its linkages. We have shown that, in a nonradiative space-time, no linkage for a Killing field is changed to first order by an internally generated perturbation. So, for example, internally generated perturbations in the Schwarzschild or Kerr space-times cannot radiate angular momentum to first order.

Finally, we show that even for a radiating background, an internally generated perturbation cannot change the linkage for a rotational Killing field, provided the linkage is taken over a cross section S to which that field is tangent. This also follows from (33). For simplicity, choose Ω so that $\xi^a \nabla_a \Omega = 0$. Since ξ^a is a Killing field, we have $\mathcal{L}_\xi C_{abcd} = 0$. So, by (30), the right side of (33) is $\xi^m \nabla_m (-\Omega^{-1} C_{abcd} n^b n^d Y^{ac})$. Now integrate (33) over the region of I between the initial time and any cross section S' to which ξ^a is everywhere tangent. Then, since the integral of the right side vanishes, we conclude that the integral of δF over S' must vanish. But this is true for every such S' . So, the integral of δF over the region of I between the initial time and our given cross section S must vanish. Hence, the first-order change in the linkage of ξ^a over S must vanish.

The examples above can be regarded as part of a larger question. Fix an asymptotically flat space-time, and a perturbation Y_{ab} which preserves asymptotic flatness to first order and is in the present gauge. Fix a cross section S of I in the midst of the Y -radiation. What should we take as the physical energy, momentum, or angular momentum of the perturbed system at the time represented by S ? That is, what are the appropriate generators to use in computing the linkage integrals over S ? Physically, one would expect that the appropriate generators would be those somehow adapted to the geometry near S . But no such local criterion is known in the presence of radiation. This is the essence of the fact that the asymptotic symmetry group in general relativity is the BMS and not the Poincaré group. Above, the generators were selected according to the following criterion. Assume that the perturbation is internally generated, select generators adapted to (e.g., Killing fields of) the now unperturbed background geometry at early times, and use these generators to compute linkages at all times. One can ask whether this choice is at all physically appropriate, and whether other, better choices might be available.

There is one arrangement in which one can obtain evidence on these issues. Consider an asymptotically flat space-time, with internally generated perturbation Y_{ab} in a gauge such that it vanishes in a neighborhood of I prior to some retarded time. Suppose now that this perturbation “dies out” in the limit of late times in the following sense: in this limit, $Y_{ab} = X_{ab}(\lambda)$ on I , for some BMS generator λ^a . That is, we require that, up to gauge, the perturbation vanish on I at late times. The generator λ^a is of course uniquely determined up to a generator which is a Killing field of the background at early times. By a gauge transformation, we could alternatively have Y_{ab} vanish on I at late times, and then can be $-X_{ab}(\lambda)$ at early times. Now suppose that some local geometrical criterion has been used to select a physically appropriate generator ξ^a at early times. Apply the same criterion locally to select a generator in the limit of late times. Then, since Y_{ab} differs from zero on I at late times only by

the gauge action of λ^a , the selected generator will differ from ξ^a , to first order, by $[\lambda, \xi]^a$. That is, this perturbation in the geometry induces a shift in the locally determined generators which, to first order, changes ξ^a by $[\lambda, \xi]^a$. In short, there would be an inner automorphism, determined by λ^a , on the BMS Lie algebra, which sends each generator at early times to that generator selected by the same criterion at late times.

Can such a shift actually take place? We give an example to show that it can. Let the background be Minkowski space-time, in standard spherical coordinates t, r, θ, ϕ . This space-time is of course asymptotically flat, with conformal factor $\Omega = r^{-1}$. Introduce in this space-time a conserved stress-energy field, \tilde{T}^{ab} , with the following features. Initially, \tilde{T}^{ab} represents a static, spherically symmetric fluid ball of mass m , at rest and centered at the origin. At late times, \tilde{T}^{ab} represents two fluid balls, each static and spherically symmetric and of mass $(m/2)(1 - v^2)^{1/2}$ in its own rest frame, but moving from the center in opposite directions (along $\theta = 0$ and $\theta = \pi$) with speed v . The v factor ensures total mass conservation, as required by conservation of \tilde{T}^{ab} .) Now consider an exterior solution $\tilde{\gamma}_{ab}$ of the linearized Einstein equation with this \tilde{T}^{ab} as source. Initially, we may set

$$\tilde{\gamma}_{ab} = 2m/r(\tilde{\nabla}_a t \tilde{\nabla}_b t + \tilde{\nabla}_a r \tilde{\nabla}_b r), \quad (34)$$

the first-order perturbation to Schwarzschild off Minkowski space-time. This form of the perturbation at early times preserves asymptotic flatness to first order. Indeed, setting $Y_{ab} = \frac{1}{2}\Omega\tilde{\gamma}_{ab}$, not only is Y_{ab} finite at I , it even vanishes there to second order in Ω . Similarly, at late retarded times, let $\tilde{\gamma}_{ab}$ be the sum of two terms of the form (34), appropriately adjusted for the motion of each ball. Again, the corresponding Y_{ab} will vanish at I to second order in Ω . During intermediate times, when the first-order radiation is reaching I , $\tilde{\gamma}_{ab}$ will be more complicated. In fact, $Y_{ab} = \frac{1}{2}\Omega\tilde{\gamma}_{ab}$ will not necessarily even remain finite at I then. Now perform on this $\tilde{\gamma}_{ab}$ a gauge transformation which vanishes at early times and which leads to a transformed perturbation, $\tilde{\gamma}'_{ab}$, with finite $Y'_{ab} = \frac{1}{2}\Omega\tilde{\gamma}'_{ab}$ everywhere on I . Then at early times $Y'_{ab} = Y_{ab}$, while at late times $Y'_{ab} = Y_{ab} + X_{ab}(\lambda)$, for some BMS generator λ^a . Clearly, λ^a is unique up to a generator which is a Killing field in Minkowski space-time at early times, and so we may assume without loss of generality that λ^a is a supertranslation generator. This λ^a generates the shift in the generators from early to late times.

We may determine this supertranslation λ^a for our example using (32). Here, the Weyl tensor of the background, and so the right side of (32), vanishes. Applying this equation to Y'_{ab} , we conclude that $F(Y')$ is constant along each n -integral curve on I . Initially, $Y'_{ab} = Y_{ab}$, and so we may calculate $F(Y')$ explicitly from (34). There results $F(Y') = m$. So, by (32), we must have $F(Y') = m$ everywhere on I . Similarly, at late times $Y'_{ab} = Y_{ab} + X_{ab}(\lambda)$, and so $F(Y') = F(Y) + F(X(\lambda))$. The left side is m , while $F(Y)$ at late times is again calculated explicitly. In this way, we obtain

$$F(X(\lambda)) = m - \frac{1}{2}m(1 - v^2)^2[(1 - v\cos\theta)^{-3} + (1 + v\cos\theta)^{-3}].$$

This, then, is the formula for the flux of the supertranslation λ^a . We see in particular that λ^a is not, in this example, equivalent to a translation, and so there is indeed a nontrivial supertranslation shift in the generators. [The perturbation of this example can be made to vanish near I at early times by subtracting (34) from $\tilde{\gamma}'_{ab}$ for all times.]

This particular example has, in addition, the following curious feature. At initial times, the background is Minkowski and the perturbation Y'_{ab} vanishes on I . Thus, at early times one can distinguish an entire preferred Poincaré subalgebra of the BMS Lie algebra, namely that consisting of equivalence classes containing a generator ξ^a with $X_{ab}(\xi) = 0$ on I . Similarly, at late times the background is again Minkowski and the perturbation Y'_{ab} differs only by gauge from vanishing on I . So, the same criterion again yields an entire preferred Poincaré subalgebra at late times. That is, in this example one would regard all the Poincaré quantities, energy, momentum, and angular momentum, as making physical sense at both early and late times.

In this example, the shift between the initial preferred Poincaré subalgebra and the final one (initial generator ξ^a sent to final generator ξ^a changed to first order by $[\lambda, \xi]^a$) is via a supertranslation λ^a . Thus, translations, since they commute with λ^a , are not shifted. That is, energy and momentum would be calculated at initial and final times using the same generators. But for ξ^a a rotation, $[\lambda, \xi]^a$ is a supertranslation, and in general not a translation. Thus, there arise nontrivial shifts in the rotation generators. Quite different generators would be used to determine the physical "angular momentum" at initial and final times, and, since they differ by supertranslations, these different generators would yield different linkages. The conclusion seems to be that there are two mechanisms by which a system can change its physical angular momentum. One is by simply radiating angular momentum, i.e., changing, as described by the flux, the linkage for a fixed generator between two cross sections. The other is by catalyzing a shift in the physical choice of generator to be used in determining the angular momentum. The example above shows that the second mechanism can be operative. Can it be physically important?

There is a further complication. In this example, the supertranslation λ^a is determined only up to the addition of an arbitrary translation. Thus, even the correspondence between the preferred initial and final Poincaré subalgebras is not unique. (In the case of no perturbation, the correspondence is made unique by adding an appropriate translation to achieve $\lambda^a = 0$. But, when λ^a is a nontrivial supertranslation, there is no natural way to "remove its translation part.") This lack of uniqueness means that we do not know precisely what the "same rotation" at early and late times means. The "directions" of the rotations can be compared unambiguously. The problem is in comparing their "origins." Thus, not only is there a second mechanism for a system to change its angular momentum, but even the comparison between initial and final angular momenta is subject to an ambiguity.

These observations suggest that it may be inappropriate to think of isolated systems in general relativity in terms of

the familiar “angular momentum,” i.e., in terms of the Poincaré group. Perhaps the BMS group must be faced squarely.

¹R. Penrose, Proc. R. Soc. London Ser. A **284**, 159 (1965).
²R. Penrose, in *Battelles Rencontres*, edited by C. M. DeWitt and J. A. Wheeler (Benjamin, New York, 1968).
³R. Geroch, in *Asymptotic Structure of Space-Time*, edited by F. P. Esposito and L. Witten (Plenum, New York, 1977).
⁴R. Geroch and G. Horowitz, Phys. Rev. Lett. **40**, 203 (1978).
⁵H. Bondi, M. G. J. Van der Burg, and A. W. K. Metzner, Proc. R. Soc. London Ser. A **270**, 103 (1962).
⁶R. K. Sachs, Phys. Rev. **128**, 2851 (1962).
⁷S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-Time* (Cambridge U. P., Cambridge, 1973).
⁸S. W. Hawking, in *General Relativity*, edited by S. W. Hawking and W. Israel (Cambridge U. P., Cambridge, 1979).
⁹R. Penrose, in Ref. 8.
¹⁰R. K. Sachs, in *Relativity, Groups, and Topology*, edited by C. M. DeWitt and B. DeWitt (Gordon and Breach, New York, 1964).
¹¹R. Penrose, in Ref. 10.
¹²L. Tamburino and J. Winicour, Phys. Rev. **150**, 1039 (1966).
¹³J. Winicour, J. Math. Phys. **9**, 861 (1968).
¹⁴E. T. Newman and R. Penrose, J. Math. Phys. **7**, 863 (1966).

¹⁵J. Winicour, in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980), Vol. 2.
¹⁶A. Komar, Phys. Rev. **113**, 934 (1959).
¹⁷R. Geroch and B. C. Xanthopoulos, J. Math. Phys. **19**, 714 (1978).
¹⁸Alternatively, we could have defined a BMS generator originally as a vector field $\bar{\xi}^a$ on the 3-manifold I which, when used to take the Lie derivative of the pull back to I of the field $g_{ab}n^cn^d$, yields zero. To each such $\bar{\xi}^a$ there corresponds exactly one equivalence class of generators, as here defined.
¹⁹Consider, e.g., the quadratic case. Let the vector space be that of all quadratic functions on the vector space of equivalence classes of BMS generators, and let the action of the BMS group be that which arises from the adjoint action of this group on its Lie algebra.

²⁰In fact, this same argument shows that the integral

$$\int_S [\epsilon_{abcd} \nabla^c (\Omega^{-2} \xi^d) + (l^m n_m)^{-1} l^n \nabla_n X \epsilon_{ab}] dS^{ab}$$

is invariant under any replacement of ξ^a by an equivalent generator, with no gauge conditions whatever. Note that this reduces to (10) under (9), and to (14) under (13).

²¹The weaker condition on Y_{ab} that it vanish only on I prior to some retarded time does not seem to work as a replacement for “internally generated.” It appears that the remaining gauge freedom, to change Y_{ab} by $X_{ab}(\lambda)$ such that $X_{ab}(\lambda)$ vanishes only on I at early times, can in general change linkages.

²²Note that $F(Y)$ is well defined since Ω remains a Bondi conformal factor under the perturbation (with $\delta\Omega = 0$) only if $Y_{ab}n^b = \Omega Y_a$, for some smooth Y_a .

Energy density and spatial curvature in general relativity

Theodore Frankel

Department of Mathematics, University of California, La Jolla, California 92093

Gregory J. Galloway

Department of Mathematics, University of Miami, Coral Gables, Florida 33124

(Received 16 October 1979; revised manuscript received 7 May 1980)

Positive energy density tends to limit the size of space. This effect is studied within several contexts. We obtain sufficient conditions (which involve the energy density in a crucial way) for the compactness of spatial hypersurfaces in space-time. We then obtain some results concerning static or, more generally, stationary space-times. The Schwarzschild solution puts an upper bound on the size of a static spherically symmetric fluid with density bounded from below. We derive a result of roughly the same nature which, however, requires no symmetry and allows for rotation. Also, we show that static or rotating universes with $\Lambda = 0$ require that the density along some spatial geodesic must fall off rapidly with distance from a point.

PACS numbers: 04.20.Cv

I. INTRODUCTION

The energy density of space-time is related to the curvature of a spatial hypersurface via the Einstein field equations and the Gauss equations. Energy density, being inherently positive, contributes a positive term to the Ricci curvature of a spatial hypersurface. Positive Ricci curvature tends to limit the size of the spatial hypersurface. This effect was studied for general Riemannian manifolds in a classical paper of Myers.¹ In this paper we make use of two refinements of Myers' results (due to Ambrose² and one of the authors³) to demonstrate within several different contexts the relationship between energy density and the "size of space."

A number of theorems (including some of the singularity theorems) of general relativity assume a *closed* space-time, i.e., a space-time which admits a compact spatial hypersurface. It is, therefore, of interest to establish some sufficient conditions for the compactness of spatial hypersurfaces. In Sec. II we present several such compactness results for spatial hypersurfaces of various types. (These results modify and generalize some earlier results of one of the authors.⁴)

In Sec. III we obtain generalized versions of two classical results in general relativity involving static space-times which we now briefly recall. The Schwarzschild solution puts an upper bound on the size of a static, spherically symmetric fluid ball. This bound may be expressed in terms of the average density of the ball. As Eddington⁵ has said, "The limit exists because the presence of dense matter increases the curvature of space, and makes the total volume of space smaller. Clearly the volume of the material sphere cannot be larger than the volume of space." We obtain a bound of roughly the same nature on the size of an arbitrary fluid mass in a general *stationary* space-time. Our result requires no spherical symmetry and allows for rotation.

It is well known that a static fluid filled space-time cannot be a solution to the Einstein equations (with cosmological constant $\Lambda = 0$) *provided* the time lines along which the

gravitational field is constant are geodesics.⁶ Maitra⁷ has constructed a dust filled stationary space-time (the time lines along which the gravitational field is constant rotate) which satisfies the Einstein equations (with $\Lambda = 0$). In his model, which is cylindrically symmetric, the density falls off rapidly with distance from the axis of rotation. We prove for a general static or stationary space-time which is a solution to the Einstein equations (with $\Lambda = 0$) that the density along *some* spatial geodesic must fall off rapidly (in a precise sense) with distance from a point.

II. SUFFICIENT CONDITIONS FOR THE COMPACTNESS OF SPATIAL HYPERSURFACES

Let M^4 be a space-time, i.e., a smooth four-dimensional manifold equipped with a Lorentzian metric $\langle \cdot, \cdot \rangle$ (with signature: $- + + +$). Assume, in addition, that M^4 is time orientable. Let V^3 be a spatial hypersurface in M^4 . Then V^3 in the induced metric has the structure of a Riemannian manifold. The following result of Ambrose⁸ gives a useful criterion for the compactness of a *complete* Riemannian manifold.

Lemma (Ambrose): Let M^n be a complete Riemannian manifold. If there is a point q of M^n such that along each geodesic γ emanating from q the Ricci curvature satisfies

$$\int_0^\infty \text{Ric}(X, X) d\lambda = +\infty, \quad (1)$$

where λ is arc length along γ and X is the unit tangent to γ , then M^n is compact. [$\text{Ric}(X, X) = \sum R_{ij} X^i X^j$ where R_{ij} are the components of the Ricci tensor and X^i are the components of X .]

We shall apply Ambrose's result to the case of a spatial hypersurface V^3 in space-time M^4 . The unit tangent vectors to the future directed geodesics orthogonal to V^3 define a smooth unit timelike vector field T at least in a neighborhood of V^3 . This vector field may be thought of as defining a distinguished geodesic reference frame, where the normal geodesics, which are the integral curves of T , represent the world-

lines of a distinguished class of geodesic observers.

Let X be a vector tangent to V^3 at q . Extend X along the normal geodesic γ through q by making it invariant under the flow generated by T :

$$[X, T] = \nabla_X T - \nabla_T X = 0,$$

where ∇ is the Levi-Civita connection and $[,]$ is the Lie bracket. Because the flow is geodesic, X remains perpendicular to T , i.e., remains in the "rest space" of the geodesic observer. The vector field X along γ may be thought of as a position vector tracking nearby geodesic observers. The vector fields $v(X) = \nabla_T X$ and $a(X) = \nabla_T \nabla_T X$ along γ are called the 3-velocity and 3-acceleration of X , respectively.⁹ The inequality $\langle v(X), X \rangle \geq 0$ indicates a recession of nearby geodesic observers in the direction of X , and the inequality $\langle a(X), X \rangle \leq 0$ indicates a deceleration of the recession in the direction of X . Introduce the scalar: $\Theta = \text{div} T$, where div is the space-time divergence operator. Let B be the unit "2-sphere," i.e., the set of unit vectors in the tangent space of V^3 at q . A computation shows

$$\Theta(q) \equiv \text{div} T(q) = \frac{1}{4\pi} \int_{X \in B} \langle v(X), X \rangle d\Omega,$$

where $d\Omega$ is the area element of B . Thus, Θ is a measure of the average rate of expansion of the normal geodesics. (Note: if X is a unit vector tangent to V^3 which is made invariant under the normal geodesic flow, then

$$\left. \frac{d}{ds} \right|_q \|X\| = \langle v(X), X \rangle,$$

where s is proper time and $\|X\| = \langle X, X \rangle^{1/2}$ is the length of X). Geometrically, $\Theta(q)$ is minus the trace of the second fundamental form of V^3 at q .¹⁰

Theorem 1: Let V^3 be a spatial hypersurface in M^4 . Assume V^3 is complete in the induced metric. If there is a point q in V^3 such that along each geodesic γ of V^3 emanating from q the condition

$$\int_0^\infty [\text{Ric}(X, X) - \langle a(X), X \rangle - \langle v(X), X \rangle \Theta + \langle v(X), X \rangle^2] d\lambda = +\infty \quad (2)$$

is satisfied, where λ is arc length along γ and X is the unit tangent to γ , then V^3 is compact.

Proof: A computation shows¹¹ that, for unit vectors X tangent to V^3 ,

$$\text{Ric}_V(X, X) = \text{Ric}(X, X) - \langle a(X), X \rangle - \langle v(X), X \rangle \Theta + \langle v(X), X \rangle^2 + \langle v(X), e_2 \rangle^2 + \langle v(X), e_3 \rangle^2,$$

where Ric_V is the Ricci tensor of V^3 in the induced metric and X, e_2, e_3 are orthonormal. Combining this equation with Ambrose's result yields the desired conclusion.

If space-time is filled with a perfect fluid having four-velocity u , density ρ and pressure p then the Einstein equations imply,

$$\text{Ric}(X, X) = 4\pi\kappa(\rho - p) + 8\pi\kappa(\rho + p)\langle X, u \rangle^2 \geq 4\pi\kappa(\rho - p) \quad (3)$$

(assuming $\rho + p \geq 0$) for any unit spacelike vector X . (Here κ is the gravitational constant.) Cosmologically, we expect $\langle a(X), X \rangle \leq 0$. (In fact a computation shows

$$-\frac{1}{3}\text{Ric}(T, T) = \frac{1}{4\pi} \int_{X \in B} \langle a(X), X \rangle d\Omega,$$

where, for ordinary matter, $\text{Ric}(T, T) \geq 0$. So there must at least be deceleration on the average.) Thus, roughly speaking, Theorem 1 says that if the mass-energy density on V^3 is sufficiently large relative to expansion then V^3 must be compact.

We present several corollaries to Theorem 1, corresponding to several different types of spatial hypersurfaces. First, consider the special case in which V^3 is a maximal hypersurface. A hypersurface V^3 is maximal if and only if the trace of its second fundamental form is identically zero, or, equivalently, if and only if Θ vanishes on V^3 . A maximal hypersurface is an extremal of the 3-volume functional on space-time. This class of hypersurfaces has been investigated by a number of authors.¹²

Corollary 2. Consider a perfect fluid filled space-time M^4 in which the Einstein equations are satisfied. Suppose M^4 admits a maximal spatial hypersurface V^3 which is complete in the induced metric. If

(i) the 3-acceleration (in any direction) relative to each geodesic orthogonal to V^3 is nonpositive on V^3 , i.e., $\langle a(X), X \rangle \leq 0$ for all X tangent to V^3 , and

(ii) there is a point q in V^3 such that along each geodesic γ of V^3 emanating from q the condition

$$\int_0^\infty (\rho - p) d\lambda = +\infty \quad (4)$$

is satisfied then V^3 is compact.

Proof: The vanishing of Θ , (3), and condition (i) imply that the integrand in (2) is greater than or equal to $4\pi\kappa(\rho - p)$. Thus, (4) implies (2) and V^3 is compact.

The next result applies to hypersurfaces whose normal geodesics diverge in all directions.

Corollary 3: Consider a perfect fluid filled space-time M^4 in which the Einstein equations are satisfied. Suppose M^4 admits a spatial hypersurface V^3 which is complete in the induced metric. If

(i) condition (i) of Corollary 2 holds and

(ii) the 3-velocity (in any direction) relative to each geodesic orthogonal to V^3 is nonnegative; i.e., $\langle v(X), X \rangle \geq 0$ for all X tangent to V^3 and

(iii) there is a point q in V^3 such that along each geodesic γ of V^3 emanating from q the condition

$$\int_0^\infty [4\pi\kappa(\rho - p) - 3h^2] d\lambda = +\infty \quad (5)$$

is satisfied, where $h = \frac{1}{3}\Theta$ and Θ is the expansion of the normal geodesics, then V^3 is compact.

Proof: Let $X = e_1, e_2, e_3$ be an orthonormal basis of V^3 at some point along one of the geodesics emanating from q .

Introduce the notation:

$$v_\alpha = \langle v(e_\alpha), e_\alpha \rangle, \alpha = 1, 2, 3.$$

By (ii), $v_\alpha \geq 0$. A simple computation shows

$$\Theta = \sum_{\alpha=1}^3 v_\alpha,$$

so that

$$\begin{aligned} \langle v(X), X \rangle \Theta - \langle v(X), X \rangle^2 &= v_1 v_2 + v_1 v_3 \\ &\leq v_1 v_2 + v_1 v_3 + v_2 v_3 \\ &= \frac{1}{2} \left[\left(\sum_{\alpha} v_\alpha \right)^2 - \sum_{\alpha} v_\alpha^2 \right]. \end{aligned} \quad (6)$$

Now by Schwarz's inequality,

$$\begin{aligned} \frac{1}{2} \left[\left(\sum_{\alpha} v_\alpha \right)^2 - \sum_{\alpha} v_\alpha^2 \right] &\leq \frac{1}{2} \left[\left(\sum_{\alpha} v_\alpha \right)^2 - \frac{1}{3} \left(\sum_{\alpha} v_\alpha \right)^2 \right] \\ &= \frac{2}{3} \left(\sum_{\alpha} v_\alpha \right)^2 \\ &= \frac{2}{3} \Theta^2 \\ &= 3h^2. \end{aligned} \quad (7)$$

Therefore, (3), condition (i), (6), and (7) imply that the integrand in (2) is greater than or equal to $4\pi\kappa(\rho - p) - 3h^2$. Thus, (5) implies (2) and V^3 is compact.

We emphasize that the expansion term $h = \frac{1}{3}\Theta$ appearing in (5) refers to the expansion of the normal geodesics. It refers to the expansion of the fluid (even if the fluid flow is not geodesic) if and only if the fluid flow is orthogonal to V^3 . Also, Corollary 3 is still true if instead of requiring all 3-velocities to be nonnegative, we require all 3-velocities to be nonpositive.

The final corollary refers to an arbitrary spatial hypersurface on which the acceleration condition holds. At each point of V^3 , let

$$\eta = \max_{X \in B} |\langle v(X), X \rangle|.$$

Then we have the following.

Corollary 4: Consider a perfect fluid filled space-time M^4 in which the Einstein equations are satisfied. Suppose M^4 admits a spatial hypersurface V^3 which is complete in the induced metric. If

- (i) condition (i) of Corollary 2 holds and
- (ii) there is a point q in V^3 such that along each geodesic γ of V^3 emanating from q the condition

$$\int_0^\infty [4\pi\kappa(\rho - p) - 2\eta^2] d\lambda = +\infty \quad (8)$$

is satisfied then V^3 is compact.

Proof: Using the notation introduced in the proof of Corollary 3, we have

$$\begin{aligned} \langle v(X), X \rangle \Theta - \langle v(X), X \rangle^2 &= v_1 v_2 + v_1 v_3 \\ &\leq |v_1| |v_2| + |v_1| |v_3| \\ &\leq 2\eta^2. \end{aligned} \quad (9)$$

The inequality (3), condition (i), and (9) imply that the integrand in (2) is greater than or equal to $4\pi\kappa(\rho - p) - 2\eta^2$.

Thus, (8) implies (2) and V^3 is compact.

These corollaries clearly demonstrate the role of energy density in "closing up" the universe.

III. FLUID MASSES IN STATIONARY SPACE-TIMES

We assume in this section that M^4 is a stationary space-time, i.e., that M^4 admits a future timelike Killing vector field X (which, hence, generates isometries in time). If, in addition X is an irrotational vector field then M^4 is a static space-time. We also assume that M^4 is perfect fluid filled and (unless otherwise stated) that the fluid flow is rigidly rotating, i.e., that the fluid 4-velocity u is parallel to X . Let V^3 be a spatial hypersurface in M^4 . (If M^4 is static we are assured of the existence of such global spatial sections. Indeed, if X is irrotational then through each point of M^4 there passes a maximal spatial hypersurface orthogonal to X .¹³ However, even if X is not irrotational, it has been shown by Hawking¹⁴ that unless a space-time is on the verge of causality violation, it must admit global spatial sections.)

The geometry of stationary space-times is developed in Lichnerowicz¹⁵ and Landau and Lifschitz.¹⁶ Our notation is quite close to that of Landau and Lifschitz but there are minor differences. Assume local coordinates $t = x^0, x^1, x^2, x^3$ have been chosen so that the Killing field $X = \partial/\partial t$ and V^3 is defined by $t = 0$. The metric of M^4 is of the form

$$ds^2 = -h dt^2 + 2g_{0\beta} dt dx^\beta + g_{\alpha\beta} dx^\alpha dx^\beta \quad (10)$$

where $h = -g_{00}$, all metric coefficients (g_{ij}) are independent of t , Greek indices run from 1 to 3, and the summation convention is used. The coefficient $h = -g_{00}$ is globally defined since $g_{00} = \langle X, X \rangle$.

The metric on V^3 is chosen to be, not the induced metric $g_{\alpha\beta} dx^\alpha dx^\beta$, but rather the metric orthogonal to the world lines of the fluid:

$$\begin{cases} dl^2 = \gamma_{\alpha\beta} dx^\alpha dx^\beta, \\ \gamma_{\alpha\beta} = g_{\alpha\beta} + h g_\alpha g_\beta, \end{cases} \quad (11)$$

where

$$g_\alpha = g_{0\alpha}/h.$$

All metric considerations on V^3 refer to the metric dl^2 . Physically, this metric corresponds to the metric that would be used by observers co-moving with the fluid and is more naturally associated with the kinematics of the fluid.

On V^3 we can consider the local 1-form $\varphi = g_\alpha dx^\alpha$ and its exterior derivative

$$\ell = d\varphi, \quad f_{\alpha\beta} = \partial g_\beta / \partial x_\alpha - \partial g_\alpha / \partial x_\beta.$$

The 2-form ℓ has global character since it can be shown that

$$\Omega = -(\sqrt{h}/2)\ell$$

has the property that $\Omega(Y, Z)$, for each pair of vectors Y, Z tangent to W^3 , is the vorticity form evaluated on the projections of Y and Z into u^\perp , the subspaces orthogonal to the world lines.

The angular velocity vector ω on V^3 is the pseudovector associated in the usual way with the exterior form Ω on V^3 . The angular speed ω is then determined from¹⁷

$$\|\omega\|^2 = \omega^2 = \frac{1}{8} h f_{\alpha\beta} f^{\alpha\beta}.$$

The following theorem shows that the quantity $\rho - p$ must decrease rather rapidly along some geodesic in V^3 .

Theorem 5: If the spatial hypersurface V^3 is geodesically complete and if the acceleration of the fluid world lines is bounded on V^3 ,

$$\|\nabla_u u\| < b,$$

then one cannot have that, along all geodesics emanating from a given point q of V^3 ,

$$\int_0^\infty [4\pi\kappa(\rho - p) - 2\omega^2] dl = +\infty. \quad (12)$$

We remark that the condition that the 4-acceleration $\nabla_u u$ be bounded is automatically satisfied if the fluid is pressureless ($p = 0$), for the equations of motion then imply that $\nabla_u u$ vanishes.

Proof of Theorem 5: The proof consists of showing that if (12) does hold along all geodesics of V^3 emanating from q then V^3 must be compact (by, again, invoking Ambrose's result). However, Aufenkamp has shown that there can be no compact spatial hypersurfaces in a stationary fluid filled space-time.¹⁸

The Ricci tensor (R^i_j) for M^4 and the Ricci tensor ($P^{\mu\nu}$) for V^3 are related by¹⁹

$$R^{\mu\nu} = P^{\mu\nu} - (1/\sqrt{h})(\sqrt{h})^{\mu\nu} + \frac{1}{2}h f^{\mu\sigma} f^\nu_\sigma.$$

For our fluid $R^{\mu\nu} = 4\pi\kappa(\rho - p)\gamma^{\mu\nu}$. Let σ be a dl^2 geodesic of V^3 with unit tangent $v = v^\alpha \partial/\partial x^\alpha$. Then we get from the above

$$\begin{aligned} \text{Ric}_V(v,v) &\equiv P_{\alpha\beta} v^\alpha v^\beta = 4\pi\kappa(\rho - p) \\ &\quad - \frac{h}{2} f^\alpha_\alpha f_{\beta\sigma} v^\alpha v^\beta + \frac{1}{\sqrt{h}} \frac{d^2 \sqrt{h}}{dl^2}, \end{aligned} \quad (13)$$

where l is arclength along σ . Thus,

$$\begin{aligned} \text{Ric}_V(v,v) &= 4\pi\kappa(\rho - p) - 2\|v \times \omega\|^2 + \frac{d^2}{dl^2} \log \sqrt{h} \\ &\quad + \left(\frac{d}{dl} \log \sqrt{h} \right)^2. \end{aligned} \quad (14)$$

From (14) we have

$$\begin{aligned} \int_0^R \text{Ric}_V(v,v) dl &\geq \int_0^R [4\pi\kappa(\rho - p) - 2\omega^2] dl \\ &\quad + \frac{d}{dl} \log \sqrt{h} \Big|_0^R. \end{aligned} \quad (15)$$

The geodesic curvature of the world line through a point can be written as $\nabla_u u = (\nabla_u u)_V + T$, where $(\nabla_u u)_V$ is tangent to V^3 and T is along the world line. One can show that²⁰

$$(\nabla_u u)_V = \text{grad}_V \log \sqrt{h}.$$

Thus by definition of the dl^2 metric on V^3

$$\|\nabla_u u\| = \|(\nabla_u u)_V\|_V \geq \left| \frac{d}{dl} \log \sqrt{h} \right|,$$

(where $\| \cdot \|_V$ is length in the dl^2 metric) and so by hypothesis $d/dl \log \sqrt{h} \Big|_0^R$ is bounded by $2b$ on V^3 . If, for each geodesic, the integral on the right-hand side of (15) tends to infinity as $R \rightarrow \infty$ we would conclude from Ambrose's result that V^3 is compact. But by Aufenkamp's result, this is impossible.

The same analysis, in particular inequality (15), holds when the fluid is not rigidly rotating, i.e., when the 4-velocity is not necessarily along the Killing field $\partial/\partial t$. This follows from the Einstein equations, which now show that $R^{\mu\nu} v_\mu v_\nu \geq 4\pi\kappa(\rho - p)$ for the unit vector v tangent to a spatial geodesic. However, ω still represents the vorticity of the Killing vector orbits, i.e., the time lines. Maitra's cylindrically symmetric, stationary model mentioned in the introduction is an example of a model in which the fluid is not rotating rigidly. In this model the density falls off more rapidly than the inverse cube of the distance from the axis of rotation, yielding a nice illustration of Theorem 5.

A static universe is a stationary universe with vorticity zero, $\omega = 0$. One then chooses the spatial sections to be orthogonal to the time lines, i.e., $g_{0\beta} = 0$ for $\beta = 1, 2, 3$. If one adopts the usual restriction $p \leq \frac{1}{3}\rho$, we then have

Corollary 6: In a static universe with bounded acceleration of world lines, one cannot have

$$\int_0^\infty \rho dl = +\infty$$

along all V -geodesics emanating from a given point q in V^3 . (The fact that there are no compact spatial sections in a static universe follows from Green's theorem and Levi-Civita's equation: $\nabla^2_V \sqrt{h} = 4\pi\kappa(\rho + 3p)\sqrt{h}$ for a static universe.)

We now turn our attention to finite mass distributions. A static spherically symmetric ball of fluid, of mass m , in an otherwise empty universe, gives rise to an exterior Schwarzschild solution that can be joined to an interior solution only if $2m/r_0 < 1$, where r_0 is the "coordinate radius" of the ball. If one assumes that the density $\bar{\rho}$ of the fluid is a nonincreasing function of r then we must have the slightly stronger restriction $2m/r_0 < \frac{3}{8}$ if we are to avoid an infinite central pressure.²¹ If we define the average density ρ by $m = \frac{4}{3}\pi r_0^3 \kappa \bar{\rho}$ then we get an upper bound for the coordinate radius of the ball,

$$r_0 < (3\pi\kappa\bar{\rho})^{-1/2}.$$

We wish to investigate the relationship between the size of an arbitrary fluid mass and its density in a general stationary space-time. Let the setting be just as in the previous theorem. Thus, M^4 is a stationary fluid filled space-time and V^3 is a spatial hypersurface in M^4 . By an arbitrary fluid mass in M^4 we shall simply mean a bounded connected open subset \mathcal{D} of V^3 . We do not assume that the density vanishes on (or in the vicinity of) the boundary of \mathcal{D} , i.e., \mathcal{D} may be a portion of an even larger fluid mass. A nonzero vorticity ($\omega \neq 0$) on \mathcal{D} suggests that \mathcal{D} is rotating with respect to distant matter in the universe.

Recall that a ball of radius R in V^3 is the set of all points in V^3 whose distance from some given point in V^3 is less than or equal to R . The following theorem puts a bound on the size of the largest ball which a fluid mass \mathcal{D} can contain.

Theorem 7: Let V^3 be a complete spatial hypersurface and let \mathcal{D}_c be a bounded connected open subset of V^3 on which the inequality

$$4\pi\kappa(\rho - p) - 2\omega^2 \geq c \quad (16)$$

holds for some positive constant c . Suppose that

$$\|\nabla_u u\| \leq b$$

on \mathcal{D}_c . Then \mathcal{D}_c contain no ball of radius greater than

$$l_0 \equiv \pi[b + (b^2 + 2c)^{1/2}]/c.$$

Before proceeding to the proof, we would like to make a few comments. One expects (16) to be satisfied for some $c > 0$ provided the rotation is not too great or vanishes altogether (the static case). The inequality (16) is reminiscent of a condition that has arisen in classical physics. Poincaré²² had shown that a classical fluid body rotating with constant angular velocity ω must satisfy

$$2\pi\kappa\bar{\rho} - \omega^2 \geq 0,$$

where $\bar{\rho}$ is the average density of matter. A general relativistic version of this condition may be easily derived *via* the generalized Levi-Civita equation for a stationary space-time,²³

$$\nabla_{\nu}^2 \sqrt{h} = \text{Ric}(u, u)\sqrt{h} - 2\omega^2 \sqrt{h}.$$

If \mathcal{D} is a compact mass of fluid whose boundary $\partial\mathcal{D}$ is a surface across which the pressure is nonincreasing as one leaves \mathcal{D} , then since $\text{Ric}(u, u) = 4\pi\kappa(\rho + 3p)$,

$$\int_{\mathcal{D}} [4\pi\kappa(\rho + 3p) - 2\omega^2](\sqrt{h}) \, d\text{vol}_{\nu} = \int_{\partial\mathcal{D}} \text{grad}_{\nu} \sqrt{h} \, dS.$$

The general relativistic equation of hydrodynamic equilibrium gives

$$(\rho + p)\text{grad}_{\nu} \log \sqrt{h} = -\text{grad}_{\nu} p.$$

Since $\text{grad}_{\nu} p$ points inward along $\partial\mathcal{D}$, we have

$$\int_{\mathcal{D}} [4\pi\kappa(\rho + 3p) - 2\omega^2](\sqrt{h}) \, d\text{vol} \geq 0,$$

an analogue of the Poincaré condition. However, for the proof of Theorem 7 we require the stronger condition (16).

Proof of Theorem 7: Let q be any point in \mathcal{D}_c . and let B_q be a ball in V^3 centered at q having radius larger than l_0 . Let σ be a *minimizing* geodesic from q to a point on the boundary of B_q ; σ is necessarily contained in B_q . Along this geodesic, from (14),

$$\text{Ric}_{\nu} \left(\frac{d\sigma}{dl}, \frac{d\sigma}{dl} \right) \geq 4\pi\kappa(\rho - p) - 2 \left| \frac{d\sigma}{dl} \times \omega \right|^2 + \frac{d^2}{dl^2} \log \sqrt{h}. \quad (17)$$

Note that the function $f(l) = d \log \sqrt{h} / dl$ defined along σ satisfies, as we have already seen, $|f(l)| \leq b$. We now invoke the following refinement of Myer's lemma.

Lemma (Galloway²⁴): If along a geodesic σ of a Riemannian V^n there is a constant $c > 0$ and a smooth function f such that $|f| \leq b$ and

$$\text{Ric} \left(\frac{d\sigma}{dl}, \frac{d\sigma}{dl} \right) \geq c + \frac{df}{dl},$$

then σ cannot be minimizing if the length of σ is greater than

$$\pi\{b + [b^2 + (n-1)c]^{1/2}\}/c.$$

If σ were contained in \mathcal{D}_c then by (16),

$$4\pi\kappa(\rho - p) - 2 \left| \frac{d\sigma}{dl} \times \omega \right|^2 \geq c$$

along σ , and so by (17) and the preceding lemma σ would not be minimizing. Thus σ and, hence, B_q are not contained in \mathcal{D}_c .

Note that, roughly speaking, $\|d\sigma/dl \times \omega\|^2 \leq \omega^2$ allows the rotating body to "bulge out" in directions orthogonal to the "axis of rotation". Also note that if the gravitational-centrifugal acceleration in the radial direction increases as one moves out from q , i.e., if $d^2 \log \sqrt{h} / dl^2 \geq 0$, then we may apply the above argument with $b = 0$ to conclude that $\pi\sqrt{2/c}$ is an upper bound for the radius of a ball contained in \mathcal{D}_c .

Note added in proof: Recently one of the authors (G.J.G.) has obtained a generalization of the Ambrose theorem which can be used to improve some of the results presented here.

¹S. B. Myers, *Duke Math. J.* **8**, 401 (1941).

²W. Ambrose, *Duke Math. J.* **24**, 345 (1957).

³G. J. Galloway, *J. Differential Geom.* **14**, 105 (1979).

⁴See Ref. 3 and, G. J. Galloway, *J. Math. Phys.* **18**, 250 (1977).

⁵A. S. Eddington, *The Mathematical Theory of Relativity* (Cambridge U. P., Cambridge, 1957).

⁶See, for example, the discussion in T. Frankel, *Gravitational Curvature* (Freeman, San Francisco, 1979), p. 140.

⁷S. C. Maitra, *J. Math. Phys.* **7**, 1025 (1966).

⁸See Ref. 2.

⁹See Ref. 6, pp. 71–89.

¹⁰See Ref. 6, p. 41.

¹¹See the proof of the lemma on p. 161 on Ref. 6.

¹²See, for example, the article by Fisher, Marsden, and Choquet-Bruhat appearing in *Isolated Gravitating Systems in General Relativity*, edited by J. Ehlers (Italian Physical Society, 1979), pp. 396–456.

¹³See Ref. 6, p. 76.

¹⁴S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of the Universe* (Cambridge U. P., Cambridge, 1973), p. 198.

¹⁵A. Lichnerowicz, *Theories relativistes de la gravitation et de l'électromagnétisme* (Masson, Paris, 1955).

¹⁶L. Landau and E. Lifschitz, *Classical Theory of Fields*, 4th Eng. ed. (Pergamon, New York, 1975).

¹⁷See Ref. 16, p. 253.

¹⁸See the proof by Avez in *Ann. Inst. Fourier (Grenoble)*, **13** 127 (1963).

¹⁹See Ref. 16, p. 280.

²⁰This can be derived using results of Ref. 16 on p. 252.

²¹S. Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1972).

²²H. Poincaré, *Leçons sur les hypothèses cosmogoniques* (Herman, Paris, 1913).

²³See Ref. 16, p. 280.

²⁴See Ref. 3.

Asymptotically flat \mathcal{H} spaces

M. Ludvigsen

Math Department, University of York, York, England

E. T. Newman

Physics Department, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

K. P. Tod

Math Institute, Oxford University, Oxford, England

(Received 5 August 1980; accepted for publication 13 November 1980)

A definition of an asymptotically flat \mathcal{H} -space is given. Using a technique for solving for all null geodesics in an \mathcal{H} -space, it is shown that the construction of another \mathcal{H} -space from a given asymptotically flat \mathcal{H} -space is idempotent, i.e., the new \mathcal{H} -space is isometric to the first one.

PACS numbers: 04.20.Cv

I. INTRODUCTION

The theory of complex space-times has received a great deal of study in recent years for a variety of reasons. An important class of these space-times are those with a metric tensor which is holomorphic (in some domain) in the complex space-time coordinates z^a such that they satisfy the vacuum Einstein equations with a Weyl tensor that is self-(or anti-self) dual. These space-times are referred to as left (right) flat. An important subset of the left flat space-times, called \mathcal{H} -spaces, arises naturally in the study of the asymptotic shear of null surfaces in real asymptotically flat space-times. More specifically, an \mathcal{H} -space is the space of asymptotically shear free complex null cones of an asymptotically flat space-time. Several natural questions arise in the study of \mathcal{H} -spaces, namely, what is the meaning of an asymptotically flat \mathcal{H} -space, what relationship does an asymptotically flat \mathcal{H} -space have with the original real space it was obtained from, and finally what conditions must be imposed on the original real space so that the associated \mathcal{H} -space is asymptotically flat?

This paper is devoted to answering the first two questions. Though we have what seems like reasonable conjectures the last question is as yet unanswered.

In Sec. II we review the basic ideas behind \mathcal{H} -space theory and discuss the general theory of null geodesics in an \mathcal{H} -space. In Sec. III we give the definition of an asymptotically flat \mathcal{H} -space and prove that the asymptotic shear of the asymptotically flat \mathcal{H} -space is identical of that of the original real space.

II. \mathcal{H} -SPACE

Let (M, g_{ab}) be an asymptotically flat solution^{1,2} of Einstein's equations, with future null infinity \mathcal{I}^+ . In the neighborhood of \mathcal{I}^+ we will use a Bondi coordinate system $(u, r, \zeta, \bar{\zeta})$ and a conformal factor Ω such that the rescaled metric on \mathcal{I}^+ has the (degenerate) form

$$ds^2 = \frac{d\zeta d\bar{\zeta}}{2P^2}, \quad (2.1)$$

with

$$P = \frac{1}{2}(1 + \zeta\bar{\zeta}). \quad (2.2)$$

If $\sigma^0(u, \zeta, \bar{\zeta})$ is the asymptotic shear of the $u = \text{const.}$ null

surfaces then the asymptotic shear σ^0 of an arbitrary null surface which intersects \mathcal{I}^+ on the cut

$$u = Z(\zeta, \bar{\zeta}) \quad (2.3)$$

is given³ by

$$\sigma^0 = \sigma^0(Z, \zeta, \bar{\zeta}) - \delta^2 Z, \quad (2.4)$$

where σ^0 and Z are respectively spin-weight 2 and 0 quantities. The condition for the new surface to have vanishing asymptotic shear is therefore

$$\delta^2 Z = \sigma^0(Z, \zeta, \bar{\zeta}). \quad (2.5)$$

However, since σ^0 is complex and Z is real (2.5) in general has no solutions and hence, in general, there are no (asymptotically) shear free cuts (good cuts) of \mathcal{I}^+ .

If, however, we make \mathcal{I}^+ complex ($C\mathcal{I}^+$) by allowing u to assume complex values and by releasing $\bar{\zeta}$ from being the complex conjugate of ζ (and calling it $\bar{\zeta}$), the situation is different. If $\sigma^0((Z, \zeta, \bar{\zeta}))$ is real analytic in u , $\text{Re}\zeta$, and $\text{Im}\zeta$, and is sufficiently small, it can be shown⁴ that there exists a four-complex parameter set of complex solutions to (2.5), i.e., a four-complex dimensional set of good cuts given by

$$u = Z(z^a, \zeta, \bar{\zeta}), \quad (2.6)$$

which is holomorphic in the four complex parameters z^a and ζ and $\bar{\zeta}$ (near $\bar{\zeta} = \bar{\zeta}$). In other words the solution space of (2.5) defines (locally) a four-complex-dimensional manifold with local coordinates z^a which is referred to as an \mathcal{H} -space.

Though $\sigma^0(u, \zeta, \bar{\zeta})$ is (by assumption) well behaved on \mathcal{I}^+ , it will inevitably develop singularities in $C\mathcal{I}^+$ when continued too far into the complex. To avoid such difficulties, we restrict our attention to a finite complex "thickening" $C'\mathcal{I}^+$ of $C\mathcal{I}^+$, i.e., to a region of $C\mathcal{I}^+$ where $\text{Im}u$ is bounded and $\bar{\zeta}$ is close to $\bar{\zeta}$. The details of this region will emerge later.

For fixed but arbitrary values of $(\zeta, \bar{\zeta})$ in $C'\mathcal{I}^+$ the functions Z , δZ , $\delta^2 Z$, and $\delta\delta Z$ are scalar fields on \mathcal{H} and hence their gradients $Z_{,a}$, $\delta Z_{,a}$, $\delta^2 Z_{,a}$, and $\delta\delta Z_{,a}$ are covector fields on \mathcal{H} . In Refs. 4 and 5 it has been shown that \mathcal{H} is naturally endowed with a nondegenerate (complex) holomorphic metric

$$ds^2 = g_{ab}(z^c) dz^a dz^b \quad (2.7)$$

which satisfies (and can actually be uniquely defined by) the two conditions

$$Z_a Z_b g^{ab} = Z_a Z^a = 0 \quad (2.8)$$

$$\delta Z_a \delta Z_b g^{ab} = \delta Z_a \tilde{Z}^a = -1 \quad (2.9)$$

for ζ and $\tilde{\zeta}$ in $C^1 \mathcal{S}^+$ and for each z^a . Condition (2.8) determines g_{ab} up to conformal factor and (2.9) determines the factor. Eqs. (2.8) and (2.9) immediately imply

$$Z_a \delta Z^a = Z_a \tilde{\delta Z}^a = 0 \quad (2.10)$$

and

$$Z_a \delta \tilde{\delta Z}^a = 1. \quad (2.11)$$

From the metric (2.7) it can be shown^{4,5} that the \mathcal{H} -space has the following curvature properties:

$$R_{ab} = 0, \quad (2.12)$$

$$W_{abcd} = 0, \quad (2.13)$$

and in general

$$\tilde{W}_{abcd} \neq 0,$$

where W_{abcd} and \tilde{W}_{abcd} are the anti-self-dual and self-dual parts of the Weyl tensor C_{abcd} . Associated with the vanishing of W_{abcd} is the fact that the anti-self-dual bivector $Z_{[a} \delta Z_{b]}$ is covariantly constant, i.e.,

$$\nabla_c (Z_{[a} \delta Z_{b]}) = 0, \quad (2.14)$$

from which it follows that $\nabla_a Z_b$ has the form

$$\nabla_b Z_a = \alpha Z_a Z_b + 2\beta Z_{(a} \delta Z_{b)} + \gamma \delta Z_a \delta Z_b \quad (2.15)$$

Using (2.8) and (2.10), we have from (2.15)

$$Z^a \nabla_a Z_b = 0 \quad \text{and} \quad \delta Z^a \nabla_a Z_b = 0, \quad (2.16)$$

which will be needed later.

As mentioned earlier, for fixed but arbitrary $\zeta, \tilde{\zeta}$,

$$\begin{aligned} u &= Z(z^a, \zeta, \tilde{\zeta}), & r &= \delta \delta Z, \\ w &= \delta Z, & \tilde{w} &= \tilde{\delta Z}, \end{aligned} \quad (2.17)$$

are four independent scalar functions on \mathcal{H} and thus could be thought of as forming a natural coordinate system (parametrized by the $\zeta, \tilde{\zeta}$) on \mathcal{H} . Since the system of scalars will form an important element in our study of asymptotic flatness, we now investigate its properties.^{6,7}

Since, from (2.8), $Z_a Z^a = 0$, the points of \mathcal{H} defined by $u = \text{const}$ form a null hypersurface whose generators are null geodesics and whose tangent vectors are given by $Z^a = g^{ab} Z_b$. Either directly from (2.16) or from the fact that $Z_a = Z_a$ it follows that Z^a is affinely parametrized and can be written

$$Z^a = \frac{dz^a}{d\lambda}, \quad (2.18)$$

with λ affine. Calculating $dr/d\lambda$ from (2.17), we have

$$\frac{dr}{d\lambda} = \delta \delta Z_a \cdot \frac{dz^a}{d\lambda} = \delta \delta Z_a \cdot Z^a, \quad (2.19)$$

and from (2.11) $dr/d\lambda = 1$. Thus r is an affine parameter. Furthermore, from (2.10) it is immediately seen that w and \tilde{w} are constant along the generators and hence (on a $u = \text{const}$ surface) can be used to label them.

What we have just shown is that the equations for any null geodesic $z^a = z^a(r)$ can be written implicitly (with $\zeta, \tilde{\zeta}$ constant) as

$$Z(z^a, \zeta, \tilde{\zeta}) = u = \text{const}, \quad \delta Z(z^a, \zeta, \tilde{\zeta}) = \omega = \text{const},$$

$$\tilde{\delta Z}(z^a, \zeta, \tilde{\zeta}) = \tilde{\omega} = \text{const}, \quad r = \delta \delta Z(z^a, \zeta, \tilde{\zeta}).$$

By introducing the special null tetrad system $(l^a, n^a, m^a, \tilde{m}^a)$, with $l \cdot n = -m \cdot \tilde{m} = 1$ and all other products vanishing, with

$$l_a = Z_a \quad \text{and} \quad m_a = \delta Z_a, \quad (2.20)$$

the (complexified) optical scalars

$$\rho = m^a \tilde{m}^b \nabla_a l_b, \quad \tilde{\rho} = \tilde{m}^a m^b \nabla_a l_b \quad (2.21)$$

associated with the $u = Z = \text{const}$ surfaces are seen, from (2.16) and the fact that Z_a is a gradient, to vanish. The $Z = \text{const}$ surfaces are therefore divergence free and resemble null hyperplanes.

In order to gain further insight into the nature of the scalars $(u, r, \omega, \tilde{\omega})$ we consider the trivial case of an \mathcal{H} -space arising from a σ^0 which vanishes, e.g., the \mathcal{H} -space of flat space-time. In this case the general solution of (2.5) is

$$Z = z^a L_a(\zeta, \tilde{\zeta}), \quad (2.22)$$

where $z^a = (t, x, y, z)$ are four-complex parameters and the four functions L_a are given by

$$L_a = \frac{1}{\sqrt{2}} \left(1, \frac{\zeta + \tilde{\zeta}}{1 + \zeta \tilde{\zeta}}, i \frac{\zeta - \tilde{\zeta}}{1 + \zeta \tilde{\zeta}}, \frac{-1 + \zeta \tilde{\zeta}}{1 + \zeta \tilde{\zeta}} \right). \quad (2.23)$$

The metric in this case is flat, and the z^a form a Minkowski coordinate system. For an arbitrary but fixed $\zeta = \zeta_0$ and $\tilde{\zeta} = \tilde{\zeta}_0$ (note we are now restricting ourselves to the real sphere $\zeta = \tilde{\zeta}$) we have

$$\begin{aligned} u &= z^a L_a(\zeta_0, \tilde{\zeta}_0), \\ r &= z^a \delta \delta L_a(\zeta_0, \tilde{\zeta}_0), \end{aligned} \quad (2.24)$$

$$\omega = z^a \delta L_a(\zeta_0, \tilde{\zeta}_0),$$

$$\tilde{\omega} = z^a \tilde{\delta} L_a(\zeta_0, \tilde{\zeta}_0).$$

Slightly abusing the term, we will refer to "real" values of the set $(u, r, \omega, \tilde{\omega})$ when u and r are real and $\tilde{\omega} = \bar{\omega}$. It is not hard to show that "real" values for $(u, r, \omega, \tilde{\omega})$ in (2.24) imply real values for z^a as well as the converse, real values for z^a imply "real" values for $(u, r, \omega, \tilde{\omega})$. [In particular for $\zeta_0 = \tilde{\zeta}_0 = 0$, (2.24) yields

$$\begin{aligned} u &= (1/\sqrt{2})(t - z), \\ r &= (\sqrt{2})z, \\ \omega &= (1/\sqrt{2})(x - iy), \\ \tilde{\omega} &= (1/\sqrt{2})(x + iy), \end{aligned} \quad (2.25)$$

which is seen to be the conventional null plane coordinates. For other values of $(\zeta_0, \tilde{\zeta}_0)$ we would simply have rotated versions of (2.25).] Note that in this case the good cut function $Z(z^a, \zeta, \tilde{\zeta})$ for real values of z^a remains entirely in the real \mathcal{S}^+ .

It appears likely that a similar (though slightly more general) situation to the $\sigma^0 = 0$ case occurs for a wide class of \mathcal{H} -spaces, i.e., for a wide class of σ^0 : Namely for σ^0 in this class there is defined a subset \mathcal{H}' of \mathcal{H} that is a finite complex thickening of a real 4-manifold, R^4 . \mathcal{H}' is defined as those points of \mathcal{H} for which the set $(u, r, \omega, \tilde{\omega})$ is "real" for

some $(\zeta, \bar{\zeta})$. (For other values of $\zeta, \bar{\zeta}$ at the same point, the set would not in general be real. The $\sigma^0 = 0$ case is the exception.) From this it would follow that the good cut function $Z(z^a, \zeta, \bar{\zeta})$ for z^a in \mathcal{H}' defines a finite complex thickening $C^+\mathcal{I}$ of \mathcal{I} .

III. ASYMPTOTICALLY FLAT \mathcal{H} -SPACE

In this section we first give a definition of asymptotic flatness appropriate to \mathcal{H} -space and then show that, if \mathcal{H} is asymptotically flat and \mathcal{H}^* is the \mathcal{H} -space constructed from the asymptotic shear of \mathcal{H} , then \mathcal{H} and \mathcal{H}^* are naturally isometric.

Definition: \mathcal{H} will be said to be asymptotically flat if:

(1) On the region \mathcal{H}' , defined above, the functions

$$u = Z(z^a, \zeta, \bar{\zeta}), \quad \omega = \delta Z(z^a, \zeta, \bar{\zeta})$$

$$r = \delta\bar{\delta}Z(z^a, \zeta, \bar{\zeta}), \quad \bar{\omega} = \bar{\delta}Z(z^a, \zeta, \bar{\zeta})$$

are holomorphic in z^a for each ζ and have the following properties;

(a) $|\text{Im}u|$, $|\text{Im}r|$, and $|\bar{\omega} - \omega|$ are uniformly bounded for all points z^a in \mathcal{H}' and all $(\zeta, \bar{\zeta})$, and in addition vanish for some $(\zeta_0, \bar{\zeta}_0)$ which depends on the point z^a .

(b) As z^a ranges over the whole of \mathcal{H}' , $\text{Re}u$, and $\text{Re}r$ range over the whole of the real line and ω ranges over the complex plane.

(2) A "real" null geodesic is one with "real" tangent vector $dz^a/d\lambda = Z^a(z^a, \zeta, \bar{\zeta})$ for some ζ and real increments in r [see Eq. (2.17) *et seq.*]. We define \mathcal{H}'' as the set of points of \mathcal{H} reached by real null geodesics from \mathcal{H}' and we demand that the functions u, r, ω , and $\bar{\omega}$ be holomorphic at a point in \mathcal{H}'' for all $\zeta, \bar{\zeta}$. (It should be noted that since on null geodesics $u, \omega, \bar{\omega}$ are constant the points of \mathcal{H}'' are determined by the range of $u, \omega, \bar{\omega}$ on \mathcal{H}' as $\zeta, \bar{\zeta}$ varies.)

(3) There exists a nonzero holomorphic function Ω on \mathcal{H}'' such that:

(a) There is a larger space $(\hat{\mathcal{H}}'', \hat{g}_{ab})$ such that \mathcal{H}'' is diffeomorphic to a region U of $\hat{\mathcal{H}}''$, with $\hat{g}_{ab} = \Omega^2 g_{ab}$ on U and $\Omega = 0$ on $\hat{\mathcal{H}}'' - U \equiv C^+\mathcal{I}(\mathcal{H})$. On $C^+\mathcal{I}(\mathcal{H})$, $\nabla_a \Omega \neq 0$, $\hat{g}^{ab} \nabla_a \Omega \nabla_b \Omega = 0$, and \hat{g}_{ab} is holomorphic.

(b) If $z^a = z^a(r)$ is the equation of a null geodesic in \mathcal{H}'' affinely parametrized by r and given by u, ω , and $\bar{\omega}$ constant, then $\lim_{r \rightarrow \infty} \Omega[z^a(r)] \cdot r$ exists and is nonvanishing as $\text{Re}r \rightarrow \infty$. The limits in the positive and negative $r = \delta\bar{\delta}Z$ directions define $C^+\mathcal{I}$ and $C^-\mathcal{I}$.

Condition 1 says that \mathcal{H}' is a complex thickening of R^4 . Condition 2 says that the null geodesic $z^a = z^a(r)$ specified by

$$u = u_0 = Z(z_0^a, \zeta_0, \bar{\zeta}_0),$$

$$\omega = \omega_0 = \mathcal{H}Z(z_0^a, \zeta_0, \bar{\zeta}_0),$$

$$\bar{\omega} = \bar{\omega}_0 = \bar{\mathcal{H}}Z(z_0^a, \zeta_0, \bar{\zeta}_0),$$

[i.e., the null geodesic through some fixed point z_0^a in \mathcal{H}' with direction $Z^a(z_0^a, \zeta_0, \bar{\zeta}_0)$ remains entirely within \mathcal{H}'' if r is sufficiently close to the real line. Thus through each point of \mathcal{H}'' there exists at least a spheres worth (as $\zeta, \bar{\zeta}$ varies) of null geodesics which escape to $C^+\mathcal{I}(\mathcal{H})$.

Condition 3, which is essentially a complex version of Penrose's conformal asymptotic condition, states that com-

plex (future) null infinity ($C^+\mathcal{I}(\mathcal{H})$) exists for \mathcal{H}' .

That these conditions are not empty is shown by the example of the Sparling-Tod \mathcal{H} -spaces.⁸

We now investigate some of the consequences of these conditions.

We wish to first show that the $(\zeta, \bar{\zeta})$ in the scalar $u = Z(z^a, \zeta, \bar{\zeta})$ can be used to parametrize the null generators of $C^+\mathcal{I}(\mathcal{H})$ and that the u parametrizes the points on each generator. We now state the basic lemma which is proved in Appendix A.

Lemma 1: If N_u is a one parameter set of divergence free ($\rho = \bar{\rho} = 0$) null hypersurfaces in \mathcal{H}' given by $u = \text{const} = Z(z^a, \zeta_0, \bar{\zeta}_0)$ (fixed ζ_0), then, when viewed from $\hat{\mathcal{H}}''$, the null generators of each N_u converge to distinct points (parametrized by u) which all lie on the same null generator of $C^+\mathcal{I}(\mathcal{H})$. (See Fig. 1.)

Due to the analyticity of Z we can slightly generalize this by allowing pairs $(\zeta, \bar{\zeta})$ instead of $(\zeta_0, \bar{\zeta}_0)$ where $\bar{\zeta}$ is close to $\bar{\zeta}_0$.

The importance of this lemma lies in its use in defining a natural coordinate system $(u, \zeta, \bar{\zeta})$ constructed on $C^+\mathcal{I}(\mathcal{H})$. Eventually we will show that it is a Bondi system.

An important property of spaces satisfying $W_{abcd} = R_{ab} = 0$ is that for any null surface which has a vanishing left-shear $\sigma^{9,10}$ at any point of a generator, the shear must vanish along the entire generator. This can be seen from the spin coefficient equation

$$\frac{d\sigma}{d\lambda} = 2\rho\sigma + \psi_0, \quad (3.1)$$

where λ is a affine parameter along the generator and ψ_0 is a component of the anti-self-dual Weyl tensor and hence vanishes when $W_{abcd} = 0$. Thus, if $\sigma = 0$ for some value of λ then by (3.1) $\sigma = 0$ for all λ . In particular the complex null cone $C(p)$, whose vertex is some point p in \mathcal{H}' , will be left-shear free for all points of $C(p)$ since σ vanishes at the vertex on all generators. (In general $\bar{\sigma}$ will be nonzero apart from its zero value at p .) Thus $C(p)$ will intersect $C^+\mathcal{I}(\mathcal{H})$ in a shear-free or good cut of $C^+\mathcal{I}(\mathcal{H})$. Conversely, using (3.1) and

$$\frac{d\rho}{d\lambda} = \rho^2 + \sigma\bar{\sigma}, \quad (3.2)$$

a second spin-coefficient equation,¹⁰ one can easily see that a good cut of $C^+\mathcal{I}(\mathcal{H})$ gives rise to a null hypersurface in \mathcal{H}''

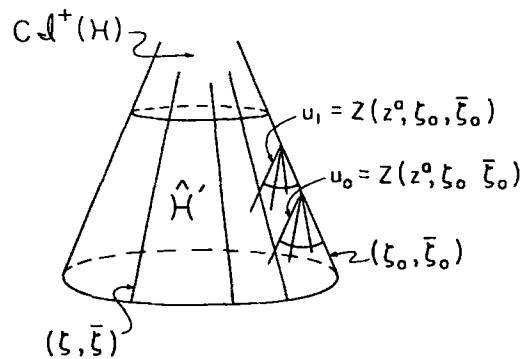


FIG. 1. The hypersurfaces $u_0 = Z(z^a, \zeta_0, \bar{\zeta}_0)$ and $u_1 = Z(z^a, \zeta_0, \bar{\zeta}_0)$ in \mathcal{H}'' are null cones whose vertices lie on the same generator of $C^+\mathcal{I}(\mathcal{H})$.

which converges to a unique point of \mathcal{H}^* and is hence a $C(p)$.

Thus we have

Lemma 2: On an asymptotically flat \mathcal{H}^* -space there exists a natural one to one mapping between the set of good cuts on $C\mathcal{I}^+(\mathcal{H})$ (i.e., the points of \mathcal{H}^*) and points of \mathcal{H}^* , the mapping being defined by the intersections of $C(p)$ and $C\mathcal{I}^+(\mathcal{H})$.

We now state and prove our main result:

Theorem:

(1) The (natural) coordinates on $C\mathcal{I}^+(\mathcal{H})$ introduced in Lemma 1 are Bondi coordinates.

(2) The asymptotic shear of the \mathcal{H} -space associated with these Bondi coordinates is the same as the asymptotic shear of the original real asymptotically flat space.

(3) The \mathcal{H}^* space constructed from the asymptotic shear of the \mathcal{H} space is isometric to the \mathcal{H} -space, i.e., \mathcal{H} of $\mathcal{H} = \mathcal{H}$.

Part (3) follow immediately from (2) since an \mathcal{H} -space is uniquely defined from the asymptotic shear in (2.5).

Assuming for the moment that part (1) is correct, part (2) follows immediately from Lemma 2; $u = Z(z^a, \zeta, \bar{\zeta})$ is the asymptotically shear-free cut of $C\mathcal{I}^+$, and by inverting the argument which led to (2.5) we have that the asymptotic shear of the $u = \text{const}$ cuts of $C\mathcal{I}^+(\mathcal{H})$ is $\sigma^0 = \delta^2 Z$. The basic idea for the proof of (1) is to take a world line in \mathcal{H} , i.e., $x^a = x^a(\tau)$, and construct the null cones centered on the line and then use the directions $(\zeta, \bar{\zeta})$ of the null lines, the affine length R along the null lines, and τ as the coordinates of a null (NU) coordinate system. One then calculates the conformally rescaled metric and easily finds the transformation on $C\mathcal{I}^+(\mathcal{H})$ to the Bondi coordinates. The actual construction is as follows

We assume that $Z(z^a, \zeta, \bar{\zeta})$ and hence $g_{ab}(z^a)$ are known. We wish to find the coordinate transformation

$$z^a = z^a(\tau, R, \zeta, \bar{\zeta}). \quad (3.3)$$

From the results of Sec. II, we have the following four implicit relationships which are equivalent to (3.3):

$$Z(z^a, \zeta, \bar{\zeta}) = Z(x^a(\tau), \zeta, \bar{\zeta}) \equiv Z^0, \quad (3.4a)$$

$$\delta Z(z^a, \zeta, \bar{\zeta}) = \delta Z^0, \quad (3.4b)$$

$$\bar{\delta} Z(z^a, \zeta, \bar{\zeta}) = \bar{\delta} Z^0, \quad (3.4c)$$

$$\delta\bar{\delta} Z(z^a, \zeta, \bar{\zeta}) = R + \delta\bar{\delta} Z^0. \quad (3.4d)$$

By differentiating implicitly each of (3.4) with respect to $R, \zeta, \bar{\zeta}$ and τ , sixteen equations are obtained from which $\partial z^a / \partial R, \partial z^a / \partial \zeta, \partial z^a / \partial \bar{\zeta}$, and $\partial z^a / \partial \tau$ can be explicitly found. For example, if each of (3.4) is differentiated with respect to R , we have

$$\begin{aligned} Z_{,a} \frac{\partial z^a}{\partial R} &= 0, & \delta Z_{,a} \frac{\partial z^a}{\partial R} &= 0, \\ \bar{\delta} Z_{,a} \frac{\partial z^a}{\partial R} &= 0, & \delta\bar{\delta} Z_{,a} \frac{\partial z^a}{\partial R} &= 1. \end{aligned} \quad (3.5)$$

Using the known scalar products (Appendix B) between $Z_{,a}, \delta Z_{,a}$, and $\delta\bar{\delta} Z_{,a}$, we have

$$\frac{\partial z^a}{\partial R} = Z^a. \quad (3.6)$$

Continuing this process we obtain

$$\frac{\partial z^a}{\partial \zeta} = R \delta Z^a P^{-1}, \quad (3.7)$$

$$\begin{aligned} \frac{\partial z^a}{\partial \bar{\zeta}} &= [(-\delta X + R \delta \mathcal{F}) Z^a \\ &+ (X - 2\mathcal{F}R) \delta Z^a + R \bar{\delta} Z^a] P^{-1} \end{aligned} \quad (3.8)$$

$$\begin{aligned} \frac{\partial z^a}{\partial \tau} &= (\delta \bar{\delta} V - \delta V \delta \mathcal{F} + V \mathcal{E}) Z^a \\ &+ (2\delta V \mathcal{F} - V \delta \mathcal{F} - \bar{\delta} V) \delta Z^a \\ &- \delta V \bar{\delta} Z^a + V \delta \bar{\delta} Z^a, \end{aligned} \quad (3.9)$$

where

$$V = \frac{\partial Z^0}{\partial \tau} = Z^0_{,a} \frac{dx^a}{d\tau} \quad (3.10)$$

and (see Appendix B)

$$\mathcal{F} = \frac{1}{2} \delta^2 Z_{,a} Z^a = \frac{1}{2} \frac{\partial X}{\partial R}, \quad (3.11a)$$

$$\mathcal{E} = \delta \bar{\delta} Z_{,a} \delta \bar{\delta} Z^a, \quad (3.11b)$$

$$X = \bar{\delta}^2 (Z - Z^0). \quad (3.11c)$$

Equations (3.6)–(3.9) are essentially the transformation matrix between the coordinates z^a and $z'^a = (\tau, R, \zeta, \bar{\zeta})$ and hence in the new coordinates the metric becomes

$$g'_{ab} = \frac{\partial z^c}{\partial z'^a} \frac{\partial z^d}{\partial z'^b} g_{cd}. \quad (3.12)$$

In Appendix B this is written out explicitly. The new radial coordinate $r = VR$ puts the metric into the NU form. Alternatively, we may conformally rescale the metric with conformal factor $\Omega = R^{-1}$ and substitute $\hat{r} = R^{-1}$ to obtain near $C\mathcal{I}^+(\mathcal{H})$,

$$d\hat{s}^2 = \Omega^2 ds^2 = -2Vd\hat{r}d\hat{r} - (1/2P^2) d\zeta d\bar{\zeta} + O(\hat{r}), \quad (3.13)$$

provided the following quantities are $O(\hat{r})$:

$$\hat{r}X, \hat{r}^2 \delta X, F, \hat{r} \delta F, \text{ and } \hat{r}^2 \delta^2 F. \quad (3.14)$$

(These are necessary consequences of condition 2a and presumably result from appropriate conditions on σ^0 .) Finally, to put (3.13) in Bondi coordinates, we introduce¹¹ u_B by

$$\begin{aligned} du_B &= Vd\tau \text{ at } C\mathcal{I}^+(\mathcal{H}), \\ \text{so that } d\hat{s}^2 &= -2du_B dr - (1/2P^2) d\zeta d\bar{\zeta} \text{ at } C\mathcal{I}^+(\mathcal{H}). \text{ This} \\ &\text{integrates to give} \\ u_B &= Z(x^a(\tau), \zeta, \bar{\zeta}) + \alpha(\zeta, \bar{\zeta}), \end{aligned} \quad (3.15)$$

where α is an arbitrary additive function (supertranslation) which we may set to zero. Thus the u coordinate of Lemma 1 is indeed a Bondi coordinate on $C\mathcal{I}^+(\mathcal{H})$ and we may drop the subscript B . This proves part 1 of the theorem.

To investigate $C\mathcal{I}^-(\mathcal{H})$, we repeat the whole procedure but with past-pointing null geodesics. That is, we replace (3.6) by

$$\frac{\partial z^a}{\partial R}(R) = Y^a(z^b(R), \zeta_0, \bar{\zeta}_0),$$

where $Y^a = -Z^a$.

This will lead to

$$d\bar{s}^2 = 2V'd\tau d\bar{r} - (1/2P^2) d\bar{\xi} d\bar{\zeta} \quad \text{on } C\mathcal{I}^-(\mathcal{H}),$$

where $V' = (dx^a/d\tau) Y_a(x^b(\tau), \bar{\xi}, \bar{\zeta})$.

Introducing a Bondi v coordinate by

$$dv = V'd\tau$$

or

$$v = Y(x^b(\tau), \bar{\xi}, \bar{\zeta}),$$

we find the asymptotic shear of the Bondi system on $C\mathcal{I}^-(\mathcal{H})$ to be

$$\begin{aligned} \Sigma^0(v, \bar{\xi}, \bar{\zeta}) &= \bar{\delta}^2 Y(x^b(\tau), \bar{\xi}, \bar{\zeta}) \\ &= -\bar{\delta}^2 Z(x^b(\tau), \bar{\xi}, \bar{\zeta}) \\ &= -\sigma^0(Z[x^b(\tau), \bar{\xi}, \bar{\zeta}], \bar{\xi}, \bar{\zeta}) \\ &= -\sigma^0(-Y[x^b(\tau), \bar{\xi}, \bar{\zeta}], \bar{\xi}, \bar{\zeta}), \end{aligned}$$

i.e.,

$$\Sigma^0(v, \bar{\xi}, \bar{\zeta}) = -\sigma^0(-v, \bar{\xi}, \bar{\zeta}). \quad (3.16)$$

Thus there is a very simple relation between the Bondi shears at $C\mathcal{I}^-$ and $C\mathcal{I}^+$. We may interpret this as the statement that the classical S matrix for the self-dual Einstein equations is the identity. This is also a reflection of Huyghen's Principle and demonstrates the solitonlike behavior of the self-dual Einstein equations.

However, $C\mathcal{I}^+$ and $C\mathcal{I}^-$ are not identical in all respects. Denote the right asymptotic shear in a Bondi frame on $C\mathcal{I}^+$ (respectively $C\mathcal{I}^-$) by $\bar{\sigma}_+$ (respectively $\bar{\sigma}_-$). Then in a self-dual space-time these are independent of u , i.e., are functions of ξ and $\bar{\xi}$ only) but the two functions are different in general. To see this, we remark that it can be shown that, with the conditions of Sec. 3, $\bar{\psi}_2^0 \rightarrow 0$ as $u \rightarrow \infty$ on $C\mathcal{I}^+$, $\bar{\psi}_2^0 \rightarrow 0$ as $u \rightarrow -\infty$ on $C\mathcal{I}^-$, and $\bar{\sigma}^0 \rightarrow 0$ and σ^0 has finite limits as $u \rightarrow \pm \infty$. Now one of the Bianchi identities at \mathcal{I} gives

$$\bar{\psi}_2^0 = \bar{\delta}^2 \bar{\sigma}^0 - \bar{\delta}^2 \sigma^0 - \bar{\sigma}^0 \sigma^0.$$

Thus $\bar{\sigma}_+$ is determined by the limit of σ^0 as $u \rightarrow \infty$:

$$\bar{\delta}^2 \bar{\sigma}_+^0 = \lim_{u \rightarrow +\infty} \bar{\delta}^2 \sigma^0$$

while $\bar{\sigma}_-$ is determined by the limit of σ^0 as $u \rightarrow -\infty$:

$$\bar{\delta}^2 \bar{\sigma}_-^0 = \lim_{u \rightarrow -\infty} \bar{\delta}^2 \sigma^0,$$

and these two limits will be different for a σ^0 which comes from a generic radiating space-time.

$C\mathcal{I}^+$ and $C\mathcal{I}^-$ are therefore distinguished by having identical σ^0 but different $\bar{\sigma}^0$.

IV. DISCUSSION

From the results in the last section we saw that the asymptotic shear σ^0 of the \mathcal{H} -space Bondi cones is identical to that of the original space. This implies that the radiation parts (r^{-1} and r^{-2}) of the self-dual part of the original Weyl tensor agree exactly with the r^{-1} and r^{-2} parts of the \mathcal{H} -space Weyl tensor. In other words the \mathcal{H} -space construction reproduces (in the radiation zone) the self-dual part of the original Weyl tensor and factors out the anti-self-dual part.

The longitudinal parts are also factored out.

From the general theory of asymptotically flat spaces and the fact that \mathcal{H} -spaces are self-dual, one has immediately that the right asymptotic shear $\bar{\sigma}^0$ of the Bondi cuts must have the property that $\partial \bar{\sigma}^0 / \partial u = 0$, and hence by an appropriate supertranslation we can have $\bar{\sigma}^0 = 0$. Thus if one constructs the $\hat{\mathcal{H}}$ -space from the right shear of an \mathcal{H} -space, i.e., $\hat{\mathcal{H}}$ of \mathcal{H} , the result would be to factor out the other part (self-dual) of the Weyl tensor and produce Minkowski space.

The main remaining question about asymptotically flat \mathcal{H} -spaces is, What are the conditions on σ^0 in (2.5) so that the solutions $Z(z^a, \bar{\xi}, \bar{\zeta})$ determine an asymptotically flat $\hat{\mathcal{H}}$ -space? Though there are several reasonable approaches to this question it has largely remained intractable. Arguments, however, from linear theory indicate that (mod supertranslation freedom) minimum conditions should be $\sigma^0 = O(u^{-3})$ as $u \rightarrow \infty$.

APPENDIX A

We will here prove the lemma: If N_u is a one parameter set of divergence-free (i.e., $\rho = \bar{\rho} = 0$) null hypersurfaces in \mathcal{H}' (which is asymptotically flat) given by $u = \text{const} = Z(z^a, \bar{\xi}_0, \bar{\zeta}_0)$, (for fixed $\bar{\xi}_0$) then, when viewed from $\hat{\mathcal{H}}'$, the null generators of each N_u converge to distinct points (parametrized by u) which all lie on the same null generator of $C\mathcal{I}^+(\mathcal{H})$.

Suppose some null generator of N_u (fixed u) intersects $C\mathcal{I}^+(\mathcal{H})$ at p . In the neighborhood of p a null tetrad system in $\hat{\mathcal{H}}'$ ($\hat{l}^a, \hat{n}^a, \hat{m}^a, \hat{\bar{m}}^a$) can be formed so that \hat{l}^a are the affinely parametrized tangent vectors to N_u and $\hat{n}_a = \hat{\nabla}_a \Omega$. Since $\hat{g}_{ab} = \Omega^2 g_{ab}$ with $\Omega \sim r^{-1}$ (from Condition 2b), a corresponding null tetrad system can be defined in \mathcal{H}' by

$$\begin{aligned} l^a &= \Omega^2 \hat{l}^a, & n^a &= \hat{n}^a, \\ m^a &= \Omega \hat{m}^a, & \bar{m}^a &= \Omega \hat{\bar{m}}^a, \\ l_a &= \hat{l}_a. \end{aligned} \quad (A1)$$

From the relationship between the covariant derivatives ∇_a and $\hat{\nabla}_a$ we have

$$\nabla_a \Omega = \hat{\nabla}_a \Omega$$

and

$$\nabla_a l_b = \hat{\nabla}_a \hat{l}_b + 2\Omega^{-1} \hat{l}_{[a} \hat{\nabla}_{b]} \Omega - \Omega \hat{g}_{ab} \hat{l}^c \nabla_c \Omega.$$

Since $\rho = \bar{\rho} = m^a \bar{m}^b \nabla_a l_b = 0$, we have in the neighborhood of p

$$\begin{aligned} 0 &= \hat{m}^a \hat{\bar{m}}^b (\hat{\nabla}_a \hat{l}_b + 2\Omega^{-1} \hat{l}_{[a} \hat{\nabla}_{b]} \Omega - \Omega^{-1} \hat{g}_{ab} \hat{l}^c \hat{\nabla}_c \Omega) \\ &= \hat{\rho} + \hat{m}^a \hat{\bar{m}}^b (2\Omega^{-1} \hat{l}_{[a} \hat{n}_{b]} - \Omega^{-1} \hat{g}_{ab} \hat{l}^c \hat{n}_c) \\ &= \hat{\rho} + \Omega^{-1}. \end{aligned}$$

Therefore, $\hat{\rho} = -1/\Omega$ and if we use $\Omega = 1/r$ [which makes Ω an affine parameter for $(\mathcal{H}', \hat{g}_{ab})$] then we have the condition for ρ to be the vertex of a null cone and hence the null generators of N_u in $\hat{\mathcal{H}}'$ converge to p .

Two surfaces of the set N_u , i.e., N_{u_0} and N_{u_1} ($u_0 \neq u_1$), do not intersect in \mathcal{H}' (since $u, r, \omega, \bar{\omega}$ form a covering of \mathcal{H}') hence the connecting vector between two neighboring ones has the form

$$k^a = \alpha n^a + \beta \bar{m}^a + \gamma l^a,$$

where $\alpha \neq 0$. (If $\alpha = 0$ the surfaces would intersect since l^a, m^a, \tilde{m}^a are tangent to a surface.) From (A.1) we thus have in $\hat{\mathcal{H}}'$ near $C\mathcal{F}^+(\mathcal{H})$

$$k^a = \alpha \hat{n}^a + \tilde{\beta} \Omega \hat{m}^a + \tilde{\beta} \Omega \tilde{m}^a + \gamma \Omega^2 \hat{l}^a,$$

and hence as $\Omega \rightarrow 0$

$$k^a = \alpha n^a.$$

Since n^a points along the null generators of $C\mathcal{F}^+(\mathcal{H})$, N_{u_0} and N_{u_1} intersect the same generator at distinct points.

APPENDIX B

We give for completeness the scalar products between the four gradients $Z_{,a}, \partial Z_{,a}, \tilde{\partial} Z_{,a}$, and $\partial \tilde{\partial} Z_{,a}$:

$$\begin{aligned} Z^a Z_a &= 0, & \partial Z^a \partial Z_a &= 0, & \tilde{\partial} Z^a \tilde{\partial} Z_a &= -2\mathcal{F}, \\ Z^a \partial Z_a &= 0, & \partial Z^a \tilde{\partial} Z_a &= -1, & \tilde{\partial} Z^a \partial \tilde{\partial} Z_a &= -\partial \mathcal{F}, \\ Z^a \tilde{\partial} Z_a &= 0, & \partial Z^a \partial \tilde{\partial} Z_a &= 0, & \tilde{\partial} Z^a \partial \tilde{\partial} Z_a &= \mathcal{E}, \\ Z^a \partial \tilde{\partial} Z_a &= 1, \end{aligned} \quad (\text{B1})$$

where \mathcal{F} is obtained from

$$Z^a \tilde{\partial}^2 Z_a = 2\mathcal{F} \quad (\text{B2})$$

and

$$\mathcal{E} = -\partial^2 \mathcal{F} + 2\mathcal{F} \partial^0 - 2. \quad (\text{B3})$$

Knowledge of the metric determines (B1) and knowledge of (B1) determines the metric. \mathcal{F} (in B2) can be determined⁵ algebraically from $Z(z^a, \zeta, \tilde{\zeta})$ and its derivatives.

The metric expressed in the coordinates $(\tau, R, \zeta, \tilde{\zeta})$ of (3.4) is

$$\begin{aligned} ds^2 &= A d\tau^2 + 2V d\tau dR + 2RV_{,\zeta} d\tau d\zeta + 2B d\tau d\tilde{\zeta} \\ &\quad - (R^2/2P^2) d\zeta d\tilde{\zeta} - (R/2P^2)(X - R\mathcal{F}) d\tilde{\zeta}^2, \end{aligned} \quad (\text{B4})$$

where

$$\begin{aligned} A &= 2[V^2 + V\partial\tilde{\partial}V - \partial V\tilde{\partial}V + 2F(\partial V)^2 - V\partial V\partial\mathcal{F} \\ &\quad + V^2(\partial^2\mathcal{F} - 2\mathcal{F}\partial)], \end{aligned} \quad (\text{B5})$$

$$B = (1/2P)[X\partial V - V\partial X - R(2\mathcal{F}\partial V - V\partial\mathcal{F} - \tilde{\partial}V)]. \quad (\text{B6})$$

The substitution $r = VR$ puts the metric into the NU form:

$$\begin{aligned} ds^2 &= (A - 2r \frac{V_{,u}}{V}) d\tau^2 + 2d\tau dr + 2(B - \frac{r\tilde{\partial}V}{2P}) d\tau d\tilde{\zeta} \\ &\quad - \frac{r^2}{2P^2 V^2} d\zeta d\tilde{\zeta} - \frac{r^2}{2P^2 V^2} \left(\frac{VX}{r} - \mathcal{F} \right) d\tilde{\zeta}^2. \end{aligned} \quad (\text{B7})$$

¹R. Geroch, in *Asymptotic Structure of Space-Time*, edited by P. Esposito and L. Witten (Plenum, New York, 1976).

²E. T. Newman and K. P. Tod in *General Relativity and Gravitation*, Vol. 2, Edited by A. Held (Plenum, New York, 1980).

³M. Ko, M. Ludvigsen, E. T. Newman, and K. P. Tod, *The Theory of \mathcal{H} -Space*, to be published in Phys. Rep.

⁴R. Hansen, E. T. Newman, R. Penrose, and K. P. Tod, Proc. Soc. London A 363 445 (1978).

⁵M. Ko, E. T. Newman, and K. P. Tod in *Asymptotic Structure of Space-Time*, edited by P. Esposito and L. Witten (Plenum, New York, 1976).

⁶K. P. Tod, Gen. Relativ. Gravit. (to be published).

⁷K. P. Tod in *Complex Manifold Techniques in Theoretical Physics*, D. E. Lerner and P. D. Sommers (Pitman, San Francisco, 1979).

⁸Sparling and K. P. Tod, J. Math. Phys. 2 331 (1981).

⁹Complex spaces have two types of shear, left shear σ and right shear $\tilde{\sigma}$, defined by

$$\sigma = m^a m^b \nabla_a l_b,$$

$$\tilde{\sigma} = \tilde{m}^a \tilde{m}^b \nabla_a l_b,$$

which are in general independent of each other though in real spaces they are complex conjugates of each other.

¹⁰E. T. Newman and R. Penrose, J. Math. Phys. 3, 566 (1962).

¹¹Some care must be exercised in this transformation; strictly speaking one must transform (off \mathcal{F}) ζ and $\tilde{\zeta}$ as well as τ and r . Their transformations must have the form

$$\zeta \rightarrow \zeta + K\tilde{\tau}, \quad \tilde{\zeta} \rightarrow \tilde{\zeta} + \tilde{K}\tilde{\tau},$$

where K and \tilde{K} are chosen to cancel terms in $d\tilde{\tau} d\zeta$ and $d\tilde{\tau} d\tilde{\zeta}$ arising from the τ and r transformation.

Static charged perfect fluid in a conformally flat spacetime

A. Banerjee^{a)} and N. O. Santos

Instituto de Física, Ilha do Fundão, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

(Received 16 July 1980; accepted for publication 21 October 1980)

For static charged perfect fluid in a spacetime where $g_{\mu\nu} = e^{2\sigma}\eta_{\mu\nu}$ and $\sigma = \sigma(x,y,z)$, assuming σ functionally related to the electrostatic potential, we prove that such solutions must be spherically symmetric. We consider two special cases. The first one is for an equation of state $\rho = \alpha p$, with α being a constant. We arrive at exact solutions which have singularities. These include the case of a charged dust where $\alpha = 0$, but the solution in this case is singular and is not included in those discussed previously by Das [A. Das, Proc. R. Soc. London, Ser. A **267**, 1 (1962)] and also by De and Raychaudhuri [U. K. De and A. K. Raychaudhuri, Proc. R. Soc. London Ser. A **303**, 97 (1968)]. The second case is for a constant mass density, where we prove that although the Schwarzschild interior solution is regular everywhere, the corresponding analog in the charged fluid case is not everywhere free from singularity.

PACS numbers: 04.20.Jb

1. INTRODUCTION

In view of the existence of a conformally flat solution of Einstein's field equations corresponding to a perfect fluid, which is the well-known Schwarzschild interior solution,¹ it may be interesting to extend such solutions in the case of charged perfect fluid. There are some static solutions in the literature,²⁻⁴ which depend on *a priori* restrictions of different characters on the distributions but are not conformally flat. We have started in our paper with a metric depending only on space coordinates and conformal to a Minkowskian metric. In other words, the metric has the form

$$ds^2 = e^{2\sigma}(dt^2 - dx^2 - dy^2 - dz^2), \text{ where } \sigma = \sigma(x,y,z).$$

There is, however, an assumption that the conformal factor is functionally related to the electrostatic potential, but there is no assumption at the beginning about the symmetry of the distribution. Eventually we arrive at the result that such solutions must be spherically symmetric.

In the next section we consider two special cases. In one we assume an equation of state such that the mass density is linearly related to the pressure ($\rho = \alpha p$), which includes the case of charged dust in the limit when the proportionality constant is zero. We find that all such solutions are singular at the origin. In the absence of matter the solution reduces to that of static electrovac obtained previously by Das.⁵ In the second case we attempt to find solutions for constant mass density. Complete solutions for this case are presented and one of the classes includes the special case of Schwarzschild interior solution. It is found that, although the special form of Schwarzschild interior solution is regular everywhere, the corresponding charged fluid solution is not free from singularity.

It may be remarked that the form of the metric we have chosen for simplicity in our discussions is not the most general conformally flat form of the metric. In the most general form the conformal factor should be a function of time co-

ordinate also, which, however, could be reduced to a static form by a suitable coordinate transformation. This is the reason why we get only a special case of Schwarzschild interior solution for constant mass density and vanishing electromagnetic field. Thus one can conclude that with a static metric conformal to Minkowski metric, there are no physically reasonable solutions for charged perfect fluid with either $\rho = \alpha p$ or $\rho = \text{const}$ and with a metric being functionally related to the electrostatic potential.

In the last section we have considered the most general static spherically symmetric metric and have found exact conformally flat solutions for charged perfect fluid with constant mass density. It is observed that the solutions in this case contain singularities within the distribution. So we cannot have an analog of Schwarzschild interior solution with a physically reasonable distribution of charged perfect fluid.

2. THE FIELD EQUATIONS

Einstein's field equations when both matter and electric charge are present assume the form

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -k(T_{\mu\nu} + E_{\mu\nu}). \quad (2.1)$$

$T_{\mu\nu}$ is the energy-momentum tensor for the matter which we consider as a perfect fluid, so that

$$T_{\mu\nu} = (\rho + p)u_\mu u_\nu - pg_{\mu\nu}, \quad (2.2)$$

where ρ is the mass density, p is the pressure, and u_μ is the 4-velocity satisfying

$$u_\mu u^\mu = 1. \quad (2.3)$$

$E_{\mu\nu}$ is the energy-momentum tensor for the electromagnetic field and is given by

$$E_{\mu\nu} = (1/4\pi)(-F_{\mu\alpha}F_\nu^\alpha + \frac{1}{2}g_{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}), \quad (2.4)$$

where $F_{\mu\nu}$ is the electromagnetic field tensor. Maxwell's equations may be written

$$F_{\mu\nu} = A_{\mu,\nu} - A_{\nu,\mu}, \quad (2.5)$$

$$F^{\mu\nu}{}_{;\nu} = 4\pi J^\mu, \quad (2.6)$$

where A_μ is the 4-potential and J_μ the 4-current tensor. In this paper we want to study static conformally flat solutions

^{a)}On leave from the Department of Physics, Jadavpur University, Calcutta-700032, India.

of (2.1), given in the form,

$$g_{\mu\nu} = e^{2\sigma}\eta_{\mu\nu}, \quad \sigma = \sigma(x_1, x_2, x_3), \quad (2.7)$$

where $\eta_{\mu\nu}$ is the flat metric having signature -2 . The Weyl tensor,⁶

$$C^{\mu}_{\nu\alpha\beta} = R^{\mu}_{\nu\alpha\beta} + \frac{1}{2}(\delta^{\mu}_{\alpha}R_{\nu\beta} - \delta^{\mu}_{\beta}R_{\nu\alpha} + g_{\nu\beta}R^{\mu}_{\alpha} - g_{\nu\alpha}R^{\mu}_{\beta}) + \frac{1}{6}R(\delta^{\mu}_{\beta}g_{\nu\alpha} - \delta^{\mu}_{\alpha}g_{\nu\beta}), \quad (2.8)$$

for such a spacetime (2.7) vanishes. The Einstein tensor for (2.7) assumes the form⁶

$$G_{\mu\nu} = 2\sigma_{\mu\nu} - \eta_{\mu\nu}[2\Box\sigma + \Delta^*\sigma], \quad (2.9)$$

where

$$\begin{aligned} \sigma_{\mu\nu} &= \sigma_{,\mu\nu} - \sigma_{,\mu}\sigma_{,\nu}, \\ \Delta^*\sigma &= \eta^{\mu\nu}\sigma_{,\mu}\sigma_{,\nu}, \\ \Box\sigma &= \eta^{\mu\nu}\sigma_{,\mu\nu}. \end{aligned} \quad (2.10)$$

In the rest frame of the charge distribution the 4-potential becomes

$$A_{\mu} = [\phi(x_1, x_2, x_3), 0, 0, 0], \quad (2.11)$$

and the nonvanishing components of $F_{\mu\nu}$ are only

$$F_{0i} = -F_{i0} = \phi_{,i}. \quad (2.12)$$

The Latin indices assume the values 1, 2, and 3.

Assuming that the conformal factor σ is functionally related to the electrostatic potential

$$\sigma = \sigma(\phi), \quad (2.13)$$

the field Equations (2.1) with (2.2), (2.4), and (2.9) become: for $\mu = \nu = 0$,

$$(2\sigma'' + \sigma'^2 + e^{-2\sigma})\Delta^*\phi + 2\sigma'\Box\phi = k\rho e^{2\sigma}; \quad (2.14)$$

for $\mu = i$ and $\nu = j$, where $i \neq j$,

$$(\sigma'' - \sigma'^2 - e^{-2\sigma})\phi_{,i}\phi_{,j} + \sigma'\phi_{,ij} = 0; \quad (2.15)$$

for $i = j$,

$$2(\sigma'' - \sigma'^2 - e^{-2\sigma})\phi_{,i}^2 + 2\sigma'\phi_{,ii} + 2(\sigma'' + \sigma'^2 - e^{-2\sigma})\Delta^*\phi + 2\sigma'\Box\phi = -k\rho e^{2\sigma}; \quad (2.16)$$

by contraction of (2.1) we obtain $R = kT$, giving

$$6[(\sigma'' + \sigma'^2)\Delta^*\phi + \sigma'\Box\phi] = k(\rho - 3p)e^{2\sigma}. \quad (2.17)$$

The prime indicates differentiation with respect to ϕ .

Defining

$$A(\phi) = \sigma'' - \sigma'^2 - e^{-2\sigma}, \quad (2.18)$$

$$B(\phi) = \sigma',$$

we obtain from Eqs. (2.14), (2.16), and (2.17)

$$A(\phi)(\phi_{,i})^2 + B(\phi)\phi_{,ii} = -\frac{1}{3}A(\phi)\Delta^*\phi + \frac{1}{3}B(\phi)\Box\phi. \quad (2.19a)$$

We observe that the right-hand side of (2.19a) is always negative and nonnull because with Eq. (2.14) we can write

$$\frac{1}{3}A(\phi)\Delta^*\phi + \frac{1}{3}B(\phi)\Box\phi = \frac{1}{2}k\rho e^{2\sigma} - \frac{3}{2}(\sigma'^2 + e^{-2\sigma})\Delta^*\phi = V^2 > 0. \quad (2.19b)$$

Now the system of equations (2.14)–(2.17) in view of (2.19b)

reduces to

$$A(\phi)\phi_{,i}\phi_{,j} + B(\phi)\phi_{,ij} = 0, \quad (2.20)$$

where $i \neq j$, and

$$A(\phi)\phi_{,i}^2 + B(\phi)\phi_{,ii} = -V^2. \quad (2.21)$$

Integrating Eq. (2.20) we obtain

$$\ln\phi_{,i} = \xi(\phi) + A_i(x^i), \quad (2.22)$$

where ξ is a function of ϕ and A_i is a function of x^i alone.

Again (2.22) on further integration yields

$$\int \exp[-\xi(\phi)] d\phi = X + Y + Z, \quad (2.23)$$

where $X = X(x^1)$, $Y = Y(x^2)$, and $Z = Z(x^3)$. In other words, the electrostatic potential ϕ must be a function of u , where $u = X + Y + Z$. In view of the above discussions we can write explicitly Eq. (2.21) as

$$\begin{aligned} (A\phi_{,u}^2 + B\phi_{,uu})X_{,1}^2 + B\phi_{,u}X_{,11} &= -V^2, \\ (A\phi_{,u}^2 + B\phi_{,uu})Y_{,2}^2 + B\phi_{,u}Y_{,22} &= -V^2, \\ (A\phi_{,u}^2 + B\phi_{,uu})Z_{,3}^2 + B\phi_{,u}Z_{,33} &= -V^2. \end{aligned} \quad (2.24)$$

Equations (2.24) lead us to conclude that none of $X_{,1}$, $Y_{,2}$, and $Z_{,3}$ can be zero, because if any one of them is zero $V^2 = 0$, which is contrary to our case unless we have a trivial solution for flat space. Again from (2.20), writing explicitly, we have

$$\begin{aligned} (A\phi_{,u}^2 + B\phi_{,uu})X_{,1}Y_{,2} &= 0, \\ (A\phi_{,u}^2 + B\phi_{,uu})Y_{,2}Z_{,3} &= 0, \\ (A\phi_{,u}^2 + B\phi_{,uu})Z_{,3}X_{,1} &= 0, \end{aligned} \quad (2.25)$$

and since $X_{,1} \neq 0$, $Y_{,2} \neq 0$, and $Z_{,3} \neq 0$ the only conclusion is that

$$A\phi_{,u}^2 + B\phi_{,uu} = 0. \quad (2.26)$$

Equations (2.24) and (2.26) lead us to

$$X_{,11} = Y_{,22} = Z_{,33} = a. \quad (2.27)$$

It can be shown by elementary arguments that a must be a constant and u is therefore given by

$$u = X + Y + Z = \frac{1}{2}a[(x^1)^2 + (x^2)^2 + (x^3)^2] + b_1x^1 + b_2x^2 + b_3x^3 + c, \quad (2.28)$$

where b_1 , b_2 , b_3 , and c are other constants of integration. It is evident that one can write u as

$$u = \frac{1}{2}a[(x^1 + \xi^1)^2 + (x^2 + \xi^2)^2 + (x^3 + \xi^3)^2], \quad (2.29)$$

ξ^1 , ξ^2 , and ξ^3 being constants, and by a suitable coordinate transformation u can be interpreted as a radial coordinate; thus ϕ is dependent only on a radial coordinate. One can therefore conclude that a static charged perfect fluid with metric given in the form $g_{\mu\nu} = e^{2\sigma}\eta_{\mu\nu}$, which is functionally related to the electrostatic potential, must be spherically symmetric.

3. SOLUTIONS OF THE FIELD EQUATIONS

The most general static spherically symmetric line element can be written in the isotropic form

$$ds^2 = e^{\nu}dt^2 - e^{\omega}(dr^2 + r^2d\Omega^2). \quad (3.1)$$

The field equations for such a metric are

$$kp - \phi'^2 e^{-(\nu+\omega)} = e^{-\omega}[(\nu' + \omega')/r + \frac{1}{2}\omega'(\nu' + \frac{1}{2}\omega')],$$

$$kp + \phi'^2 e^{-(\nu+\omega)} = \frac{1}{2}e^{-\omega}[\omega'' + \nu'' + \frac{1}{2}\nu'^2 + (\omega' + \nu')/r], \quad (3.2)$$

$$kp + \phi'^2 e^{-(\nu+\omega)} = -e^{-\omega}[\omega'' + \frac{1}{4}\omega'^2 + 2\omega'/r].$$

The conformally flat metric we have used in Sec. 2 can be written without the loss of generality in the form (3.1) when $\nu = \omega = 2\sigma$. The functional relationship between σ and the electrostatic potential ϕ is trivially satisfied in the spherically symmetric case. We have two special cases of the charged fluid.

Case I

Assuming the equation of state $p = \alpha\omega$, where $\alpha = \text{const}$, the general solution of the field equations for the metric $g_{\mu\nu} = e^{2\sigma}\eta_{\mu\nu}$ is

$$e^{2\sigma} = [h/r^{2\beta} - l]^{1/\beta}, \quad (3.3)$$

where h and l are constants of integration and $\beta = 1/(1 + 3\alpha)$.

The matter density is

$$k\rho = \frac{6hl\beta r^{-2(\beta+1)}}{(h/r^{2\beta} - l)(1/\beta + 2)}. \quad (3.4)$$

It is clear that in order to have $\rho > 0$ both h and l must be greater than zero. But we see that when $r^{2\beta} = h/l$, $\rho \rightarrow \infty$ and so the density increases indefinitely at a point where the area of the spherical surface given by $4\pi r^2 = 4\pi r^2 e^{2\sigma}$ vanishes, or in other words, we get a singularity within the distribution. It can be shown that the charge density is also singular within the distribution.

The case of charged dust is obtained by putting $\alpha = 0$. The solution is singular in this case also and the results previously obtained for static charged dust in^{7,8} are not applicable in this case, because in obtaining regularity everywhere they excluded the solution obtained here. It can be further seen that in our case $(j^4/u^4\rho)^2 \neq 1$ everywhere, unlike the case discussed by Das, De, and Raychaudhuri.

One can conclude, therefore, that it is not possible to obtain nonsingular solutions for either charged perfect fluid with $p = \alpha\rho$ or for charged dust having the metric

$$g_{\mu\nu} = e^{2\sigma}\eta_{\mu\nu},$$

where σ is independent of t and functionally related to the electrostatic potential. When we put $l = 0$, it implies $\rho = p = 0$, which is the electrovac solution given by Das⁵ in the form

$$e^{2\sigma} = h/[(x^1 + \xi^1)^2 + (x^2 + \xi^2)^2 + (x^3 + \xi^3)^2], \quad (3.5)$$

where h , ξ^1 , ξ^2 , and ξ^3 are constants.

Case II

In this case we assume $\rho = \text{const}$. The field equations (3.2) for a metric $g_{\mu\nu} = e^{2\sigma}\eta_{\mu\nu}$ give

$$\frac{2}{3}k\rho e^{2\sigma} + 2\sigma'' + 2\sigma'/r = 0. \quad (3.6)$$

The alternative solutions of (3.6) are given by

$$e^{2\sigma} = [\gamma^1 + D_1 + C_1 r^{1-D_1}]^{-2}, \quad (3.7)$$

$$e^{2\sigma} = r^{-2}[\cos(D_2 \ln r) + C_1 \sin(D_2 \ln r)]^{-2}, \quad (3.8)$$

and

$$e^{2\sigma} = r^{-2}[1 + C_1 \ln r]^{-2}, \quad (3.9)$$

where D_1 , D_2 , and C_1 are constants and the density ρ is related with them in three different solutions. All these solutions are singular within the distribution except for $D_1 = 1$ in the solution (3.7). This gives us a special case of the Schwarzschild interior solution for perfect fluid with constant density in isotropic coordinates (see Ref. 8).

We can thus make one more conclusion that one can not get a physically reasonable solution for a charged perfect fluid with constant mass density and having a metric conformal to Minkowski metric, the metric being functionally related with the electric potential.

4. GENERAL SPHERICALLY SYMMETRIC AND CONFORMALLY FLAT STATIC CHARGED PERFECT FLUID

In order to get a general static spherically symmetric solution we can use the line element (3.1). Applying the condition of conformal flatness (Ref. 5) $C^\mu{}_{\nu\alpha\beta} = 0$ one gets a relation like

$$e^{(\nu-\omega)}(1 + Ar^2)^2. \quad (4.1)$$

In view of the field equations (3.2) one can obtain in a straight forward manner

$$\frac{3}{2}(\omega'' + \omega'/r) = -k\rho e^\omega, \quad (4.2)$$

which is the same differential equation as (3.6). For $\rho = \text{const}$ one obtains the same solutions for e^ω as those given in (3.7)–(3.9) for $e^{2\sigma}$. Once e^ω is known e^ν is obtained from (4.1). Here again the first solution for e^ω with $D_1 = 1$ gives exactly the general Schwarzschild interior solution in isotropic coordinates. All other solutions with $D_1 \neq 1$ have singularities at $r = 0$. We can therefore conclude that there is no physically reasonable solution, which may be said to be an analog of the Schwarzschild interior solution, for a conformally flat, static, and charged perfect fluid sphere having constant mass density.

ACKNOWLEDGMENTS

The authors thank Professor A. Papapetrou and Professor M. M. Som for useful discussions. They acknowledge the financial aid from F.I.N.E.P. and CNPq of Brazil.

¹M. Gurses and Y. Gurse, Nuovo Cimento B 25, 786 (1975).

²W. B. Bonnor, Z. Phys. 160, 59 (1960).

³K. D. Krori and J. Barua, J. Phys. A: Gen. Phys. 8, 508 (1975).

⁴S. J. Wilson, Can. J. Phys. 47, 2401 (1969).

⁵A. Das, J. Math. Phys. 12, 232 (1971).

⁶L. P. Eisenhart, *Riemannian Geometry* (Princeton U.P., Princeton, N. J., 1949).

⁷A. Das, Proc. R. Soc. London, Ser. A 267, 1 (1962).

⁸U. K. De and A. K. Raychaudhuri, Proc. R. Soc. London, Ser. A 303, 97 (1968).

Self-gravitating anisotropic fluids with plane symmetry

P. S. Letelier and R. Machado^{a)}

Departamento de Física, Universidade de Brasília, 70 910 Brasília, D.F., Brazil

(Received 23 September 1980; accepted for publication 20 November 1980)

The general solution to Einstein's equations coupled to an anisotropic fluid described by two perfect-fluid components is obtained in the case that (1) the space-time is plane-symmetric, (2) each fluid component is irrotational, and (3) each one obeys the equation of state pressure = energy density. The method used consists of solving the equivalent problem of the Einstein's equations coupled to a complex massless-scalar field. The space-time singularities are studied using the concept of velocity-dominated singularity. Comoving systems of coordinates are also studied.

PACS numbers: 04.20.Jb

1. INTRODUCTION

In a recent paper¹ it was found that Einstein's field equations for a self-gravitating anisotropic fluid described by two irrotational perfect-fluid components with a stiff equation of state (pressure = energy density) are equivalent to Einstein's field equations coupled to a complex massless-scalar field, A , i.e.,

$$R_{\mu\nu} = -\frac{1}{2}(A_{,\mu}\bar{A}_{,\nu} + \bar{A}_{,\mu}A_{,\nu}), \quad (1.1)$$

$$\partial_\mu(\sqrt{-g}g^{\mu\nu}A_{,\nu}) = 0, \quad (1.2)$$

$$A \equiv \phi + i\psi, \quad \bar{A} \equiv \phi - i\psi. \quad (1.3)$$

The units are so chosen that we have for the velocity of light $c = 1$ and Newton's constant of gravitation $G = 1/8\pi$.

The 4-velocity of the perfect-fluid components are related to the real and imaginary part of A by

$$U_\mu = \phi_{,\mu}/(\phi_{,\alpha}\phi^{,\alpha})^{1/2}, \quad (1.4a)$$

$$v_\mu = \psi_{,\mu}/(\phi_{,\alpha}\psi^{,\alpha})^{1/2}. \quad (1.4b)$$

The anisotropic fluid variables are connected to A by

$$T_{\mu\nu} = \rho U_\mu U_\nu + (\sigma - \pi)\chi_\mu\chi_\nu - \pi(g_{\mu\nu} - U_\mu U_\nu), \quad (1.5)$$

$$\rho = \sigma = \frac{1}{2}|A^{,\mu}A_{,\mu}|, \quad (1.6)$$

$$\pi = \frac{1}{2}\bar{A}_{,\mu}A^{,\mu}, \quad (1.7)$$

$$U_\mu = \text{Re}(e^{i\alpha}\bar{A}_{,\mu})/[\text{Re}(e^{i\alpha}\bar{A}_{,\nu})\text{Re}(e^{i\alpha}A^{,\nu})]^{1/2}, \quad (1.8)$$

$$\chi_\mu = -\text{Im}(e^{i\alpha}\bar{A}_{,\mu})/[-\text{Im}(e^{i\alpha}\bar{A}_{,\nu})\text{Im}(e^{i\alpha}A^{,\nu})]^{1/2}, \quad (1.9)$$

$$(\sqrt{2})e^{i\alpha} = \left[1 + \frac{\text{Re}(A_{,\mu}A^{,\mu})}{|A_{,\beta}A^{,\beta}|}\right]^{1/2} + i\left[1 - \frac{\text{Re}(A_{,\mu}A^{,\mu})}{|A_{,\beta}A^{,\beta}|}\right]^{1/2}, \quad (1.10)$$

where U^μ is the anisotropic fluid flux velocity, χ^μ is a space-like unit 4-vector that points in the direction of anisotropy, ρ is the usual rest energy density of the fluid, π is the pressure on a plane perpendicular to the anisotropy direction, and σ is the pressure along the anisotropy direction.

The purpose of this paper is to study Einstein's equations (1.1) or better its anisotropic fluid interpretation, for the space-time that admits the three-parameter groups of

motions that characterize plane symmetry.

In Sec. 2 we present the general solution to Eq. (1.1) for a plane symmetric space-time, and we compute the anisotropic fluid variables for this case. In Sec. 3 different comoving systems of coordinates are studied. In Sec. 4 the concept of velocity-dominated singularity^{2,3} is used to study the fluid and space-time singular behavior. The rather surprising result is found that near the singularity the anisotropic fluid becomes isotropic.

2. THE SOLUTION

The most general plane-symmetric metric can be written as⁴

$$ds^2 = e^\omega du dv - e^\mu(dx^2 + dy^2), \quad (2.1)$$

where ω and μ are functions of u and v

The field equations (1.1) and (1.2) for the metrics (2.1) reduce to

$$\mu_{++} + \frac{1}{2}\mu_+^2 - \omega_+\mu_+ = -\phi_+^2 - \psi_+^2, \quad (2.2a)$$

$$\mu_{--} + \frac{1}{2}\mu_-^2 - \omega_-\mu_- = -\phi_-^2 - \psi_-^2, \quad (2.2b)$$

$$\omega_{+-} + \mu_{+-} + \frac{1}{2}\mu_+\mu_- = -\phi_+\phi_- - \psi_+\psi_-, \quad (2.2c)$$

$$(e^\mu)_{+-} = 0, \quad (2.2d)$$

$$-2\phi_{+-} = \mu_-\phi_+ + \mu_+\phi_-, \quad (2.3a)$$

$$-2\psi_{+-} = \mu_-\psi_+ + \mu_+\psi_-, \quad (2.3b)$$

where we have introduced the notation $\mu_+ \equiv (\partial\mu/\partial v)$, $\mu_- \equiv (\partial\mu/\partial u)$, etc.

The general solution (2.2d) is

$$e^\mu = t \equiv f(u) + h(v), \quad (2.4)$$

where f and h are functions of their arguments. From (2.4), (2.2a), and (2.2b) we find

$$\omega_+ = f_{++}/f_+ - f_+/2t + t(\phi_+^2 + \psi_+^2)/f_+, \quad (2.5a)$$

$$\omega_- = h_{--}/h_- - h_-/2t + t(\phi_-^2 + \psi_-^2)/h_-. \quad (2.5b)$$

Note that from (2.3) and either (2.5a) or (2.5b) we recover (2.2c) and that (2.3) and (2.5) imply $\omega_{+-} = \omega_{-+}$. So the integral

$$\omega = \ln(4t^{1/2}f_+h_-) + \Omega[\phi, \psi], \quad (2.6a)$$

$$\Omega[\phi, \psi] \equiv M[\phi] + M[\psi] + \omega_0, \quad (2.6b)$$

$$M[\phi] \equiv \int t[(\phi_+^2/f_+)du + (\phi_-^2/h_-)dv], \quad (2.6c)$$

^{a)}Partially supported by CNPq, Brazil.

is exact and ω_0 is an integration constant. Now defining $Z \equiv h(v) - f(u)$ we can cast the metric (2.1) as

$$ds^2 = (e^{\Omega/\sqrt{t}})(dt^2 - dz^2) - t(dx^2 + dy^2). \quad (2.7)$$

In the system of coordinates (t, z, x, y) the relations (2.6c) and (2.3) can be written as

$$M[\phi] = \int t [(\phi_t^2 + \phi_z^2)dt + 2\phi_t\phi_z dz], \quad (2.8)$$

$$\phi_{tt} + \phi_t/t - \psi_{zz} = 0, \quad (2.9a)$$

$$\psi_{tt} + \psi_t/t - \psi_{zz} = 0, \quad (2.9b)$$

where we have introduced the notation

$\psi_t \equiv (\partial\psi/\partial t)$, $\psi_z \equiv (\partial\psi/\partial z)$, etc. Thus, the general solution to the system of equations (2.2) and (2.3) can be found computing $M[\phi]$ and $M[\psi]$, where ϕ and ψ are general solutions of the linear equation (2.9). The general solution to (2.9) is well known.

From (2.7) and (1.6)–(1.9) we find that the anisotropic fluid variables can be expressed as

$$\rho = \sigma = \frac{1}{2}(\sqrt{t})e^{-\Omega}r_1, \quad (2.10)$$

$$\pi = \frac{1}{2}(\sqrt{t})e^{-\Omega}s_2, \quad (2.11)$$

$$U_\mu = \frac{e^{\Omega/2}}{t^{1/4}} \frac{(r_1 + s_1)^{1/2}\phi_{,\mu} + (r_1 - s_2)^{1/2}\psi_{,\mu}}{(s_1^2 + r_2^2 + r_1s_2)^{1/2}}, \quad (2.12)$$

$$\chi_\mu = -\frac{e^{\Omega/2}}{t^{1/4}} \frac{(r_1 - s_1)^{1/2}\phi_{,\mu} - (r_1 + s_1)^{1/2}\psi_{,\mu}}{(s_1^2 + r_2^2 - r_1s_2)^{1/2}}, \quad (2.13)$$

where

$$s_1 \equiv \phi_t^2 - \phi_z^2 - \psi_t^2 + \psi_z^2, \quad (2.14)$$

$$s_2 \equiv \phi_t^2 - \phi_z^2 + \psi_t^2 - \psi_z^2, \quad (2.15)$$

$$r_1 \equiv (s_1^2 + r_2^2)^{1/2}, \quad (2.16)$$

$$r_2 \equiv 2(\phi_t\psi_t - \phi_z\psi_z). \quad (2.17)$$

The fluid anisotropy for the present model can be described by the quantity

$$\delta \equiv (\sigma - \pi)/\pi > 0, \quad (2.18)$$

$$= (r_1 - r_2)/s_2. \quad (2.19)$$

Note that, letting $\psi \rightarrow 0$ in (2.6b), (2.7), (2.10), (2.11), and (2.12), we recover the corresponding expression for the Taubensky and Taub solutions.⁴

3. COMOVING COORDINATES

For the fluid under consideration we have two types of comoving coordinates, first the comoving coordinates with respect to one of the fluid components and second the comoving coordinates with respect to the fluid flux velocity. Comoving coordinates of the first type can be easily found. The coordinates

$$T = \phi(t, z), \quad (3.1a)$$

$$dZ = t(\phi_z dt + \phi_t dz), \quad (3.1b)$$

$$X = x, \quad Y = y, \quad (3.1c)$$

are comoving to the fluid component described by $u^\mu = \phi^{,\mu}/(\phi_{,\alpha}\phi^{,\alpha})^{1/2}$, since

$$u^\mu \rightarrow u^{\mu'}(T, Z) = (\phi_{,\alpha}\phi^{,\alpha})^{1/2}(1, 0, 0, 0), \quad (3.2)$$

The integrability of (3.1b) is guaranteed by (2.9a).

The metric in this new system of coordinates can be cast

$$ds^2 = (\phi_{,\alpha}\phi^{,\alpha})^{-1}(dT^2 - t^{-2}dZ^2) - t(dX^2 + dY^2). \quad (3.3)$$

Note that this metric, as well as the Jacobian of the transformation (3.1), is singular at $t = 0$.

The velocity $v^\mu = \psi^{,\mu}/(\psi_{,\mu}\psi^{,\mu})^{1/2}$ of the other fluid component, as well as the flux velocity U^μ and the anisotropy direction χ^μ , transform under (3.1) as

$$v^\sigma = \phi_{,\alpha}\psi^{,\alpha}/(\psi_{,\beta}\psi^{,\beta})^{1/2}, \quad (3.4a)$$

$$v^1 = t^{3/2}e^{-\Omega}(\phi_z\psi_t - \phi_t\psi_z)/(\psi_{,\alpha}\psi^{,\alpha})^{1/2}, \quad (3.4b)$$

$$U^\sigma = \frac{e^{\Omega/2}}{t^{1/2}} \frac{(r_1 + s_1)^{1/2}\phi_{,\alpha}\phi^{,\alpha} + (r_1 - s_1)^{1/2}\phi_{,\alpha}\psi^{,\alpha}}{(s_1^2 + r_2^2 + r_1s_2)^{1/2}}, \quad (3.5a)$$

$$U^1 = t^{5/4}e^{-\Omega/2}(r_1 - s_1)^{1/2} \frac{(\phi_z\psi_t - \phi_t\psi_z)}{(s_1^2 + r_2^2 + r_1s_2)^{1/2}}, \quad (3.5b)$$

$$\chi^\sigma = \frac{-e^{\Omega/2}}{t^{1/2}} \frac{(r_1 - s_1)^{1/2}\phi_{,\alpha}\phi^{,\alpha} - (r_1 + s_1)^{1/2}\phi_{,\alpha}\psi^{,\alpha}}{(s_1^2 + r_2^2 - r_1s_2)^{1/2}}, \quad (3.6a)$$

$$\chi^1 = t^{5/4}e^{-\Omega/2}(r_1 + s_1)^{1/2} \frac{(\phi_z\psi_t - \phi_t\psi_z)}{(s_1^2 + r_2^2 - r_1s_2)^{1/2}}. \quad (3.6b)$$

In a similar way one can define comoving coordinates to the fluid component $v^\mu = \psi^{,\mu}/(\psi^{,\alpha}\psi_{,\alpha})$. Formally, we can also define a comoving system of coordinates with respect to the fluid flux velocity

$$T = T(t, z), \quad (3.7a)$$

$$dZ = H[(\phi_z + \tan\beta\psi_z)dt + (\phi_t + \tan\beta\psi_t)dz], \quad (3.7b)$$

$$X = x, \quad Y = y, \quad (3.7c)$$

$$\tan\beta \equiv (r_1 - s_1)^{1/2}/(r_1 + s_2)^{1/2}, \quad (3.7d)$$

where H is an integrating factor. It happens that the field equations (2.9) are not sufficient to guarantee the integrability of (3.7b) for any ϕ and ψ solutions of (2.9).

4. SINGULARITIES

The singular behavior of the field equations (2.9) near $t = 0$ is described³ by

$$\phi \simeq E(z)\ln t, \quad (4.1a)$$

$$\psi \simeq F(z)\ln t, \quad (4.1b)$$

where E and F are arbitrary functions of their arguments. Now we shall compute all the relevant quantities keeping only the term of dominant singular behavior.³

From (4.1) and (2.7) we get

$$ds^2 \simeq t^{4(E^2 + F^2) - 1/2}(dt^2 - dz^2) - t(dx^2 + dy^2), \quad (4.2)$$

and, performing a simple change of variables, we find

$$ds \simeq d\tau^2 - \tau^{2P_1}d\xi^2 - \tau^{2P_2}dx^2 - \tau^{2P_3}dy^2, \quad (4.3)$$

where the symbol $P \equiv (P_1, P_2, P_3)$ is given by

$$P = \left(\frac{2(E^2 + F^2) - \frac{1}{4}}{2(E^2 + F^2) + \frac{3}{4}}, \frac{\frac{1}{2}}{2(E^2 + F^2) + \frac{3}{4}}, \frac{\frac{1}{2}}{2(E^2 + F^2) + \frac{3}{4}} \right), \quad (4.4)$$

Note that $\Sigma P_i = 1$ and $\Sigma P_i^2 \neq 1$. Thus the metric (2.7) has a velocity-dominated semi-Kasner-like singularity.^{2,3}

From (2.10), (2.11), (2.19), and (4.1) we find

$$\pi = \sigma \simeq \rho \simeq \frac{1}{2}(E^2 + F^2)t^{-4(E^2 + F^2) - 3/2}, \quad (4.5)$$

$$\delta \simeq 0; \quad (4.6)$$

thus, near the singularity the fluid became isotropic, a result

that was not expected.

¹P. S. Letelier, "Anisotropic fluids with two perfect-fluid components," *Phys. Rev. D*, **27**, 807 (1980).

²E. P. T. Liang, *J. Math. Phys.* **13**, 386 (1972), and references therein.

³P. S. Letelier and R. Tabensky, *J. Math. Phys.* **16**, 8 (1975).

⁴R. Tabensky and A. H. Taub, *Commun. Math. Phys.* **29**, 61 (1973).

Homothetic solutions of Einstein's equations and shock waves

Gaetano Moschetti

Seminario Matematico, Universita' di Catania, Viale A. Doria, 6, I 95125 Catania, Italy

(Received 16 July 1980; accepted for publication 23 October 1980)

The problem of joining two homothetic solutions of Einstein's field equations with a perfect fluid is considered in general, without any symmetry requirements.

PACS numbers: 04.20.Jb, 95.30.Lz

The study of self-similar (or homothetic) solutions of the Einstein equations for a perfect fluid is of fundamental importance in many areas of relativistic astrophysics and cosmology. In particular some models try to explain the variability of quasars and radiosources on the basis of relativistic blast waves.¹ These waves correspond to self-similar solutions of the special relativistic fluid-dynamical equations generalizing the well-known classical Taylor-Sedov solutions.¹⁻³ However, in astrophysical applications, it is doubtful whether the self-gravity of the waves can be neglected⁴ and therefore a fully general-relativistic treatment is warranted.

In this area a fundamental problem is that of matching two self-similar solutions of Einstein's equations with a perfect fluid across a shock front. In the case of spherical symmetry, Cahill and Taub⁵ have shown that self-similarity is preserved by the Einstein equations, provided the initial data are self-similar. They have also shown that, in the case of spherical symmetry, if the perfect fluid is thermodynamic and the space-time is self-similar behind a shock front, the matching conditions imply that it is self-similar ahead of the shock front.

In this paper, Cahill and Taub's results are proved without imposing the restriction of spherical symmetry. The proof hinges upon some results of Coll⁶ on the Cauchy problem of Killing vectors suitably generalized to homothetic Killing vectors.

In Sec. 1 the formalism which will be used is expounded. In Sec. 2 one treats the problem of the preservation of self-similarity by the Einstein equations in the case of perfect fluids.

Finally in Sec. 3 one considers the problem of the continuation of a homothetic vector across a shock front.

1. GENERALITY

Let M be a differentiable manifold with a metric g with signature $(1 - 1 - 1 - 1)$. Consider on M the hypersurfaces $\phi(p) = \text{const}$, such that $g(d\phi, d\phi) \neq 0$. If one writes $\epsilon = \pm 1$, according to whether the hypersurfaces are space-like or timelike, the unity normal is

$$n = \lambda d\phi, \quad (1)$$

where $\lambda = \pm 1/(\epsilon g(d\phi, d\phi))^{1/2}$.

Let h_{ab} and K_{ab} be the first and the second fundamental forms of $\phi(p) = \text{const}$, namely⁷

$$h_{ab} = g_{ab} - \epsilon n_a n_b, \quad (2)$$

$$K_{ab} = h_a^c h_b^d \nabla_c n_d, \quad (3)$$

where ∇_c is the covariant derivative with respect to the metric g_{ab} .

Let \tilde{R}_{ab} be the Ricci tensor of the metric h_{ab} ,⁷ and T_{ab} the energy-momentum tensor, which is characterized by⁶

$$\tau = T_{ab} n^a n^b, \quad (4)$$

$$t_c = T_{ab} n^a h^b_c, \quad (5)$$

$$H_{ab} = h_a^c h_b^d T_{cd}. \quad (6)$$

If one writes

$$K = h^{ab} K_{ab}, \quad (7)$$

$$\tilde{R} = h^{ab} \tilde{R}_{ab}, \quad (8)$$

$$H = h^{ab} H_{ab}, \quad (9)$$

$$S_{ab} = K_{ac} K_b^c - \frac{1}{2} K K_{ab}, \quad (10)$$

$$P_{ab} = H_{ab} - \frac{1}{2} (H + \epsilon \tau) h_{ab}, \quad (11)$$

and takes a coordinate congruence adapted to a vector field $V = \epsilon \lambda n + \mu$,⁸ where μ is a vector field tangent to the hypersurfaces $\phi(p) = \text{const}$, then the Einstein equations read⁶

$$D^a K_{ab} - D_b K = t_b, \quad (12)$$

$$K^{ab} K_{ab} - K^2 + \epsilon \tilde{R} = -2\tau, \quad (13)$$

$$\partial h_{ab} = 2\epsilon \lambda K_{ab}, \quad (14)$$

$$\partial K_{ab} = \lambda (\tilde{R}_{ab} + 2\epsilon S_{ab} - P_{ab}) - D_a D_b \lambda, \quad (15)$$

where $\partial = L_{V-\mu}$ denotes the Lie derivative along $V - \mu$ and D_a denotes the covariant derivative with respect to the metric h_{ab} defined⁷ by $D_a A^b_c = h_a^i h_j^b h_c^k \nabla_i A^j_k$. Equations (12) and (13) are the constraint equations and Eqs. (14) and (15) are the evolution equations.

Let

$$B_{ab} = L_{\xi} g_{ab} - c g_{ab}, \quad (16)$$

where $\xi = \epsilon \alpha n + \beta$, α is a function, β a vector field tangent to the hypersurfaces $\phi(p) = \text{const}$, and c a constant. If one writes $\sigma = B_{ab} n^a n^b$, $1_c = B_{ab} n^a h_c^b$, $L_{ab} = h_a^c h_b^d B_{cd}$, then

$$\sigma = (\epsilon/\lambda) \{ 2(\partial\alpha + L_{\beta}\lambda) - c\lambda \}, \quad (17)$$

$$1_a = (\epsilon/\lambda) \{ h_{ab} \partial\beta^b + \epsilon \lambda D_a \alpha - \epsilon \alpha D_a \lambda \}, \quad (18)$$

$$L_{ab} = L_{\beta} h_{ab} + 2\epsilon \alpha K_{ab} - c h_{ab}. \quad (19)$$

By choosing α and β such that $\sigma = 1_a = 0$, namely

$$\partial\alpha = \frac{1}{2} c\lambda - L_{\beta}\lambda, \quad (20)$$

$$\partial\beta^a = \epsilon (\alpha D^a \lambda - \lambda D^a \alpha), \quad (21)$$

the derivative of Eq. (19) along $V - \mu$ is

$$\partial L_{ab} = 2\epsilon \lambda \{ L_{\beta} K_{ab} + \alpha (\tilde{R}_{ab} + 2\epsilon S_{ab} - P_{ab}) - D_a D_b \alpha - \frac{1}{2} c K_{ab} \}, \quad (22)$$

and by Eq. (A1)

$$\begin{aligned} (L_{\xi} T_{cd})h_a^c h_b^d &= (1/\lambda) \{ \lambda L_{\xi} H_{ab} + \epsilon \lambda (t_b D_a \alpha + t_a D_b \alpha) \\ &- \epsilon \alpha (t_b D_a \lambda + t_a D_b \lambda) \}. \end{aligned} \quad (23)$$

ξ is a homothetic Killing vector iff $B_{ab} = 0$.

From the Einstein equations, if ξ is a homothetic Killing vector, it follows that $L_{\xi} T_{ab} = 0$, hence the following.

Proposition 1: If ξ is a homothetic Killing vector, Eqs. (20) and (21) and the following equations

$$L_{\beta} h_{ab} + 2\epsilon \alpha K_{ab} - ch_{ab} = 0, \quad (24)$$

$$L_{\beta} K_{ab} + \alpha (\tilde{R}_{ab} + 2\epsilon S_{ab} - P_{ab}) - D_a D_b \alpha - \frac{1}{2} c K_{ab} = 0, \quad (25)$$

$$\lambda L_{\xi} H_{ab} + \epsilon \lambda (t_b D_a \alpha + t_a D_b \alpha) - \epsilon \alpha (t_b D_a \lambda + t_a D_b \lambda) = 0 \quad (26)$$

hold true.

Remark: If $\alpha = 0$, Eqs. (19)–(23) become

$$L_{ab} = L_{\beta} h_{ab} - ch_{ab}, \quad (19')$$

$$L_{\beta} \lambda = \frac{1}{2} c \lambda, \quad (20')$$

$$\partial \beta^a = 0, \quad (21')$$

$$\partial L_{ab} = 2\epsilon \lambda (L_{\beta} K_{ab} - \frac{1}{2} c K_{ab}), \quad (22')$$

$$(L_{\beta} T_{cd})h_a^c h_b^d = L_{\beta} H_{ab}. \quad (23')$$

If $\alpha \neq 0$, by starting from a non-null hypersurface Σ , a convenient set of hypersurfaces, such that $\Sigma = \{ \phi(p) = 0 \}$ and $V = \xi$, can be found.⁸ Therefore by using Eqs. (14) and (15), Eqs. (19)–(23) become

$$L_{ab} = L_{\nu} h_{ab} - ch_{ab}, \quad (19'')$$

$$L_{\nu} \lambda = \frac{1}{2} c \lambda, \quad (20'')$$

$$\partial \mu^a = 0, \quad (21'')$$

$$\partial L_{ab} = 2\epsilon \lambda (L_{\nu} K_{ab} - \frac{1}{2} c K_{ab}), \quad (22'')$$

$$(L_{\nu} T_{cd})h_a^c h_b^d = L_{\nu} H_{ab}. \quad (23'')$$

Equations (19'')–(23'') can be formally obtained from Eqs. (19')–(23') by replacing β by V . From this it is easily seen that all the calculations given in the appendix, where the vector field β is used, also hold true when β is replaced by V . Hence any theorem involving the vector field $\xi = \epsilon \alpha n + \beta$, will be proved only in the simpler case $\alpha = 0$.

2. THE CAUCHY PROBLEM FOR THE HOMOTHETIC KILLING VECTORS

Theorem 1: Under suitable differentiability conditions,⁶ a vector field ξ is a homothetic Killing vector if and only if Eqs. (24) and (25) are satisfied on a non-null hypersurface Σ and Eqs. (20), (21), and (26) hold true in a neighborhood of Σ .

Proof: The necessity of these conditions is a trivial consequence of Proposition 1. In the $\alpha = 0$ case the conditions of the theorem are the Eqs. (20') and (21') and

$$L_{ab}|_{\Sigma} = 0, \quad \partial L_{ab}|_{\Sigma} = 0, \quad (27)$$

$$G_{ab} = L_{\beta} H_{ab} = 0. \quad (28)$$

By deriving Eq. (22') along $V - \mu$ and by using Eqs. (15), (20'), (21'), and (A2)–(A5) one finds

$$\begin{aligned} \frac{1}{2} \epsilon \partial^2 L_{ab} &= (\epsilon/2\lambda) (\partial \lambda) (\partial L_{ab}) + \lambda \{ \lambda L_{\beta} (\tilde{R}_{ab} + 2\epsilon S_{ab} - P_{ab}) \\ &- L_{\beta} D_a D_b \lambda + \frac{1}{2} c D_a D_b \lambda \} \\ &= D_{ab} \{ L(2), \partial L(0) \} + D_{ab} \{ G(0) \}, \end{aligned} \quad (29)$$

where $D_{ab} \{ L(r), \partial L(m) \}$, and $D_{ab} \{ G(n) \}$ are homogeneous polynomials in L_{ab} , ∂L_{ab} , G_{ab} , and their tangential derivatives of order r , m , and n respectively. The Eqs. (27)–(29) prove the sufficiency.

Henceforth solutions of Einstein's equations with a perfect fluid energy-momentum tensor will be considered, namely

$$T_{ab} = (\rho + p) u_a u_b - p g_{ab}, \quad (30)$$

where p is the pressure, ρ the rest energy density and u_a the fluid unit velocity. Put $v = g(u, n)^9$ and $\chi_a = h_a^b u_b$,

$$u_a = \epsilon v n_a + \chi_a, \quad (31)$$

$$\tau = (\rho + p) v^2 - \epsilon p, \quad (32)$$

$$t_a = (\rho + p) v \chi_a, \quad (33)$$

$$H_{ab} = (\rho + p) \chi_a \chi_b - p h_{ab}. \quad (34)$$

The last equation can be written⁶

$$H_{ab} = (1/\pi) t_a t_b - p h_{ab}, \quad (35)$$

where

$$\pi = \epsilon p + \tau. \quad (36)$$

Theorem 2: The vector field ξ is a homothetic Killing vector if and only if Eqs. (24) and (25) are satisfied on a non-null hypersurface Σ , and Eqs. (20) and (21) together with

$$L_{\xi} p + cp = 0 \quad (37)$$

hold in a neighborhood of Σ .

Proof: From Eqs. (35), (A5), and (A6)

$$\begin{aligned} L_{\beta} H_{ab} &= D_{ab} \{ L(2), \partial L(1) \} - \epsilon ((1/\pi^2) t_a t_b + h_{ab}) \\ &\quad \times (L_{\beta} p + cp). \end{aligned} \quad (38)$$

Equations (29), (27), and (38) prove the sufficiency. The necessity is trivially obtained from proposition 1 and Eq. (38).

Lemma 1: If ξ is a vector field subject to Eqs. (20) and (21), then

$$L_{\xi} p + cp = D \{ L(2), \partial L(1) \} + (1/\bar{V}^2) (L_{\xi} p + cp), \quad (39)$$

$$L_{\xi} (p/\rho) = D \{ L(2), \partial L(1) \} + (1/\rho^2) (\rho - p/\bar{V}^2) (L_{\xi} p + cp), \quad (40)$$

where $\bar{V}^2 = v^2/(v^2 - \epsilon)$.

Proof: From Eqs. (32) and (33)

$$\rho = \epsilon \tau + \eta/\pi, \quad (41)$$

where $\eta = t^a t_a$; from Eqs. (41), (A7), and (A9) it follows that

$$L_{\beta} \rho + cp = D \{ L(2), \partial L(1) \} - \epsilon (\eta/\pi^2) (L_{\beta} p + cp); \quad (42)$$

and from Eqs. (32) and (33) it follows that

$$\epsilon \eta/\pi^2 = -1/\bar{V}^2. \quad (43)$$

Equation (40) is trivially obtained from Eq. (39).

Theorem 3: If the perfect fluid is barotropic, namely, $p = f(\rho)$, space-time is homothetic in a neighborhood of a non-null hypersurface Σ if and only if a scalar α and a tangent vector field β exist on Σ , such that Eqs. (24) and (25)

hold, and $f(\rho) = k\rho$, where $k = \text{const}$, holds in a neighborhood of Σ .

Proof: Construct, by Eqs. (20') and (21'), β and λ in a neighborhood of Σ . If $p = k\rho$, Eq. (40) becomes

$$(1/\rho)(1 - k/\bar{V}^2)(L_\beta p + cp) = D\{L(2), \partial L(1)\}. \quad (44)$$

If $k = \bar{V}^2$, then from Eq. (43)

$$\pi^2 + \epsilon k \eta = 0, \quad (45)$$

which once differentiated with respect to β , gives

$$D\{L(2), \partial L(1)\} + 2\epsilon\rho\pi(L_\beta p + cp) = 0. \quad (46)$$

The necessity follows trivially from Theorem 2 and Lemma 1 by differentiating $p = f(\rho)$ along β .

Henceforth the term "barotropic perfect fluid" will be used to indicate the $p = k\rho$, where $k = \text{const}$.

Now we will consider a thermodynamic perfect fluid, namely a perfect fluid characterized by a function θ , the proper temperature, and a function S , the specific proper entropy, such that⁵

$$\rho + p = (p/\theta)G(\theta), \quad (47)$$

$$\theta dS = dG - (1/r)dp, \quad (48)$$

$$r = (p/\theta), \quad (49)$$

where r is the rest mass density. If a is the velocity of sound then^{5,10}

$$\gamma = 1/a^2 = (G(\theta)/\theta)(1 - 1/\dot{G}(\theta)), \dot{G}(\theta) = dG/d\theta, \quad (50)$$

$$\omega dS = d\rho - \gamma dp, \quad (51)$$

$$u(S) = 0. \quad (52)$$

Henceforth the case $\dot{G}(\theta) = G(\theta)/\theta$, corresponding to a barotropic perfect fluid, will be excluded.

By differentiating Eq. (32) along β and from Eq. (A7) it follows that

$$2v(\rho + p)L_\beta v + (v^2 - \epsilon)(L_\beta p + cp) + v^2(L_\beta \rho + c\rho) = D\{L(2), \partial L(0)\}. \quad (53)$$

From Eq. (47) it follows that:

$$L_\beta \frac{p}{\rho} = \frac{G(\theta) - \theta \dot{G}(\theta)}{(G(\theta) - \theta)^2} L_\beta \theta. \quad (54)$$

Then

Proposition 2: If ξ is a homothetic Killing vector one has

$$L_\xi v = 0 \quad (55)$$

and if the fluid is thermodynamic one has

$$L_\xi \theta = 0. \quad (56)$$

Theorem 4: A space-time with a thermodynamic perfect fluid is homothetic in a neighborhood of a non-null hypersurface Σ if and only if a scalar α and a tangent vector field β exist on Σ such that Eqs. (24) and (25) and

$$\alpha \partial \theta + \lambda L_\beta \theta = 0 \quad (57)$$

hold true.

Proof: Construct, by Eqs. (20') and (21'), β and λ in a neighborhood of Σ . From Eq. (52)

$$\tilde{\partial} S = 0, \quad (58)$$

where $\tilde{\partial} = \pi \partial + \lambda L_\beta$, and from Eq. (51),

$$\tilde{\partial} \rho - \gamma \tilde{\partial} p = 0. \quad (59)$$

Equations (36), (41), and (59) give

$$(1 + \gamma)\epsilon\pi^2 \tilde{\partial} \tau - (\epsilon\gamma\pi^2 + \eta)\tilde{\partial} \pi + \pi \tilde{\partial} \eta = 0, \quad (60)$$

which can be written

$$\partial \pi = (1/[\epsilon\gamma\pi^2 + \eta])\{\epsilon(1 + \gamma)\pi^2 \partial \tau + \epsilon\lambda(1 + \gamma)\pi L_\beta \tau - (\lambda/\pi)(\epsilon\gamma\pi^2 + \eta)L_\beta \pi + \tilde{\partial} \eta\}, \quad (61)$$

from which

$$\partial p = (\epsilon/\epsilon\gamma\pi^2 + \eta)\{(\epsilon\pi^2 - \eta)\partial \tau + \epsilon\lambda(1 + \gamma)\pi L_\beta \tau - (\lambda/\pi)(\epsilon\gamma\pi^2 + \eta)L_\beta \pi + \tilde{\partial} \eta\}, \quad (62)$$

but from Eq. (59)

$$\partial p = \gamma \partial p - (\lambda/\pi)(L_\beta \rho - \gamma L_\beta p), \quad (63)$$

hence

$$\partial \frac{p}{\rho} = \frac{\epsilon(\rho - \gamma p)}{\rho^2(\epsilon\gamma\pi^2 + \eta)}\{(\epsilon\pi^2 - \eta)\partial \tau + \epsilon\lambda(1 + \gamma)\pi L_\beta \tau - \frac{\lambda p}{\pi}(\epsilon\gamma\pi^2 + \eta)L_\beta \pi + \tilde{\partial} \eta\} + \frac{\lambda p}{\pi \rho^2}(L_\beta \rho - \gamma L_\beta p). \quad (64)$$

Let $N = L_\beta \theta$; from Eqs. (A9) and (A14)-(A17)

$$L_\beta \epsilon(\rho - \gamma p)/\rho^2(\epsilon\gamma\pi^2 + \eta) = D\{N(0)\} + D\{L(2), \partial L(1)\} + 3c\epsilon(\rho - \gamma p)/\rho^2(\epsilon\gamma\pi^2 + \eta). \quad (65)$$

From Eqs. (A7), (A9), (A10), (A13), (A17), (A18), and (A23)

$$L_\beta \{(\epsilon\pi^2 - \eta)\partial \tau\} = D\{N(0)\} + D\{L(2), \partial L(2)\} - 3c(\epsilon\pi^2 - \eta)\partial \tau. \quad (66)$$

From Eqs. (A9), (A14), (A17), and (A21)

$$L_\beta \{(\lambda/\pi)(\epsilon\gamma\pi^2 + \eta)L_\beta \pi\} = D\{N(1)\} + D\{L(3), \partial L(2)\} - 3c(\lambda/\pi)(\epsilon\gamma\pi^2 + \eta)L_\beta \pi. \quad (67)$$

From Eqs. (A11), (A14), and (A17)

$$L_\beta \{\epsilon\lambda(1 + \gamma)\pi L_\beta \tau\} = D\{N(0)\} + D\{L(3), \partial L(1)\} - 3c\epsilon\lambda(1 + \gamma)\pi L_\beta \tau. \quad (68)$$

From Eqs. (A7)-(A10), (A12), (A17), (A.22), and (A24)

$$L_\beta \tilde{\partial} \eta = D\{N(1)\} + D\{L(3), \partial L(2)\} - 3c\tilde{\partial} \eta. \quad (69)$$

From Eqs. (A14)-(A20)

$$L_\beta \{(\lambda p/\pi \rho^2)(L_\beta \rho - \gamma L_\beta p)\} = D\{N(1)\} + D\{L(3), \partial L(2)\}. \quad (70)$$

Equations (64)-(70) give

$$L_\beta \partial(p/\rho) = D\{N(1)\} + D\{L(3), \partial L(2)\}. \quad (71)$$

But from Eq. (47)

$$L_\beta \partial \frac{p}{\rho} = \frac{G(\theta) - \theta \dot{G}(\theta)}{(G(\theta) - \theta)^2} \partial N + D\{N(0)\}, \quad (72)$$

therefore

$$\partial N = D\{N(1)\} + D\{L(3), \partial L(2)\}, \quad (73)$$

and from Eqs. (29), (38), and (40)

$$\partial^2 L_{ab} = D_{ab}\{N(0)\} + D_{ab}\{L(2), \partial L(1)\}. \quad (74)$$

Equations (73) and (74) and the hypothesis

$L_{ab}|_\Sigma = 0$, $\partial L_{ab}|_\Sigma = 0$, $N|_\Sigma = 0$ prove the theorem.

3. ON THE CONTINUATION OF A HOMOTHETIC KILLING VECTOR THROUGH A SHOCK

Let (\tilde{M}, g) be a space time with a perfect fluid, divided into two parts M_- and M^+ by a timelike noncharacteristic hypersurface Σ , through which the second derivative of the metric g (therefore the energy-momentum tensor) are discontinuous, but the first and the second fundamental form h_{ab} and K_{ab} are continuous.

These hypotheses and Eqs. (12) and (13) imply the continuity of the vector field

$$m_a = -\tau n_a + t_a \quad (75)$$

through Σ .

The continuity of m_a together with the continuity of ν through Σ are the Rankine-Hugoniot equations.^{5,10}

In this case ν and \bar{V} , defined in the previous section, are respectively the frequency and the velocity of the shock front relative to the fluid.^{5,10}

From the continuity of h_{ab} and K_{ab} through Σ and from Theorem 3 follows

Theorem 5: If a homothetic Killing vector β tangent to Σ exists on M_- and if the fluid is barotropic on M , then M is homothetic in a neighborhood of Σ .

Theorem 6: If the fluid is thermodynamic on $\tilde{M} = M_- \cup M^+$ and a homothetic Killing vector β tangent to Σ exists on M_- , then M is homothetic in a neighborhood of Σ .

Proof: Since β is tangent to Σ and h_{ab} and K_{ab} are continuous through Σ , it follows that

$$L_{ab}|_{\Sigma} = 0, \quad \partial L_{ab}|_{\Sigma} = 0. \quad (76)$$

The theorem follows from Theorem 4 if $L_{\beta}\theta|_{\Sigma} = 0$.

Since M_- is homothetic, from Theorem 2 and Proposition 2

$$L_{\beta}p_- = -cp_-, \quad L_{\beta}\nu_- = 0, \quad L_{\beta}\theta_- = 0. \quad (77)$$

These equations and the continuity of ν through Σ give

$$L_{\beta}p\nu/\theta = -cp\nu/\theta \quad (78)$$

on Σ . But

$$L_{\beta}p\nu/\theta = (\theta(\nu L_{\beta}p + pL_{\beta}\nu) - p\nu L_{\beta}\theta)/\theta^2, \quad (79)$$

hence

$$L_{\beta}\nu/\nu = -(1/p)(L_{\beta}p + cp) + L_{\beta}\theta/\theta \quad (80)$$

on Σ . From Eqs. (39) and (53)

$$(1/\bar{V}^2)(L_{\beta}p + cp) + (\rho + p)(L_{\beta}\nu/\nu) = 0 \quad (81)$$

on Σ . Equations (80) and (81) give

$$\left(\frac{1}{\bar{V}^2} - \frac{\rho + p}{p}\right)(L_{\beta}p + cp) + \frac{\rho + p}{\theta}L_{\beta}\theta = 0, \quad (82)$$

and by Eq. (40)

$$\left(\frac{1}{\bar{V}^2} - 1 - \frac{\rho}{p}\right)\frac{\rho\bar{V}^2}{\bar{V}^2 - \frac{\rho}{p}}L_{\beta}\frac{p}{\rho} + \frac{\rho}{\theta}\left(1 + \frac{p}{\rho}\right)L_{\beta}\theta = 0. \quad (83)$$

The case $\bar{V}^2 = p/\rho$, together with Eq. (40), implies $L_{\beta}p/\rho|_{\Sigma} = 0$, hence $L_{\beta}\theta|_{\Sigma} = 0$. By using Eq. (47), Eq. (83) becomes

$$\begin{aligned} 0 &= \rho \left[\frac{G(\theta) - \theta}{\bar{V}^2(G(\theta) - \theta) - \theta} \left(1 - \bar{V}^2 \frac{G(\theta)}{\theta}\right) \right. \\ &\quad \times \frac{G(\theta) - \theta \dot{G}(\theta)}{(G(\theta) - \theta)^2} + \left. \frac{G(\theta)}{\theta(G(\theta) - \theta)} \right] L_{\beta}\theta \\ &= \frac{\rho}{(\theta - G(\theta))\{\bar{V}^2(G(\theta) - \theta) - \theta\}} \\ &\quad \times \{(\theta - \bar{V}^2 G(\theta))\dot{G}(\theta) + \bar{V}^2 G(\theta)\} L_{\beta}\theta. \end{aligned} \quad (84)$$

But if

$$(\theta - \bar{V}^2 G(\theta))\dot{G}(\theta) + \bar{V}^2 G(\theta) = 0, \quad (85)$$

then by Eq. (50) $\gamma = 1/\bar{V}^2$, namely Σ would be a characteristic of Einsteins's equations,¹⁰ therefore $L_{\beta}\theta|_{\Sigma} = 0$.

ACKNOWLEDGMENTS

I am indebted to A. M. Anile and M. Francaviglia for helpful discussions.

APPENDIX

Henceforth the vector field $V = \epsilon\lambda n + \mu$ will be chosen as the coordinate congruence with respect to a family of non-null hypersurfaces $\phi(p) = \text{const}$.

If $\xi = \epsilon\alpha n + \beta$ is another vector field, one has

$$L_{\xi}h_a^b = (\epsilon/\lambda)\{\alpha D_a\lambda - \lambda D_a\alpha\}n^b. \quad (A1)$$

If $\alpha = 0$ and Eqs. (20') and (21') hold true, by using Einstein's equations (12)-(15) and the equations

$$L_{\beta}h_{ab} = D_{ab}\{L(0)\} + ch_{ab}, \quad (A2)$$

$$L_{\beta}K_{ab} = D_{ab}\{L(0)\} + \frac{1}{2}cK_{ab}, \quad (A3)$$

it follows that

$$L_{\beta}D_a\psi = D_aL_{\beta}\psi, \quad (A4)$$

where ψ is a scalar.

$$\begin{aligned} L_{\beta}D_aU^b_c &= D_aL_{\beta}U^b_c + \frac{1}{2}h^{bd}K^e_c \\ &\quad \times \{D_e(L_{\beta}h_{ad}) + D_a(L_{\beta}h_{de}) - D_d(L_{\beta}h_{ae})\} \\ &\quad - \frac{1}{2}K^{bd}\{D_c(L_{\beta}h_{ad}) + D_a(L_{\beta}h_{cd}) - D_d(L_{\beta}h_{ac})\} \\ &= D_a^b\{L(1)\} + D_aL_{\beta}U^b_c, \end{aligned} \quad (A5)$$

where U^b_c is a tensor field tangent to $\phi(p) = \text{const}$.

$$\begin{aligned} L_{\beta}\tilde{R}_{ab} &= \frac{1}{2}h^{cd}\{D_cD_a(L_{\beta}h_{bd}) + D_cD_b(L_{\beta}h_{ad}) - D_cD_d(L_{\beta}h_{ab}) \\ &\quad - D_bD_a(L_{\beta}h_{cd})\} = D_{ab}\{L(2)\}, \end{aligned} \quad (A6)$$

$$L_{\beta}\tau = D\{L(2), \partial L(0)\} - c\tau, \quad (A7)$$

$$L_{\beta}t_a = D_a\{L(1), \partial L(1)\} - \frac{1}{2}ct_a, \quad (A8)$$

$$L_{\beta}\eta = D\{L(1), \partial L(1)\} - 2c\eta, \quad (A9)$$

$$L_{\beta}L_i\lambda = D\{L(1), \partial L(1) - cL_i\lambda\}, \quad (A10)$$

$$L_{\beta}L_i\tau = D\{L(3), \partial L(1)\} - \frac{5}{2}cL_i\tau, \quad (A11)$$

$$L_{\beta}L_i\eta = D\{L(2), \partial L(2)\} - \frac{7}{2}cL_i\eta, \quad (A12)$$

$$L_{\beta}D_a t^a = D\{L(2), \partial L(2)\} - \frac{3}{2}cD_a t^a. \quad (A13)$$

Putting $N = L_{\beta}\theta$ one has

$$L_{\beta}\gamma = D\{N(0)\}, \quad (A14)$$

$$L_{\beta}p = D\{N(0)\} + D\{L(2), \partial L(1)\} - cp, \quad (A15)$$

$$L_{\beta}\rho = D\{N(0)\} + D\{L(2), \partial L(1)\} - c\rho, \quad (A16)$$

$$L_\beta \pi = D \{N(0)\} + D \{L(2), \partial L(1)\} - c\pi, \quad (\text{A17})$$

$$L_\beta H_{ab} = D_{ab} \{N(0)\} + D_{ab} \{L(2), \partial L(1)\}, \quad (\text{A18})$$

$$L_\beta L_i p = D \{N(1)\} + D \{L(3), \partial L(2)\} - \frac{5}{2} c L_i p, \quad (\text{A19})$$

$$L_\beta L_i \rho = D \{N(1)\} + D \{L(3), \partial L(2)\} - \frac{5}{2} c L_i \rho, \quad (\text{A20})$$

$$L_\beta L_i \pi = D \{N(1)\} + D \{L(3), \partial L(2)\} - \frac{5}{2} c L_i \pi, \quad (\text{A21})$$

$$L_\beta D^b (\lambda H_{ab}) = D_a \{N(1)\} + D_a \{L(3), \partial L(2)\} - \frac{1}{2} c D^b (\lambda H_{ab}), \quad (\text{A22})$$

$$\partial \tau = \lambda H_{ab} K^{ab} - \epsilon \lambda \tau K - 2L_i \lambda - \lambda D^a t_a, \quad (\text{A23})$$

$$\partial \eta = 2 \{ -\epsilon \lambda K_{ab} t^a t^b - \epsilon \lambda \eta K + \epsilon \tau L_i \lambda - t^a D^b (\lambda H_{ab}) \}. \quad (\text{A24})$$

Equations (A23) and (A24) can be obtained by Einstein's equations and by

$$\begin{aligned} \partial D_a U^b_c &= D_a \partial U^b_c + \frac{1}{2} h^{bd} K_c^e \{ D_e (\partial h_{ad}) + D_a (\partial h_{de}) - D_d (\partial h_{ae}) \} \\ &\quad - \frac{1}{2} K^{bd} \{ D_c (\partial h_{ad}) + D_a (\partial h_{cd}) - D_d (\partial h_{ac}) \}, \end{aligned} \quad (\text{A25})$$

where U^b_c is a tensor field tangent to $\phi(p) = \text{const}$.

$$\begin{aligned} \partial \tilde{R}_{ab} &= \frac{1}{2} h^{cd} \{ D_c D_a (\partial h_{bd}) + D_c D_b (\partial h_{ad}) - D_c D_d (\partial h_{ab}) \\ &\quad - D_a D_b (\partial h_{cd}) \}. \end{aligned} \quad (\text{A26})$$

¹R. D. Blandford and C. F. McKee, *Phys. Fluids* **19**, 1130 (1976).

²G. Deb Ray and T. K. Chakraborty, *Astrophys. Space Sci.* **56**, 119 (1978).

³G. Deb Ray and T. K. Chakraborty, *Astrophys. Space Sci.* **61**, 81 (1979).

⁴E. P. T. Liang, *Ap. J.* **211**, 361 (1977).

⁵M. E. Cahill and A. H. Taub, *Commun. Math. Phys.* **21**, 1 (1971).

⁶B. Coll, *Ann. Inst. Henry Poincaré*, **XXV**, 393 (1976).

⁷S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-time* (Cambridge U. P., London, 1973).

⁸B. K. Berger, *J. Math. Phys.* **17**, 1268 (1976).

⁹The hypersurfaces $\phi(p) = \text{const}$ are chosen to be noncharacteristic surfaces of Einstein's equations, so that $v \neq 0$ (Ref. 10).

¹⁰A. Lichnerowicz, *Relativistic Hydrodynamics and Magnetohydrodynamics* (Benjamin, New York, 1967).

¹¹Henceforth the variables on M_- will be marked with a minus sign.

Liouville dynamics and Poisson brackets

G. Marmo ^{a)}

Istituto Nazionale di Fisica Nucleare, Sezione di Napoli
Istituto di Fisica Teorica, Università di Napoli, Naples, Italy

E. J. Saletan

Physics Department, Northeastern University, Boston, Massachusetts 02115

A. Simoni ^{a)}

Istituto di Fisica Teorica, Università di Napoli
Istituto Nazionale di Fisica Nucleare, Sezione di Napoli, Naples, Italy

F. Zaccaria ^{a)}

Istituto di Fisica Teorica, Università di Napoli
Istituto Nazionale di Fisica Nucleare, Sezione di Napoli, Naples, Italy

(Received 22 July 1980; accepted for publication 22 August 1980)

We study divergenceless dynamical systems from a differential geometrical point of view. The analogy with Hamiltonian mechanics is pursued even as far as Poisson brackets. In particular, we study Nambu mechanics and its generalizations.

PACS numbers: 05.20. — y, 02.40.Hw

I. INTRODUCTION

In many situations in physics one comes across divergenceless vector fields, i.e., those satisfying the equation

$$\operatorname{div} \mathbf{v} = 0. \quad (1)$$

This equation occurs, for example, in hydrodynamics, statistical mechanics, thermodynamics, magnetostatics. In the case of hydrodynamics, for instance, it expresses the condition of incompressibility (it is the continuity equation for the flow of an incompressible fluid flowing with velocity \mathbf{v}). More generally, if the fluid is compressible, the continuity equation becomes (we write ρ for the mass density)

$$\frac{d\rho}{dt} + \operatorname{div} \rho \mathbf{v} = 0, \quad (2)$$

which is of the same form, but generalized to four dimensions. Like this equation for the conservation of mass, the equation for the conservation of charge, namely

$$\frac{d\rho}{dt} + \operatorname{div} \mathbf{J} = 0,$$

is a generalization of (1) to four dimensions. Another example in higher dimension is the conservation of phase-space volume in Hamiltonian statistical mechanics:

$$\sum \left(\frac{\partial \dot{q}^i}{\partial q^i} + \frac{\partial \dot{p}_i}{\partial p_i} \right) = 0.$$

All of these examples involve a vector field $v = (v^1, \dots, v^n)$ on a manifold M . The vector field is the infinitesimal generator of a family of transformations of M , and these transformations leave a certain volume element invariant. That is, Eqs. (1), (2), and their generalizations to higher dimension express the invariance of a volume element.

Consider, for instance Eq. (2). Two equivalent ways of interpreting it are the following (in both of them the manifold M is \mathbb{R}^4). First, the vector field $v = (\rho, \rho v^1, \rho v^2, \rho v^3)$ preserves the volume element usually written as

$d\tau = dt dx^1 dx^2 dx^3$. Second, the vector field $v = (1, v^1, v^2, v^3)$ preserves the volume element $dm = \rho d\tau$. Both views, of course, merely state conservation of mass.

In this paper we discuss the generalization of these ideas to arbitrary differentiable manifolds, and we interpret the vector fields as dynamical systems on the manifolds. The volume elements, which we shall call Ω , are then regular m -forms on the manifold, where m is the dimension of the manifold under consideration. We wish to examine the geometrical properties of such volume preserving dynamical systems, which we shall call Liouville dynamical systems. This then furnishes a geometrical approach to all dynamical systems which satisfy some continuity condition. In particular, we draw the analogy between Hamiltonian and Liouville dynamics and study the analogs of Poisson brackets.

Section II discusses some of the simplest properties of Liouville dynamical systems, their invariant structures, diffeomorphisms and symmetries, and begins the discussion of Poisson brackets. Section III generalizes the idea of the Poisson bracket. Section IV applies the previous considerations to a particular class of Liouville systems which are called Nambu^{1,2} dynamical systems.

II. LIOUVILLE DYNAMICS

A. Definition and simplest properties

1. Let M be an orientable differential manifold of dimension $m = 2n$. We shall distinguish one vector field $\Delta \in \mathfrak{X}(M)$ on the manifold which we call the *dynamics*, *dynamical system*, or *dynamical vector field*. If M is a symplectic manifold, i.e., if there is given on it a symplectic (regular, closed) two-form ω , we say that Δ is Hamiltonian with respect to ω iff there exists a function $H \in \mathcal{F}(M)$ such that

$$i_{\Delta} \omega = -dH. \quad (3)$$

It is well known that ω is then *invariant under* Δ , i.e., that $L_{\Delta} \omega = 0$, where L is the Lie derivative. If, further, we write

^{a)}Postal address: Mostra d'Oltremare Pad. 19, 80125 Naples, Italy.

$$\omega^n = \omega \wedge \omega \wedge \dots \wedge \omega = \Omega, \quad (n \text{ factors})$$

it then follows that Ω is also invariant:

$$L_\Delta \Omega = 0. \quad (4)$$

This is known as Liouville's theorem; a Hamiltonian dynamics leaves the volume element invariant. (Note that ω^n , like any other regular m -form on M , is a volume element for M).

We shall say that a dynamics Δ is *Liouville with respect to an arbitrary volume Ω on M* (not necessarily with respect to ω^n) iff Ω is invariant under Δ , i.e., iff Eq. (4) is satisfied. This does not presuppose that M is symplectic; there need be no two-form on M such that $\Omega = \omega^n$, and in fact, $m = \dim M$ may from now on be odd. Then a dynamical system may be Liouville without being Hamiltonian.

In analogy with Hamiltonian systems and more precisely, Δ will be called *locally Liouville with respect to Ω* (locally Ω -Liouville) iff (4) is satisfied. Equation (4) implies that about each point $x \in M$ there is a neighborhood U in which there exists an $(m-2)$ -form $\Theta \in \Lambda^{m-2}(U)$ (or, briefly, *locally* there exists such a Θ) such that $i_\Delta \Omega|_U = -d\Theta$. If then Θ turns out to be a globally defined $(m-2)$ -form, we shall say that Δ is *globally Ω -Liouville*. That is, Δ is globally Ω -Liouville iff there exists a $\Theta \in \Lambda^{m-2}(M)$ such that

$$i_\Delta \Omega = -d\Theta. \quad (5)$$

Usually when we characterize a dynamics as Liouville we shall mean that it is locally Liouville. Clearly if a dynamics is globally Ω -Liouville it is also locally Ω -Liouville.

Let (M, ω) be a symplectic manifold, and $\mathcal{X}_{\mathcal{L}}(M)$ be the set of dynamical systems Liouville with respect to $\Omega = \omega^n$ on M . We have seen that $\mathcal{X}_{\mathcal{H}}(M) \supset \mathcal{X}_{\mathcal{L}}(M)$ (the set of Hamiltonian dynamical systems). Suppose that M is, like the phase space of a classical dynamical system, the cotangent bundle T^*Q of some configuration space Q and that F is a two-form on Q . Let ω_0 be the natural symplectic structure on T^*Q , and construct the two-form

$$\omega = \omega_0 + F.$$

(Note the abuse of notation: F is actually a two-form on Q .) Then it is clear that $\omega^n = \omega_0^n$, for each term of the form $\omega_0^k \wedge F^{n-k}$ has more than n factors of the form dq^i , and only n of them are independent. It follows then that if Δ and Δ' are Hamiltonian with respect to two different symplectic forms, they can nevertheless be Liouville with respect to the same Ω . Incidentally, symplectic structures like $\omega_0 + F$ lead to what Birkhoff³ calls *gyroscopic interactions* which, like the magnetic force, do no work. It is seen then that such interactions do not show up at the level of Liouville dynamics.

2. Suppose that Δ is *not* Liouville with respect to a given volume Ω . Is it possible to find another volume Ω' such that Δ is Ω' -Liouville? Note first that the set of m -forms on a manifold M of dimension m is itself of dimension one, and therefore that if Ω and Ω' are volumes, there exists a function f such that

$$\Omega' = f\Omega, \quad f \neq 0 \in \mathcal{F}(M). \quad (6)$$

Thus the question reduces to finding an f such that Δ is Liouville with respect to $f\Omega$.

It is not always possible to find such a function. In

general

$$\begin{aligned} L_\Delta \Omega' &= L_\Delta f \Omega = (L_\Delta f)\Omega + f L_\Delta \Omega \\ &= (L_\Delta f + f \operatorname{div}_\Omega \Delta)\Omega, \end{aligned}$$

where the function $\operatorname{div}_\Omega \Delta$ is defined by

$$L_\Delta \Omega = (\operatorname{div}_\Omega \Delta)\Omega \quad (7)$$

(observe that $L_\Delta \Omega$ is an m -form). Thus if Δ is Ω' -Liouville, f must satisfy the equation

$$L_\Delta f + f \operatorname{div}_\Omega \Delta = 0, \quad (8)$$

which does not always have solutions in $\mathcal{F}(M)$, the set of C^∞ functions over M .

As an illustration of a dynamical field which cannot be made Ω' -Liouville, consider

$$\Delta = x\partial/\partial x + y\partial/\partial y$$

on \mathbb{R}^2 , and $\Omega = dx \wedge dy$. Then $L_\Delta \Omega = 2\Omega$, and Eq. (8) becomes

$$x\partial f/\partial x + y\partial f/\partial y + 2f = 0.$$

To be a solution of this equation f would have to be "homogeneous" of degree -2 in x and y , and thus $f \notin \mathcal{F}(M)$, for f is undefined at $x = y = 0$.

B. Invariant structures

1. Let Δ be locally Liouville (henceforth when Ω is not specified, we are assuming that a given Ω has been chosen); then in the neighborhood of each $m \in M$ there exists an $(m-2)$ -form Θ such that (5) is satisfied locally, and $d\Theta$ is a local invariant of the motion, for

$$L_\Delta d\Theta = d(i_\Delta d\Theta) = -d(i_\Delta \Omega) = 0.$$

If in addition $i_\Delta \Theta$ is closed, Θ itself is invariant, for

$$L_\Delta \Theta = i_\Delta d\Theta + di_\Delta \Theta = -i_\Delta \Omega + di_\Delta \Theta.$$

Note the rough analogy with the Hamiltonian case. In that case $i_\Delta \omega = -dH$ implies that dH is an invariant, and moreover H itself is an invariant, for the analogue of $di_\Delta \Theta$ does not appear in the expression for $L_\Delta H$.

Let us write (5) in a *canonical chart*, obtaining what may be called, in analogy with Hamiltonian dynamics, Liouville's canonical equations. A canonical chart is one in which Ω can be written in the form

$$\Omega = dx^1 \wedge dx^2 \wedge \dots \wedge dx^m.$$

Let Θ in this chart be given by

$$\Theta = \sum_{i < j} \Theta_{ij} dx^1 \wedge dx^2 \wedge \dots \wedge d\bar{x}^i \wedge \dots \wedge d\bar{x}^j \wedge \dots \wedge dx^m,$$

where $d\bar{x}^k$ means that dx^k does not appear in the exterior product. Then

$$\begin{aligned} -d\Theta &= -\sum_k \frac{\partial \Theta_{ij}}{\partial x^k} dx^k \wedge dx^1 \wedge \dots \wedge d\bar{x}^i \wedge \dots \wedge d\bar{x}^j \wedge \dots \wedge dx^m \\ &= -\sum_{i < j} \left[\frac{\partial \Theta_{ij}}{\partial x^i} (-1)^{i+1} dx^1 \wedge \dots \wedge d\bar{x}^i \wedge \dots \wedge dx^m \right. \\ &\quad \left. + \frac{\partial \Theta_{ij}}{\partial x^j} (-1)^j dx^1 \wedge \dots \wedge d\bar{x}^i \wedge \dots \wedge dx^m \right]. \end{aligned}$$

This can be written in the form

$$-d\Theta = \sum_{ij} (-1)^i \frac{\partial \Theta_{ij}}{\partial x^i} dx^1 \wedge \dots \wedge d\bar{x}^j \wedge \dots \wedge dx^m, \quad (9)$$

where $\Theta_{jk} = -\Theta_{kj}$. On the other hand, if Δ^k is the k th component of Δ in the local chart, we have

$$i_{\Delta} \Omega = \sum \Delta^j (-1)^{j+1} dx^1 \wedge \dots \wedge d\bar{x}^j \wedge \dots \wedge dx^m,$$

so that Eq. (5) becomes

$$\Delta^j \equiv \dot{x}^j = (-1)^{j+1} \sum_i \frac{\partial \Theta_{ij}}{\partial x^i} (-1)^i. \quad (10)$$

These are Liouville's canonical equations. The expression on the right-hand side of this equation is sometimes called the divergence of $\Theta_{ij} (-1)^{i+j+1}$. If the bivector (actually, bivector field) with these components is called Σ (that is, $i_{\Sigma} \Omega = \Theta$), then this equation states that $\Delta = \text{div} \Sigma$.

Example 1: If $m = 2$, write $\Theta_{12} = -\Theta_{21} = -H$. Then

$$\dot{x}^1 = \frac{\partial \Theta_{21}}{\partial x^2} = \frac{\partial H}{\partial x^2},$$

$$\dot{x}^2 = -\frac{\partial \Theta_{12}}{\partial x^1} (-1) = -\frac{\partial H}{\partial x^1}.$$

This is the Hamiltonian case in dimension $m = 2$, or $n = 1$, so that, as expected, $\omega = \Omega$.

Example 2: If $m = 3$, write $\Theta_{12} = -f_3$, $\Theta_{23} = -f_1$, $\Theta_{31} = f_2$. Then

$$\dot{x}^1 = \frac{\partial f_3}{\partial x^2} - \frac{\partial f_2}{\partial x^3},$$

$$\dot{x}^2 = \frac{\partial f_1}{\partial x^3} - \frac{\partial f_3}{\partial x^1},$$

$$\dot{x}^3 = \frac{\partial f_2}{\partial x^1} - \frac{\partial f_1}{\partial x^2}.$$

In the usual notation for \mathbb{R}^3 this may be written $\dot{\mathbf{x}} = \nabla \times \mathbf{f}$.

Since Θ plays roughly the role in Liouville dynamics that H plays in Hamiltonian, one may ask what the analog is for replacing H by $H' = H + K$, where K is a constant. The analog is related to the fact that $dH' = dH$, and in the Liouville case this means that Θ can be replaced by $\Theta' = \Theta + \beta$ where $d\beta = 0$. Locally this means that β can be written in terms of an $(m-3)$ -form $\alpha: \beta = d\alpha$, and thus that the Θ_{ij} of Eq. (10) are not uniquely defined.

An example of this is classical electrodynamics. In the introduction it was pointed out that conservation of charge implies that the vector field $\Delta = (\rho, \mathbf{J})$ on \mathbb{R}^4 is a Liouville dynamics. This means that there exists a two-form \tilde{F} such that

$$i_{\Delta} \Omega = d\tilde{F},$$

where $\Omega = dt \wedge dx^1 \wedge dx^2 \wedge dx^3$. In fact it can be shown [see Eq. (10)] that if $F = F_{\mu\nu} dx^{\mu} \wedge dx^{\nu}$ is the two-form whose elements are the components of the electromagnetic field, and if \tilde{F} is its dual in the sense of Hodge (i.e., $\tilde{F}_{\mu\nu}$

$= g_{\mu\alpha} g_{\nu\beta} \epsilon^{\alpha\beta\gamma\delta} F_{\gamma\delta}$), then what is obtained are Maxwell's equations. As described above, however, \tilde{F} is not unique: one can add any two-form β such that $d\beta = 0$. This corresponds to adding an arbitrary source-free field to the electromagnetic field which corresponds to \tilde{F} .

2. Let $i_{\Delta} \Omega = \alpha$. Then Δ is (locally) Liouville, i.e., Δ satisfies (4), iff α is closed:

$$d\alpha = 0, \quad i_{\Delta} \Omega = \alpha. \quad (11)$$

Suppose now that there exist functions $f_1, \dots, f_{m-1} \in \mathcal{F}(M)$ such that

$$\alpha = df_1 \wedge \dots \wedge df_{m-1}. \quad (12)$$

Then α is closed, Δ is therefore Liouville, and moreover each of the f_k is a constant of the motion. Indeed,

$$\begin{aligned} i_{\Delta} i_{\Delta} \Omega &= 0 \\ &= \sum_k (-1)^{k+1} df_1 \wedge \dots \wedge (L_{\Delta} f_k) df_{k+1} \wedge \dots \wedge df_{m-1}. \end{aligned}$$

Since the df_k are independent (if $\alpha \neq 0$), $L_{\Delta} f_k = 0$ for each k . In a sense this is a different kind of analog of a Hamiltonian dynamics, one which is defined not by one, but by a set of $m-1$ Hamiltonians f_1, \dots, f_{m-1} . Each of these Hamiltonians is then a constant of the motion.

To a limited extent, something like the converse is true locally. Assume Δ to be Liouville and let $f \in \mathcal{F}(M)$ be a constant of the motion (i.e., satisfy $L_{\Delta} f = 0$). Then locally there exists an $(m-2)$ -form β such that $df \wedge \beta = \alpha$ is the $(m-1)$ -form of Eq. (11). Indeed,

$$0 = i_{\Delta} (df \wedge \beta) = (L_{\Delta} f) \beta - df \wedge i_{\Delta} \beta = df \wedge \alpha. \quad (13)$$

By taking f to be one of the coordinate functions in a local chart we see that α must be of the form (locally) $df \wedge \beta$. Suppose further that Δ has $m-1$ independent constants of the motion f_1, \dots, f_{m-1} . Then it follows that locally we may write

$$\alpha = F df_1 \wedge \dots \wedge df_{m-1}, \quad (14)$$

where (by closure) F depends only on f_1, \dots, f_{m-1} .

In a sense the f_k of Eqs. (12) and (14) provide an analogy with a Hamilton–Jacobi transformation for Δ , for if they are taken as $m-1$ of the m local coordinates, the integral curves of Δ lie along the m th coordinate direction.

C. Diffeomorphisms

1. A canonical diffeomorphism $\varphi: M \rightarrow M$ is a diffeomorphism which satisfies

$$\varphi^* \Omega = \Omega. \quad (15)$$

Again, the analogy with canonical diffeomorphisms (symplectomorphisms) of Hamiltonian dynamics, which satisfy $\varphi^* \omega = \omega$, is obvious. In general for a diffeomorphism φ , the function $\det_{\Omega} \varphi \in \mathcal{F}(M)$ is defined by

$$\varphi^* \Omega = (\det_{\Omega} \varphi) \Omega,$$

so that φ is canonical iff $\det_{\Omega} \varphi = 1$.

A symmetry of the field Δ is a diffeomorphism which leaves Δ invariant:

$$\varphi_* \Delta = \Delta. \quad (16)$$

If φ is a symmetry for the Liouville dynamics Δ , then $\det_{\Omega} \varphi$ is a constant of the motion, for

$$\begin{aligned} [L_{\Delta} (\det_{\Omega} \varphi)] \Omega &= L_{\Delta} [(\det_{\Omega} \varphi) \Omega] = L_{\Delta} \varphi^* \Omega \\ &= \varphi^* (L_{\varphi_* \Delta} \Omega) = \varphi^* (L_{\Delta} \Omega) = 0. \end{aligned} \quad (17)$$

This result leads to an interesting consequence. It associates the constant of the motion $\det_{\Omega}\varphi$ even with a discrete symmetry φ . Recall that a Hamiltonian system is necessarily Liouville, so this gives a way to associate a constant of the motion with a discrete symmetry also in the Hamiltonian case.

The converse of this result is not quite true, but if $\det_{\Omega}\varphi$ is a constant of the motion, then $\varphi_*\Delta$ is Ω -Liouville, as is obvious from (17), and Δ is $\varphi^*\Omega$ -Liouville. Such a diffeomorphism may be called *canonoid*.⁴

Let φ be canonical and a symmetry for $\alpha = i_{\Delta}\Omega$, i.e., $\varphi^*\alpha = \alpha$. Then φ is also a symmetry for Δ , for

$$\alpha = \varphi^*\alpha = \varphi^*(i_{\Delta}\Omega) = i_{\varphi_*\Delta}\Omega.$$

Since Ω is regular, $\varphi_*\Delta = \Delta$. The Hamiltonian analog of this is: if φ is canonical and leaves H invariant, then it is a symmetry for Δ . The Liouville case involves not the analogue of H [i.e., not \mathcal{O} of Eq. (15)], but the analog of dH .

Remark: According to Koopman⁵ a canonical diffeomorphism $\varphi:M\rightarrow M$ generates a unitary transformation on the particular Hilbert space of functions on M which is obtained from the measure Ω . Suppose now that Δ is Liouville with respect to two volumes Ω_1 and Ω_2 . The flow associated with Δ then generates a one-parameter group of transformations which is canonical with respect to both Ω_1 and Ω_2 , and therefore one obtains two unitary representations of this group on the two Hilbert spaces one can construct from the two measures. Let $\psi:M\rightarrow M$ be such that $\psi^*\Omega_1 = \Omega_2$. Then ψ generates the intertwining operator between the two representations. Thus canonoid transformations generate intertwining operators between unitary representations. A similar result is true for groups of transformations canonical with respect to two measures Ω_1 and Ω_2 also if these groups are more general than one-parameter groups.

2. The definitions and assertions concerning diffeomorphisms have infinitesimal versions, and these infinitesimal versions refer to vector fields rather than to diffeomorphisms.

Let $X\in\mathfrak{X}(M)$ generate a one-parameter group φ_t^X of diffeomorphisms. Then φ_t^X is canonical iff X is Ω -Liouville, for in an obvious way $(\varphi_t^X)^*\Omega = \Omega$ iff $L_X\Omega = 0$.

A vector field X is called an *infinitesimal symmetry* for Δ (or simply a *symmetry* when no confusion will arise) iff $[X,\Delta] = 0$. Then if X is symmetry for Δ , it follows that $L_X\Omega$ is invariant under Δ (i.e., that $\text{div}_{\Omega} X$ is a constant of the motion). Indeed,

$$L_{\Delta}L_X\Omega = L_{[\Delta,X]}\Omega = 0. \quad (18)$$

As in the finite case, the converse is not quite true, but if $L_X\Omega$ is invariant under Ω , then $[\Delta,X]$ is Ω -Liouville, as is obvious from (18).

Let X be Ω -Liouville and a symmetry for $\alpha = i_{\Delta}\Omega$, i.e. $L_X\alpha = 0$. Then X is also a symmetry for Δ , for

$$i_{[X,\Delta]}\Omega = L_Xi_{\Delta}\Omega = -L_X\alpha = 0,$$

and since Ω is regular, $[X,\Delta]$ must be zero. The Hamiltonian analogue of this statement is that if X is a Hamiltonian field which leaves invariant a function H , it is a symmetry for the

dynamical field Δ whose Hamiltonian is H . Again, in the Liouville case it is the analogue of dH which enters, rather than H itself.

The roles of X and Δ can be interchanged in this demonstration. That is, if $i_X\Omega = \beta$ and $L_{\Delta}\beta = 0$, then $[X,\Delta] = 0$ and X is a symmetry for Δ . This statement also has its obvious Hamiltonian analogue. Actually, the condition $L_{\Delta}\beta = 0$ is stronger than needed. It is easily shown that if $dL_{\Delta}\beta = 0$, then $[X,\Delta] = 0$, and a similar weaker statement works also in the Hamiltonian case.

D. First discussion of Poisson brackets

1. The usual (intrinsic) definition of Poisson brackets (PB) in Hamiltonian dynamics is the following. Let $f, g \in \mathcal{F}(M)$ and define $X_f \in \mathfrak{X}(M)$ by $i_{X_f}\omega = -df$. Then the PB of f with g is defined by

$$\{f, g\} = i_{X_f}i_{X_g}\omega = -\omega(X_f, X_g) \in \mathcal{F}(M). \quad (19)$$

As is well known, the PB so defined is antisymmetric, satisfies the Jacobi identity (because ω is closed) and is nondegenerate in the sense that if $\{f, g\} = 0 \forall g$, then f is a constant. Because this definition depends in the way it does on ω , it is difficult to generalize to Liouville dynamics, so we shall use another, also intrinsic, based on work by Cartan, Jost, Pauli, and Flanders.^{6,7}

The second definition also depends on ω , but in a different way. Let $f, g \in \mathcal{F}(M)$, as before. Then the PB of f with g may be defined by

$$\{f, g\}\omega^n = n(df \wedge dg) \wedge \omega^{n-1}, \quad (20)$$

where the powers of ω are with respect, of course, to the exterior product. It can be shown⁷ that this definition agrees with that of Eq. (19), but because it involves the volume it is more suitable for generalization to Liouville dynamics.

2. Before turning to a detailed discussion of PBs in the next section, we wish to point out a similar intrinsic definition of the Lie derivative of a function, which is related to the PB definition of Eq. (20).

Let $f \in \mathcal{F}(M)$ and $X \in \mathfrak{X}(M)$. Then since $i_X\Omega$ is an $(m-1)$ -form if Ω is a volume, there exists a function $g \in \mathcal{F}(M)$ such that

$$df \wedge i_X\Omega = g\Omega.$$

From $df \wedge \Omega = 0$ it follows that

$$0 = i_X(df \wedge \Omega) = (i_Xdf)\Omega - df \wedge i_X\Omega = (L_Xf)\Omega - g\Omega,$$

so that g is the Lie derivative of f . Thus the Lie derivative can be defined by

$$(L_Xf)\Omega \equiv df \wedge i_X\Omega. \quad (21)$$

The connection between Eqs. (20) and (21) is the following. Let $\Delta \in \mathfrak{X}(T^*Q)$ be Hamiltonian and let its Hamiltonian function be $h \in \mathcal{F}(T^*Q)$:

$$i_{\Delta}\omega = -dh.$$

Then for any $f \in \mathcal{F}(T^*Q)$ and for $\Omega = \omega^n$

$$\begin{aligned} (L_{\Delta}f)\Omega &= df \wedge i_{\Delta}\Omega = ndf \wedge i_{\Delta}\omega \wedge \omega^{n-1} \\ &= -ndf \wedge dh \wedge \omega^{n-1} = -\{f, h\}\Omega. \end{aligned}$$

In other words, $L_{\Delta}f = -\{f, h\}$, a well known result.

These definitions and the analogy between Liouville and Hamiltonian dynamics can be used to define a PB not between functions, but between a function and an $(m - 2)$ -form. Let Δ be globally Liouville, i.e. satisfy (5) with a certain $(m - 2)$ -form Θ . Then for any $f \in \mathcal{F}(M)$,

$$0 = i_{\Delta}(df \wedge \Omega) = (L_{\Delta}f)\Omega + df \wedge d\Theta.$$

If the analog of the PB of f with Θ is defined by

$$\{f, \Theta\} \Omega \equiv -df \wedge d\Theta, \quad (22)$$

it follows that

$$L_{\Delta}f = \{f, \Theta\}. \quad (23)$$

Thus if a Liouville dynamics Δ is defined by its *Liouville* $(m - 2)$ -form Θ through Eq. (4) in the same way as a Hamiltonian dynamics is defined by its Hamiltonian function through Eq. (3), the Lie derivative of any $f \in \mathcal{F}(M)$ is given by Eq. (23), which is obviously similar to its Hamiltonian analog.

In a local coordinate chart this analog of the PB may be written [use $df = \Sigma (\partial f / \partial x^k) dx^k$ and Eq. (9)]

$$\{f, \Theta\} = \sum_{ij} (-1)^{i+j} \frac{\partial f}{\partial x^j} \frac{\partial \Theta_{ij}}{\partial x^i}.$$

It is easily shown that for $m = 2$ and $\Theta_{12} = h$, this becomes the usual expression in Hamiltonian dynamics for the PB of a function f with the Hamiltonian in a canonical chart.

Remark: In terms of what was called Σ around Eq. (10), this analog of the PB may be written in the form of the scalar product $\text{grad}f \cdot \text{div}\Sigma$.

In Hamiltonian dynamics a PB on closed one-forms is sometimes defined⁸ in the following way. From Eq. (19) one has

$$i_{[X_r, X_s]}\omega = L_{X_r}i_{X_s}\omega = -d\{f, g\}.$$

The PB is then defined by $\{df, dg\} = -d\{f, g\}$, or

$$\{df, dg\} = i_{[X_r, X_s]}\omega.$$

Then it follows that

$$L_X dg = -L_X i_{X_r}\omega = -\{df, dg\}.$$

The analog in Liouville dynamics is a PB on closed $(m - 1)$ -forms defined as follows. Let

$$i_X \Omega = \Theta_X, \quad i_Y \Omega = \Theta_Y,$$

[with $d\Theta_X = d\Theta_Y = 0$, so that X and Y are in $\mathcal{X}_{\text{closed}}$. Because the map $\Omega: \mathcal{X}_{\text{closed}} \rightarrow \Lambda^{m-1}_{\text{closed}}: X \mapsto i_X \Omega$ is a Lie-algebra homomorphism, the expression

$$\{\Theta_X, \Theta_Y\} = i_{[X, Y]}\Omega$$

has the properties of a PB. Moreover

$$L_X \Theta_Y = L_X i_Y \Omega = i_{[X, Y]}\Omega = \{\Theta_X, \Theta_Y\};$$

the Lie derivative is related to this PB in the same way as the Lie derivative is related to the analogous PB of Hamiltonian dynamics.

III. POISSON BRACKETS

1. Further generalizations of the PB will be based on the observation that in both the Hamiltonian and Liouville cases the PB was defined through a volume element (m) -form Ω and an $(m - 2)$ -form which we shall call α . For two func-

tions $f, g \in \mathcal{F}(M)$, for example (and it is only this example that will be generalized), the PB was given by an equation of the form

$$\{f, g\} \Omega = (df \wedge dg) \wedge \alpha. \quad (24)$$

In Eq. (20), Ω was ω^n , and α was $n\omega^{n-1}$. Thus for fixed Ω , all one need do to define a generalized PB is to choose a suitable α and insert it into (24). What constitutes suitability for α is not yet clear. For example, although the PB defined through (24) is bound to be antisymmetric, it will not satisfy the Jacobi identity unless α fulfills certain conditions which will be discussed later in terms of bivectors. Moreover, the PB as defined through (24) will often be degenerate in the sense described after Eq. (19). This will also be discussed in what follows.

2. A volume element provides an isomorphism between k -vectors and $(m - k)$ -forms. For example, the map

$$\Omega: \mathcal{X}(M) \rightarrow \Lambda^{m-1}(M): X \mapsto i_X \Omega$$

is an isomorphism. More generally, so is the map

$$\Omega: \Lambda_k(M) \rightarrow \Lambda^{m-k}(M): \lambda \mapsto i_{\lambda} \Omega,$$

where

$$\Lambda_k(M) = \{\Sigma X_1 \wedge X_2 \wedge \dots \wedge X_k \mid X_i \in \mathcal{X}(M)\}$$

is the set of k -vectors. (The contraction of a k -vector with a form is defined in the usual way like contraction of tensors, except that one must take into account their antisymmetry.) This allows one to discuss PBs in terms of bivectors instead of $(m - 2)$ -forms: a bivector leads uniquely to an $(m - 2)$ -form which in turn leads to a PB in accordance with (24). The suitability of the $(m - 2)$ -form can then be analyzed in terms of the suitability of the bivector.

Let $\lambda \in \Lambda_2(M)$ be a bivector. Then λ can be used to define a PB between function in two different ways. First, by using (24) and choosing $\alpha = i_{\lambda} \Omega$, so that

$$(f, g)_{\lambda} \Omega \equiv (df \wedge dg) \wedge (i_{\lambda} \Omega). \quad (25)$$

Second, by using the fact that λ defines a linear antisymmetric map $\mathcal{F}(M) \times \mathcal{F}(M) \rightarrow \mathcal{F}(M)$ in accordance with

$$f, g \mapsto \{f, g\}_{\lambda} = i_{\lambda}(df \wedge dg). \quad (26)$$

At first glance (25) and (26) seem to give the same PB, and in fact they do, though the proof is not as easy as it may seem. We defer the proof until later, and for a while we use only (26). In any case, these are candidates for PBs provided only that λ satisfies the conditions of suitability yet to be discussed.

Example 3: Consider the bivector $w \in \Lambda_2(T^*Q)$ satisfying

$$i_w \Omega = n\omega^{n-1}.$$

It is easily shown that in a local canonical chart

$$w = \sum_i \frac{\partial}{\partial p_i} \wedge \frac{\partial}{\partial q^i}, \quad (27)$$

and that the PB it yields in accordance with (26) is the usual one of Hamiltonian dynamics.

Since bilinearity and antisymmetry accrue to the PB from (26) alone, the only condition to be placed on λ is that the PB satisfy the Jacobi identity. Consider two bivectors

$$\lambda = \sum_i X_i \wedge Y_i \quad \text{and} \quad \mu = \sum_i W_i \wedge Z_i,$$

and define (note that the ranges of i and j may be different)

$$\begin{aligned} [\lambda, \mu] = & \sum_{ij} \{ [X_i, W_j] \wedge Y_i \wedge Z_j + [Y_i, Z_j] \wedge X_i \wedge W_j \\ & - [X_i, Z_j] \wedge Y_i \wedge W_j - [Y_i, W_j] \wedge X_i \wedge Z_j \\ & + (\text{div} X_i) Y_i \wedge W_j \wedge Z_j - (\text{div} Y_i) X_i \wedge W_j \wedge Z_j \\ & + (\text{div} W_j) X_i \wedge Y_i \wedge Z_j - (\text{div} Z_j) X_i \wedge Y_i \wedge W_j \}. \end{aligned}$$

Then it has been shown^{9,10} that the PB of (26) satisfies the Jacobi identity iff $[\lambda, \lambda] = 0$. This condition is trivially satisfied by the bivector w of Eq. (27).

3. Having defined PBs in terms of two-forms, one can generalize further to something like a PB, but which maps not two, but a larger number k of functions into $\mathcal{F}(M)$. In fact let β be a k -vector and consider the map

$$\mathcal{F}_1(M) \times \dots \times \mathcal{F}_k(M) \rightarrow \mathcal{F}(M):$$

$$f_1, \dots, f_k \mapsto (f_1, \dots, f_k)_\beta$$

defined by

$$(f_1, \dots, f_k)_\beta \Omega \equiv i_\beta \Omega \wedge f_1 \wedge \dots \wedge f_k. \quad (28)$$

For $k = 2$ this reduces to Eq. (25). This definition of a more general PB is, incidentally, independent of Ω .

An example of such a generalized PB is the following. In a canonical chart let β be given by

$$\beta = F \frac{\partial}{\partial x^1} \wedge \dots \wedge \frac{\partial}{\partial x^k},$$

where $F \in \mathcal{F}(M)$. Then

$$(f_1, \dots, f_k)_\beta = F \sum_{i_1, \dots, i_k} \varepsilon_{i_1, \dots, i_k} \frac{\partial f_1}{\partial x^{i_1}} \dots \frac{\partial f_k}{\partial x^{i_k}}.$$

It is not immediately evident that this extension of the idea of a PB has much to do with dynamical systems, although it will be seen in the next section that it can be applied in Nambu mechanics. The PB of Eq. (28) is, in any case, antisymmetric in each pair of functions.

IV. NAMBU TYPE MECHANICS

1. Consider $m - 1$ independent functions H_1, \dots, H_{m-1} in $\mathcal{F}(M)$, where M , as before, has dimension m . These define an obviously Liouville dynamics Δ in accordance with

$$i_\Delta \Omega = dH_1 \wedge dH_2 \wedge \dots \wedge dH_{m-1}. \quad (29)$$

This shall be called a Nambu type of dynamical system.

Each of the H_j , according to the discussion around Eqs. (11) and (12), is a constant of the motion for Δ , and if $f \in \mathcal{F}(M)$, then according to (21) $df/dt = L_\Delta f$ is given by

$$(L_\Delta f) \Omega = df \wedge dH_1 \wedge \dots \wedge dH_{m-1}.$$

Now let W be the m -vector dual to Ω :

$$i_W \Omega = 1;$$

then contraction with W gives

$$L_\Delta f = df/dt = i_W (df \wedge dH_1 \wedge \dots \wedge dH_{m-1}).$$

We want to show that this is $(f, H^1, \dots, H_{m-1})_W$ in accordance with the definition (28). If f is not independent of the

H_k , both expressions are zero, so it follows trivially.

If f is independent of the H_k , then locally one can write

$$W = \rho \frac{\partial}{\partial f} \wedge \frac{\partial}{\partial H_1} \wedge \dots \wedge \frac{\partial}{\partial H_{m-1}},$$

$$\Omega = \frac{1}{\rho} df \wedge dH_1 \wedge \dots \wedge dH_{m-1},$$

$\rho \in \mathcal{F}(M)$. Then the result follows immediately. Thus we arrive at

$$L_\Delta f = (f, H_1, \dots, H_{m-1})_W. \quad (30)$$

It is seen that this is quite similar to the situation in Hamiltonian dynamics. The analog of Ω is the symplectic form ω of Hamiltonian dynamics, and the analogue of W is then the bivector w dual to ω , given by $i_w \omega = 1$. This bivector is then that of Eq. (27), and indeed the usual PB may be written in the form $\{f, g\} = (f, g)_w$. In keeping with this analogy, we shall call the H_j of Eq. (30) *Hamiltonian functions*.

Now consider any one of the Hamiltonian functions, say H_1 , and define the bivector G_1 by

$$i_{G_1} \Omega = dH_2 \wedge \dots \wedge dH_{m-1}. \quad (31)$$

It will now be shown that for $f \in \mathcal{F}(M)$

$$L_\Delta f = \{H_1, f\}_{G_1} \quad (32)$$

[see Eq. (26)].

The proof depends on showing first that

$$i_{G_1} dH_k = 0, \quad k \neq 1. \quad (33)$$

Indeed, let $\theta_1, \theta_2 \in \mathcal{X}_1^*(N)$ be such that for some $g \in \mathcal{F}(M)$

$$\Omega = g \theta_1 \wedge \theta_2 \wedge dH_2 \wedge \dots \wedge dH_{m-1}.$$

Then

$$\begin{aligned} i_{G_1} \Omega = & g (i_{G_1} \theta_1 \wedge \theta_2) dH_2 \wedge \dots \wedge dH_{m-1} \\ & + g \theta_1 \wedge \alpha_1 \\ & + g \theta_2 \wedge \alpha_2 + g \theta_1 \wedge \theta_2 \wedge (i_{G_1} dH_2 \wedge \dots \wedge dH_{m-1}), \end{aligned}$$

where neither α_1 nor α_2 have any factors of θ_1 or θ_2 . According to (31) this is equal to $dH_2 \wedge \dots \wedge dH_{m-1}$, and therefore because θ_1, θ_2 , and the dH_k are all independent, the last three terms add to zero. But then each one of these last three terms must be zero, for the first has no factor of θ_2 , the second no factor of θ_1 , and the last has both. Finally since α_1 contains no factors of θ_1 and since $g \theta_1 \wedge \alpha_1 = 0$, it follows that $\alpha_1 = 0$. Similarly $\alpha_2 = 0$, and

$$i_{G_1} (dH_2 \wedge \dots \wedge dH_{m-1}) = i_{G_1} i_{G_1} \Omega = 0, \quad (34)$$

Now consider α_1 . This $(m - 3)$ -form is a sum of terms each of which is the product of a function $r_k = i_{G_1} \theta_2 \wedge dH_k$, $k \in \{2, \dots, m - 1\}$ with an $(m - 3)$ -form $dH_{l_1} \wedge \dots \wedge dH_{l_{m-3}}$, $l_i \neq k$. These terms are all independent, since each has just one of the dH_k missing; and since their sum vanishes, each term vanishes separately. A similar statement may be made for α_2 . Thus

$$i_{G_1} \theta_2 \wedge dH_k = i_{G_1} \theta_1 \wedge dH_k = 0, \quad k \in \{2, \dots, m - 1\}.$$

A similar argument starting from Eq. (34), rather than from $\alpha_1 = 0$ then shows that

$$i_{G_1} dH_k \wedge dH_l = 0, \quad k \neq l.$$

It follows that the vector field $i_{G_k} dH_k$ satisfies Eq. (33), as asserted.

Now let us calculate $L_{\Delta} f$. According to (21)

$$(L_{\Delta} f)\Omega = df \wedge dH_1 \wedge \dots \wedge dH_{m-1} = df \wedge dH_1 \wedge i_{G_1} \Omega.$$

Note that the right-hand side is $(f, H_1)_{G_1} \Omega$ in accordance with Eq. (25). That is, in proving (32) we will also be proving the equivalence of (25) and (26), that is, that $\{f, g\}_{\lambda} = (f, g)_{\lambda}$.

Proceeding, we have

$$\begin{aligned} 0 &= i_{G_1} (df \wedge dH_1 \wedge \Omega) \\ &= i_{G_1} (df \wedge dH_1) \Omega + df \wedge dH_1 \wedge i_{G_1} \Omega + Z, \end{aligned}$$

where Z vanishes by Eq. (33) because it consists of terms all of which contain factors of $i_{G_k} dH_k$, $k \in \{2, \dots, m-1\}$. Thus

$$(L_{\Delta} f)\Omega = -i_{G_1} (df \wedge dH_1) \Omega = \{H_1, f\}_{G_1} \Omega,$$

which proves Eq. (32).

More generally, if G_k is defined by an equation similar to (31), but with dH_k missing, then

$$L_{\Delta} f = (H_k, f)_{G_k} = \{H_k, f\}_{G_k}. \quad (35)$$

It may thus seem that a Nambu–Liouville dynamics [i.e., one satisfying (29)] leads inevitably to a Hamiltonian dynamics given by (35). But in fact Eq. (35) does not yield a Hamiltonian dynamics, for the PB in it is degenerate in the sense described after Eq. (19): from $i_{G_k} dH_l = 0$ for $k \neq l$ it follows that $(H_l, f)_{G_k} = 0$ for $k \neq l$ and for all $f \in \mathcal{F}(M)$. In the seemingly Hamiltonian dynamics whose PB is $(\ , \)_{G_k}$, many Hamiltonian functions lead to null dynamical systems.

On the other hand these PBs satisfy the Jacobi identity.

$$\begin{aligned} i_{\Delta} \Omega &= \frac{B-C}{A} qrdq \wedge dr + \frac{C-A}{B} rpd r \wedge dp + \frac{A-B}{C} pqdp \wedge dq \\ &= \frac{1}{4} \left[\frac{B-C}{A} dq^2 \wedge dr^2 + \frac{C-A}{B} dr^2 \wedge dp^2 + \frac{A-B}{C} dp^2 \wedge dq^2 \right] \\ &= \frac{1}{4ABC} [A^2 dp^2 \wedge (Bdq^2 + Cdr^2) + B^2 dq^2 \wedge (Cdr^2 + Adp^2) \\ &\quad + C^2 dr^2 \wedge (Adp^2 + Bdq^2)] \\ &= \frac{1}{4ABC} [A^2 dp^2 + B^2 dq^2 + C^2 dr^2] \wedge [Adp^2 + Bdq^2 + Cdr^2] = dH_1 \wedge dH_2, \end{aligned} \quad (38)$$

where

$$\begin{aligned} H_1 &= A^2 p^2 + B^2 q^2 + C^2 r^2, \\ H_2 &= \frac{1}{4ABC} (Ap^2 + Bq^2 + Cr^2). \end{aligned}$$

[Incidentally, it is easily seen from Eq. (38) that $L_{\Delta} \Omega = di_{\Delta} \Omega$ vanishes.] It follows from (38) that H_1 and H_2 are constants of the motion [see the discussion around Eqs. (11) and (12)], i.e., that $L_{\Delta} H_1 = L_{\Delta} H_2 = 0$. These two constants are essentially the square of the angular momentum and the energy of the rotator.

The dynamical system under consideration is now seen to be of the Nambu type, and it follows that for any $f \in \mathcal{F}(M)$

This follows from Kirillov,⁹ where it is shown that if A is a k -vector and B is an l -vector, then

$$i_{[A, B]} \alpha = (-1)^{kl} i_A di_B \alpha + (-1)^k i_B di_A \alpha$$

for any closed α . If $A = B = G_k$ and $\alpha = \Omega$, this equation becomes

$$i_{[G_k, G_k]} \Omega = (-1)^{4+2} i_{G_k} di_{G_k} \Omega + (-1)^2 i_{G_k} di_{G_k} \Omega = 0,$$

since $di_{G_k} \Omega = 0$. Since Ω is an isomorphism, $[G_k, G_k] = 0$, which is a necessary and sufficient condition for the PB defined through G_k to satisfy the Jacobi identity.

2. As a more detailed example of the Nambu–Liouville formalism consider a rigid body rotating freely about a fixed point. Let the principal moments of the body be A, B, C , and let the angular velocity have components p, q, r in the principal axis systems. Then Euler's equations for the motions of a rigid body (actually for the components of the angular velocity vector) are

$$A\dot{p} = (B-C)qr, \quad B\dot{q} = (C-A)rp, \quad C\dot{r} = (A-B)pq.$$

Let M be \mathbb{R}^3 and let p, q, r form a Cartesian chart on M . Then this system is described by the dynamical vector field Δ which can be written in the form

$$\Delta = \frac{B-C}{A} qr \frac{\partial}{\partial p} + \frac{C-A}{B} rp \frac{\partial}{\partial q} + \frac{A-B}{C} pq \frac{\partial}{\partial r}. \quad (36)$$

Because M is of dimension three, it cannot be symplectic, but it is easily verified that

$$L_{\Delta} \Omega = 0,$$

where

$$\Omega = dp \wedge dq \wedge dr. \quad (37)$$

Thus Δ is Ω -Liouville. Let us calculate

$$\frac{df}{dt} \equiv L_{\Delta} f = (f, H_1, H_2)_W,$$

where

$$W = \frac{\partial}{\partial p} \wedge \frac{\partial}{\partial q} \wedge \frac{\partial}{\partial r}$$

is the trivector dual to Ω .

In accordance with (31) the bivector G_2 associated with H_2 is given by

$$i_{G_2} \Omega = dH_2 = 2(A^2 pdp + B^2 qdq + C^2 rdr), \quad (39)$$

This is an algebraic equation for the coefficients of the three basic bivector components of G_2 , whose solution is easily

found to be

$$G_2 = 2A^2 p \frac{\partial}{\partial q} \wedge \frac{\partial}{\partial r} + 2B^2 q \frac{\partial}{\partial r} \wedge \frac{\partial}{\partial p} + 2C^2 r \frac{\partial}{\partial p} \wedge \frac{\partial}{\partial q} \quad (40)$$

Equation (39) guarantees that $di_{G_2}\Omega = 0$, and then according to Lichnerowicz¹⁰ $[G_2, G_2] = 0$ and the generalized PB defined by G_2 satisfies the Jacobi identity. Equation (35) then implies that for any $f \in \mathcal{F}(M)$,

$$df/dt \equiv L_{\Delta} f = (H_2 f)_{G_2} = i_{G_2} dH_2 \wedge df.$$

This can be written in the form

$$\frac{df}{dt} = \frac{1}{2ABC} (E, f)_{G_2} \equiv \{E, f\}, \quad (41)$$

where $E = H_2/(2ABC)$ is the energy of the rotator. For example,

$$\begin{aligned} \dot{p} = \{E, p\} &= \frac{1}{2ABC} i_{G_2} dE \wedge dp \\ &= \frac{1}{2ABC} i_{G_2} (Bqdq \wedge dp + Crdr \wedge dp) \\ &= \frac{1}{2ABC} [-2C^2 Brq + 2CB^2 rq] \\ &= \frac{B-C}{A} rq, \end{aligned}$$

where reproduces, as it should, the first of Euler's equations. It is easily seen, in addition, that $\{p, q\} = Cr/(AB)$, and other similar results can be obtained by more or less obvious cyclic permutations.

It may thus seem that we have arrived at a Hamiltonian dynamics for the rigid rotator: Eq. (41) defines a generalized PB such that the time derivative (the Lie derivative along the dynamics) of any function is given by its PB with the energy function E . In fact, of course, the dynamics is not really Hamiltonian, and for two reasons. The first is that there is no two-form ω on M such that $i_{\Delta} \omega = dE$. The second is that the generalized PB is not regular. Indeed, a simple calculation will show that

$$\{H_1, f\} = \frac{1}{2ABC} i_{G_2} dH_1 \wedge df = 0 \forall f \in \mathcal{F}(M),$$

in agreement with Eq. (33).

Other proposals by Nambu¹ involve contracting a three-vector with a two-form to obtain a vector field and hence a derivation. Of course this procedure is easily generalized to contracting a k -vector with a $(k-1)$ -form.

V. CONCLUSION

We have discussed Liouville dynamics from the geometrical point of view and have investigated the way in which the idea of the PB can be generalized and then extended from the Hamiltonian to the Liouville domain. This involved defining the PB in terms of bivectors, and we were thus led in a natural way to further generalizing the PB in terms of k -vectors and thence to applications to Nambu mechanics. It may be of interest that bivectors can also be used to define foliations¹⁰ and hence that PBs can be related to

foliations.

There are, in addition, other possible ways these extended ideas about PB's can be used. For example, let Ω be a volume on $\mathbb{R} \times M$ defined by

$$\Omega = dt \wedge df_1 \wedge \dots \wedge df_m,$$

where $f_i \in \mathcal{F}(\mathbb{R} \times M)$. Let $\{g, h\}$ be given by

$$\{g, h\} \Omega \equiv dt \wedge dg \wedge df_2 \wedge \dots \wedge df_{m-1} \wedge dh.$$

If the f_i form a local canonical chart for M , this becomes

$$\{g, h\} = \frac{\partial g}{\partial f_1} \frac{\partial h}{\partial f_m} - \frac{\partial g}{\partial f_m} \frac{\partial h}{\partial f_1},$$

which one might call a *partial* PB. Other partial PB's can be generated in the same way. Without going into further detail, we state that this kind of procedure can be applied in the time-dependent Hamiltonian formalism, for which $df_1 \wedge \dots \wedge df_m \equiv \omega$.

A similar procedure may be used with constrained systems. For them, however, dt is replaced in the definition of Ω by the k -form $dr_1 \wedge \dots \wedge dr_k$, where r_1, \dots, r_k are the k constraint functions. Then Ω is given by $dr_1 \wedge \dots \wedge dr_k \wedge df_1 \wedge \dots \wedge df_{m-k}$, and the partial PB's are defined similarly to the way they are defined in the time-dependent case.

At the end of Sec. IIB1 we mentioned that $i_{\Delta} \Omega = d\Theta$ is not an equation that defines Θ uniquely, and we illustrated it with the example of Maxwell's equations for the electric-current four-vector $\Delta = (\rho, \mathbf{J})$. This can be thought of a sort of gauge transformation, and therefore begins to indicate a geometrical framework for treating certain kinds of gauge transformations.

¹Y. Nambu, "Generalized Hamiltonian Dynamics," Phys. Rev. D 7, 2405 (1973).

²F. Bayen, M. Flato, "Remarks concerning Nambu's generalized mechanics," Phys. Rev. D 11, 3049 (1975); I. Cohen, "Generalization of Nambu mechanics," Int. J. Theor. Phys. 12, 69 (1975); G. J. Ruggeri, "The Nambu Mechanics as a class of Singular Generalized Dynamical formalism," Int. J. Theor. Phys. 12, 287 (1975); N. Mukunda and E. C. G. Sudarshan, "Relation between Nambu and Hamiltonian mechanics," Phys. Rev. D 13, 2846 (1976); A. J. Kalnay and R. Tascò, "Lagrange, Hamiltonian-Dirac and Nambu mechanics," Phys. Rev. D 17, 1552 (1978).

³G. D. Birkhoff, *Dynamical Systems*, Coll. Publ. IX, 2nd ed. (Am. Math. Soc., Providence, R. I. 1966).

⁴G. Marmo and E. J. Saletan, "Ambiguities in the Lagrangian and Hamiltonian formalism: transformation properties," Nuovo Cimento 40 B, 67 (1977).

⁵B. Koopman, "Hamiltonian systems and transformations in Hilbert space," Proc. Nat. Acad. Sci. 17, 315 (1931).

⁶E. Cartan, *Leçons Sur les Invariants Intégraux* (Hermann, Paris, 1922). W. Pauli, "On the Hamiltonian Structure of non-local field theories," Nuovo Cimento 10, 648 (1953). R. Jost, "Poisson Brackets (an unpedagogical lecture)," Rev. Mod. Phys. 36, 572 (1964).

⁷H. Flanders, *Differential Forms with Applications to the Physical Sciences*, (Academic, New York, 1963).

⁸R. Abraham and J. E. Marsden, *Foundation of Mechanics* (Benjamin Reading, Mass., 1978), Chap. 3.

⁹A. A. Kirillov, "Invariant Differential operators on geometrical quantities," J. Funct. Anal. Appl. (in Russian) 11, 39 (1977).

¹⁰A. Lichnerowicz, "Fibrés vectoriels, structures unimodulaires exactes et automorphismes infinitésimaux," J. Math. Pur. Appl. 56, 183 (1977); "Les variétés de Poisson et leurs algèbres de Lie associées," J. Diff. Geom. 12, 253 (1977).

Passivity and equilibrium for classical Hamiltonian systems

H. A. M. Daniëls

Instituut voor Theoretische Natuurkunde, Rijksuniversiteit Groningen Postbus 800, Groningen, The Netherlands

(Received 3 September 1980; accepted for publication 26 November 1980)

For classical continuous n -particle systems equilibrium states are characterized by a condition of passivity.

PACS numbers: 05.20. — y, 05.70. — a

INTRODUCTION

In statistical mechanics one describes the equilibrium states by a density function (operator) of the form $\rho = e^{-\beta H}/Z$. The aim of this paper is to give for classical systems a justification for this from the second law of thermodynamics: If a system is in equilibrium, no work is performed by the system if the external parameters are varied in a certain (cyclic) way.

The notion of passivity has been introduced in Ref. 1 and the equivalence of equilibrium, KMS, and complete passivity established for abstract C^* -dynamical systems, where, as in fermion and lattice systems, the dynamics is given by a strongly continuous one-parameter group of automorphisms. In a related paper² the notion of passivity is discussed, for finite quantum spin systems, in detail. This problem has not been treated yet for continuous classical systems. Therefore we will consider here classical continuous systems consisting of n -point masses, and realize a program analogous to the one in Ref. 2.

The main part of this paper deals with the characterization of passive states by a very simple condition on the density function (Theorem 1). The rest of the results follow from this Theorem and investigations similar to those given in Ref. 2. As far as we know, the proof of Theorem 1 is essentially new. In Ref. 1 part of the results is proved relying on the algebraic structure for KMS states as obtained in Ref. 3, whereas in this paper we rather use symplectic geometry.

1. NOTATION AND DEFINITIONS

In this section we fix the notation and quote some standard results on Hamiltonian mechanics; for more details see, e.g., Refs. 4–6.

A n -particle classical system S is described by (M^{2k}, Ω, H) , where M^{2k} is a symplectic manifold (the cotangent bundle of the configuration manifold), Ω the canonical symplectic 2-form, and H the Hamilton function on M^{2k} . [One could think of $M^{2k} = (\mathbb{R}^d \times \mathbb{R}^d)^n$ if the particles move freely in d -dimensional Euclidean space or $M^{2k} = (T^d \times \mathbb{R}^d)^n$ if the particles are enclosed in a box with periodic boundary conditions.] In the following we write M for M^{2k} . λ will denote the Liouville measure on M . Let us introduce some function spaces on M .

$C^\infty(M)$ is the set of (real) C^∞ -functions on M .

$C_0^\infty(M)$ is the subset of functions with compact support.

$C^0(M)$ is the set of (complex) continuous functions which have a limit at infinity.

$L^2(M, \lambda)$ is the set of (complex) L^2 -functions with respect to λ . \mathfrak{A} is the C^* -algebra of (complex) continuous functions on \bar{M} , where \bar{M} is the one point compactification of M . As a set, \mathfrak{A} equals $C^0(M)$.

The Poisson bracket is a bilinear map $\{ , \} : C_0^\infty(M) \times C_0^\infty(M) \rightarrow C_0^\infty(M)$ such that for all $f, g, h \in C_0^\infty(M)$:

(p1) Skew-symmetry: $\{f, g\} = -\{g, f\}$;

(p2) Jacobi identity:

$\{\{f, g\}, h\} + \{\{h, f\}, g\} + \{\{g, h\}, f\} = 0$;

(p3) Leibnitz rule: $\{fg, h\} = \{f, h\} \cdot g + \{g, h\} \cdot f$.

$\lambda \text{ diff}(M)$, resp. $S \text{ diff}(M)$, is the set of λ -(volume)-preserving, resp. symplectic (Ω -preserving), diffeomorphisms on M . $\lambda \text{ diff}_0(M)$ and $S \text{ diff}_0(M)$ are the subsets consisting of diffeomorphisms which differ from the identity only in a compact subset of M . $S \text{ diff}_0(M) \sim \text{id}$ will denote the C^∞ -connected component of the identity in $S \text{ diff}_0(M)$. The following are in one-to-one correspondence:

$H \in C^\infty(M)$ locally defined;

X_H local Hamiltonian vector field with Hamilton function H ;

$L_H : C^\infty(M) \rightarrow C^\infty(M)$ the Liouville operator defined by $L_H(f) = \{f, H\}$;

$(\phi_t)_{t \in \mathbb{R}}$ the phase flow, a one-parameter group in $S \text{ diff}(M)$ defined for small $|t|$.

The correspondence is defined by the following formulas:

$$\left. \frac{d}{dt} \right|_{t=0} \phi_t(x) = X_H(x),$$

$$\left. \frac{d}{dt} \right|_{t=0} f(\phi_t(x)) = (L_H f)(x) = \{f, H\}(x).$$

Given a Hamilton function H , the equations of motion in local coordinates read:

$$\dot{p}_i = - \frac{\partial H}{\partial q_i}, \quad p_i(0) = p_{i0},$$

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad q_i(0) = q_{i0}.$$

Write $(p, q) = x$. The solution of these equations $x_t = \phi_t(x_0)$ defines the phase flow ϕ_t . If the Hamiltonian is time-dependent, we write $x_t = \phi_{t,s}(x_s)$, where $\phi_{t,s}$ is the flow from s to t along the trajectories of the solutions of the equations of motion with initial condition $x(s) = x_s$. $\phi_{t,s}$ is no longer a one-parameter group, but one has $\phi_{t,s} = \phi_{t,u} \circ \phi_{u,s}$. If $\phi \in \lambda \text{ diff}(M)$, a unitary linear operator $\phi^* : L^2(M, \lambda) \rightarrow L^2(M, \lambda)$ is defined by transposition:

$$\phi^*(f) = f \circ \phi.$$

A state ω of the system S is a normalized positive linear functional on \mathfrak{A} :

$$\omega(f) = \int_M f d\mu,$$

where μ is a normalized, regular, Borel measure on M . We only consider states which are given by a positive density function ρ :

$$\frac{d\mu}{d\lambda} = \rho \in L^1(M, \lambda).$$

Perturbations of the dynamics: We assume that the field X_H is full and therefore the corresponding phase flow ϕ_t is defined for all $t \in \mathbb{R}$. This guarantees that all differential equations considered in the sequel have solutions defined for all $t \in \mathbb{R}$.

Definition 1: A perturbation is a family $(h_t)_{t \in \mathbb{R}}$ in $C_0^\infty(M)$ such that:

- (h1) $(t, x) \rightarrow h_t(x)$ is smooth;
- (h2) $h_t = 0$ for $t \notin (0, T)$;

- (h3) $\bigcup_{t \in \mathbb{R}} \text{supp}(h_t)$ is contained in a compact subset of M .

The perturbed phase flow corresponding to the time-dependent Hamiltonian $H + h_t$ is denoted by $\psi_{t,s}$.

2. PASSIVITY, COMPLETE PASSIVITY, AND STATEMENT OF THE RESULTS

Given the perturbation h_t , we define (cf. Ref. 1)

$$l^h := \int_0^T \psi_{t,0}^* \left(\frac{dh_t}{dt} \right) dt \in \mathfrak{A}.$$

Definition 2: ω is called a passive state if for all perturbations h_t $\omega(l^h) \geq 0$.

Theorem 1: ω is passive if and only if ρ is decreasing with respect to H , i.e., for all $x, y \in M$

$$H(x) > H(y) \Rightarrow \rho(x) \leq \rho(y). \quad (1)$$

Corollary: A Gibbs state is passive.

For every $m \in \mathbb{N}$ we consider the system $S^m = (M^m, \Omega^m, H^m)$, where

$$M^m = \underbrace{M \times M \times \dots \times M}_{m \text{ times}}$$

$$\Omega^m = \bigoplus^m \Omega,$$

$$H^m = H \otimes 1 \otimes \dots \otimes 1 + \dots + 1 \otimes \dots \otimes 1 \otimes H.$$

If ω is a state on \mathfrak{A} with density ρ , then $\omega^{\otimes m}$ is the state on $\mathfrak{A}^{\otimes m}$ with density function ρ^m defined by

$$\omega^{\otimes m}(f_1 \otimes \dots \otimes f_m) = \omega(f_1) \cdot \omega(f_2) \cdot \dots \cdot \omega(f_m),$$

$$\rho^m(x_1, x_2, \dots, x_m) = \rho(x_1) \cdot \dots \cdot \rho(x_m).$$

Definition 3: ω is completely passive if $\omega^{\otimes m}$ is passive state for the system S^m for all $m \in \mathbb{N}$.

Theorem 2: ω is completely passive if and only if $\rho = e^{-\beta H} / Z$ with $0 \leq \beta < \infty$.

3. PROOFS OF THEOREMS 1 AND 2

Let us first compute $\omega(l^h)$. We start with a perturbation h_t of the Hamiltonian H . Let ϕ_t respectively $\psi_{t,s}$ denote the phase flows corresponding to H respectively $H + h_t$. Then by partial integration (cf. Ref. 1)

$$l^h = - \int_0^T \left(\frac{d}{dt} \psi_{t,0}^* \right) h_t dt. \quad (1a)$$

Define

$$\gamma_t = \phi_{-t} \psi_{t,0} \quad (2)$$

so that

$$\gamma_t^* = \psi_{t,0}^* \phi_{-t}^*.$$

Bearing in mind the formulas

$$\frac{d}{dt} \phi_t^* = \phi_t^* L_H,$$

$$\frac{d}{dt} \psi_{t,0}^* = \psi_{t,0}^* L_{H+h_t},$$

$$\phi^* \{f, g\} = \{\phi^* f, \phi^* g\} \quad \text{for symplectic } \phi,$$

we find

$$\begin{aligned} \frac{d}{dt} \gamma_t^* &= \frac{d}{dt} (\psi_{t,0}^* \phi_{-t}^*) \\ &= \psi_{t,0}^* L_{H+h_t} \phi_{-t}^* - \psi_{t,0}^* \phi_{-t}^* L_H \\ &= \gamma_t^* \phi_t^* (L_{H+h_t} - L_H) \phi_{-t}^* \\ &= \gamma_t^* L_{\phi_t^* h_t}. \end{aligned} \quad (3)$$

On the other hand,

$$\begin{aligned} \left(\frac{d}{dt} \psi_{t,0}^* \right) h_t &= \psi_{t,0}^* L_H(h_t) = \gamma_t^* \phi_t^* L_H(h_t) \\ &= \gamma_t^* L_H(\phi_t^* h_t) = -\gamma_t^* L_{\phi_t^* h_t}(H) \end{aligned}$$

and therefore

$$\left(\frac{d}{dt} \psi_{t,0}^* \right) h_t = - \frac{d}{dt} \gamma_t^*(H).$$

Hence $l^h = \gamma_T^*(H) - H$ and therefore ω is passive iff for all perturbations h_t

$$\omega(\gamma_T^*(H) - H) \geq 0. \quad (4)$$

We now characterize the set of γ_T 's which can be obtained from a perturbation h_t . From (2) it is clear that $\gamma_T \in \mathcal{S} \text{diff}(M)$, (3) and (h3) imply $\gamma_T \in \mathcal{S} \text{diff}_0(M)$. Since $\gamma_0 = \text{id}$ and γ_t satisfies (3), $\gamma_T \in \mathcal{S} \text{diff}_0(M) \sim \text{id}$. A very large class \mathcal{E} of γ_T 's is obtained by choosing h_t suitable. Take $T = 1$ and let $p \in \mathbb{N}$. Let $\alpha_j \in C^\infty([0, 1] \rightarrow \mathbb{R})$ such that $\text{supp}(\alpha_j) \subset [(j-1)/p, j/p]$, $\alpha_j((j-1)/p) = 0$ and $\alpha_j(j/p) = 1$, $j = 1, 2, \dots, p$. Let $g_j \in C_0^\infty(M)$ arbitrary, $j = 1, \dots, p$. Define

$$h_t = \sum_{j=1}^p \alpha_j'(t) \phi_{-t}^*(g_j).$$

Now (3) reads

$$\frac{d}{dt} \gamma_t^* = \gamma_t^* \alpha_j'(t) L_{g_j} \quad \text{for } \frac{j-1}{p} \leq t \leq \frac{j}{p}.$$

This equation can easily be solved, yielding

$$\gamma_1 = \exp(X_{g_p}) \circ \dots \circ \exp(X_{g_1}) \quad (5)$$

and

$$\gamma_t^* = \exp(L_{g_t}) \circ \dots \circ \exp(L_{g_0}),$$

where $\exp(X_g)$ denotes the time 1 map corresponding to the Hamiltonian vector field X_g with Hamilton function g . Let \mathcal{E} denote the set of γ 's defined by (5). It follows from Ref. 7 that if $H^{-1}(M, \mathbb{R}) = 0$, the class \mathcal{E} coincides with $S \text{diff}_0(M) \sim \text{id}$. So, if $H^{-1}(M, \mathbb{R}) = 0$ [e.g., $M = (\mathbb{R}^d \times \mathbb{R}^d)^n$], one has:

Lemma 1: ω is passive iff (4) holds for all $\gamma \in S \text{diff}_0(M) \sim \text{id}$.

Lemma 2: A passive state is invariant.

Proof: Take $\gamma_t = \exp(tX_g)$, then (4) implies

$$\left. \frac{d}{dt} \right|_{t=0} \omega(\gamma_t^*(H) - H) = 0.$$

Hence

$$\omega(\{g, H\}) = 0 \quad \text{for all } g \in C_0^\infty(M),$$

yielding the invariance of ω .

Remark: Using the classical KMS condition one shows easily that (4) holds for small $|t|$ if ω is a KMS state.⁸ The problem is to show that (4) holds for all t .

Proof of Theorem 1: For the sake of simplicity we restrict ourselves to the case where $M = (\mathbb{R}^d \times \mathbb{R}^d)^n$ and ρ continuous. (This proof can be generalized easily for $\rho \in L^1$.)

Suppose ρ satisfies (1). To prove passivity, we prove, in view of Lemma 1, the *a priori* stronger assertion $\omega(\gamma^*(H) - H) \geq 0$ for all $\gamma \in \lambda \text{diff}_0(M)$. Suppose there exists $\gamma \in \lambda \text{diff}_0(M)$ such that

$$\int_M [H(x) - H(\gamma(x))] d\mu(x) = \epsilon > 0.$$

This will lead to a contradiction. Let K be a cube in M such that $\gamma = \text{id}$ on K^c . Define

$$A := \sup_{\substack{x, y \in K \\ x \neq y}} \frac{\|\gamma^{-1}(x) - \gamma^{-1}(y)\|}{\|x - y\|}$$

and

$$B := \max_{x \in K} |H(x)|$$

$\|\cdot\|$ is just the Euclidean norm on $(\mathbb{R}^d \times \mathbb{R}^d)^n$. Since ρ is uniformly continuous on the cube K , there exists $\eta > 0$ such that for all $x, y \in K$

$$\|x - y\| < \eta \Rightarrow |\rho(x) - \rho(y)| < \frac{\epsilon}{8B\lambda(K)}.$$

Divide the cube K into N small cubic cells C_1, C_2, \dots, C_N of equal λ -measure: $\lambda(C_i) = \lambda(K)/N$ and $\text{diam}(C_i) = [\lambda(K)/N]^{1/r} \cdot \sqrt{r}$, where $r = 2dn$. Choose N large enough to ensure

$$\left| \int_K H(x) d\mu(x) - \sum_{i=1}^N H(x_i) \mu(C_i) \right| < \frac{\epsilon}{8}, \quad (6)$$

$$\left| \int_K H(\gamma(x)) d\mu(x) - \sum_{i=1}^N H(x_i) \mu(\gamma^{-1}(C_i)) \right| < \frac{\epsilon}{8}, \quad (7)$$

$$(A+1) \left(\frac{\lambda(K)}{N} \right)^{1/r} \cdot \sqrt{r} \leq \eta \quad (8)$$

for all possible choices of x_1, \dots, x_N with $x_i \in C_i$. Using the

mean value theorem, we can choose $x_i \in \text{int}(C_i)$ such that

$$\mu(C_i) = \int_{C_i} \rho(x) d\lambda = \rho(x_i) \lambda(C_i), \quad i = 1, \dots, N,$$

and then for all i and j

$$H(x_i) > H(x_j) \Rightarrow \mu(C_i) < \mu(C_j). \quad (9)$$

We claim that there exists a permutation Π of $\{1, 2, \dots, N\}$ such that

$$C_{\Pi(i)} \cap \gamma^{-1}(C_i) \neq \emptyset, \quad i = 1, 2, \dots, N.$$

Indeed this follows from a theorem of Hall.¹⁰ The Hall condition is fulfilled because γ^{-1} is λ -(volume)-preserving. Define $U_i = C_{\Pi(i)} \cup \gamma^{-1}(C_i)$; then, using (8),

$$\text{diam}(U_i) \leq (A+1) \text{diam}(C_i) \leq \eta$$

and therefore

$$\begin{aligned} & \left| \mu(\gamma^{-1}(C_i)) - \mu(C_{\Pi(i)}) \right| \\ &= \left| \int_{\gamma^{-1}(C_i)} \rho(x) d\lambda - \int_{C_{\Pi(i)}} \rho(x) d\lambda \right| \\ &\leq \sup_{x, y \in U_i} |\rho(x) - \rho(y)| \cdot \lambda(U_i) \\ &\leq \frac{\epsilon}{8B\lambda(K)} \cdot 2\lambda(C_i) = \frac{\epsilon}{4BN}. \end{aligned}$$

Hence

$$\left| \sum_{i=1}^N H(x_i) \mu(\gamma^{-1}(C_i)) - \sum_{i=1}^N H(x_i) \mu(C_{\Pi(i)}) \right| \leq \frac{\epsilon}{4}. \quad (10)$$

Combining (6), (7), and (10), we obtain

$$\sum_{i=1}^N H(x_i) \mu(C_i) - \sum_{i=1}^N H(x_i) \mu(C_{\Pi(i)}) \geq \frac{\epsilon}{2} > 0.$$

This contradicts (9) (cf. Refs. 11 and 2).

Conversely, let ω be passive and suppose ρ does not satisfy (1). Then there exist $x, y \in M$ such that

$$H(x) > H(y) \quad \text{and} \quad \rho(x) > \rho(y). \quad (11)$$

Since both H and ρ are continuous, there are small cells $C_1 \ni x, C_2 \ni y$ such that (11) holds on the cells. We now construct $\gamma \in \lambda \text{diff}_0(M)$ which interchanges the cells C_1 and C_2 and $\lambda(\{x \in M \mid \gamma(x) \neq x \text{ and } x \notin C_1 \cup C_2\})$ is very small. If $M = \mathbb{R}^2$, one takes $\gamma = \exp X_g$, where X_g is the vector field of Fig. 1. The field goes to zero in the shaded area. In the higher dimensional case one proceeds as follows. Take a 2-dim sym-

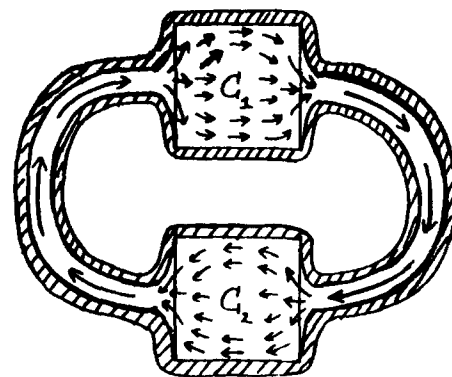


FIG. 1.

plectic plane P through x and y on which the symplectic 2-form is nondegenerate. Then take C_1 and C_2 to be two symplectomorphic cylindrical cells symplectic-orthogonal to the 2-dim plane P . Take γ restricted to P the transformation of Fig. 1 and then extend to the symplectic orthogonal complement of the plane P . Following this method, one can construct γ with

$$\omega(\gamma^*(H) - H) < 0,$$

which contradicts the passivity of ω . □

Proof of Theorem 2: In view of Theorem 1 complete passivity is equivalent to the condition

$$H(x_1) + \dots + H(x_m) > H(y_1) + \dots + H(y_m) \\ \Rightarrow \rho(x_1) \cdots \rho(x_m) \leq \rho(y_1) \cdots \rho(y_m)$$

for all $(x_1, \dots, x_m) \in M^m$ and $(y_1, \dots, y_m) \in M^m$. The rest of the proof runs like the proof of Theorem 7 in Ref. 2. □

Remark: The “only if” part of Theorem 2 can also be proved using the techniques of Ref. 8.

Note added in proof: After this paper was submitted we received a preprint by J. Górecki and W. Pusz containing similar results obtained by different methods.

ACKNOWLEDGMENTS

I am indebted to H. Maassen and F. Takens for useful

discussions and to M. Winnink for asking the relevant questions and a critical reading of the manuscript.

- ¹W. Pusz and S. L. Woronowicz, “Passive States and KMS states for General Quantum Systems,” *Commun. Math. Phys.* **58**, 273–90 (1978).
- ²A. Lenard, “Thermodynamical Proof of the Gibbs Formula for Elementary Quantum Systems,” *J. Stat. Phys.* **19**, 575–86 (1978).
- ³R. Haag, N. M. Hugenholtz, and M. Winnink, *Commun. Math. Phys.* **5**, 215 (1967).
- ⁴V. I. Arnold, *Mathematical Methods of Classical Mechanics*, *Gretu Texts in Mathematics Vol. 60* (Springer, New York-Heidelberg-Berlin, 1978).
- ⁵W. Thirring, *Lehrbuch der Mathematischen Physik* (Springer, New York-Vienna, 1977).
- ⁶A. M. Vinogradov and B. A. Kupershmidt, “The Structures of Hamiltonian Mechanics,” *Russian Math. Surv.* **32**, 177–243 (1977).
- ⁷A. Banyaga, “Sur la structure du groupe des difféomorphismes qui préservent une forme symplectique,” *Comment Math. Helv.* **53**, 174–227 (1978).
- ⁸B. Demoen, P. Vanheuverzwijn, and A. Verbeure, “Energetically stable systems,” *J. Math. Phys.* **19** 2256–9 (1978).
- ⁹After we finished the manuscript M. Aizenman showed us a proof of the “if” part of the theorem for $\rho, H \in L^2(X, \mu)$ (X measure space), using integral representations for measurable functions.
- ¹⁰M. Hall, Jr., *Combinatorial Theory* (Blaisdell, Waltham, Mass., 1967), Theorem 5.1.1.
- ¹¹G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities* (Cambridge U. P., Cambridge, 1934), Theorem 368.

General solutions to inverse transport problems

R. Sanchez and N. J. McCormick

Department of Nuclear Engineering, University of Washington, Seattle, Washington 98195

(Received 25 March 1980; accepted for publication 22 August 1980)

A new approach is developed for solving time-independent inverse problems for particle or radiation transport described by the monoenergetic linear Boltzmann equation. For a homogeneous plane geometry medium, the approach leads to a set of inverse solutions which can be obtained by purely algebraic means; previously derived methods appear as special cases. All these solutions require only measurements of the angular intensities at the surface. The analysis is extended to time-dependent and energy-dependent problems.

PACS numbers: 05.60. + w

I. INTRODUCTION

The purpose of solutions to inverse transport problems is to characterize the unknown material properties of a target. Such solutions fall into two categories, invasive and non-invasive. Only the former requires that the intensity be measured at all interior positions, while both depend upon the ingoing and outgoing intensities on the surface. The noninvasive solutions show much more promise for use in experiments.

Invasive solutions have been developed for multienergy group transport in a homogeneous slab target¹ and for one-group transport in a homogeneous target of any shape.² For a slab target with at most quadratically-anisotropic scattering, the coefficients of the scattering function can be determined in a noninvasive manner from azimuthally-independent moments of the incident and emerging distributions,^{3,4} any number of scattering coefficients can be obtained with an azimuthally-dependent incident beam,⁵ as verified numerically for up to 19th-order scattering.⁶

A general approach for obtaining inverse methods for the monoenergetic transport of photons and neutrons is developed in Sec. II. The homogeneous plane geometry medium is analyzed in Sec. III, and results obtained earlier³⁻⁵ are recovered in a direct manner; procedures for obtaining other solutions by purely algebraic manipulations are also presented. Using the Laplace transform method in Sec. IV, the analysis is broadened in scope to include the possibility that the incident distribution is time-dependent; the resulting inverse method does not require any inverse Laplace transform.

The plane geometry analysis is also extended in Sec. IV to energy-dependent problems for neutrons. Solutions are developed for the slowing down region and for cases where the scattering kernel obeys detailed balance. For the slowing down case, only in-group cross sections can be obtained; for the thermal energy problem, only spectrum-averaged coefficients result. The multigroup model is also analyzed, and results analogous to those of Siewert¹ are obtained.

II. GENERAL INVERSE SOLUTIONS

Consider the monochromatic linear transport equation describing the stationary equilibrium of particles or radiation within a convex domain D with incident radiation on its surface ∂D :

$$\left. \begin{aligned} B\psi &= S, & \text{in } D \times (4\pi) \\ \psi &= \psi_b, & \text{on } \partial D \times (2\pi)_- \end{aligned} \right\} \quad (1)$$

Here $\psi(\mathbf{r}, \Omega)$ represents the radiation intensity at point \mathbf{r} in direction Ω , $S(\mathbf{r}, \Omega)$ is the external volumetric source, and ψ_b is the incident radiation on the boundary; the term $(2\pi)_-$ refers to the hemisphere of incident directions, i.e., $\Omega \cdot \mathbf{n} < 0$, at a given point of the surface ∂D with external normal \mathbf{n} .

The linear transport operator

$$B = \Omega \cdot \nabla + K \quad (2)$$

is composed of the streaming operator $\Omega \cdot \nabla$ and of the operator K which describes the interaction of radiation with the medium,

$$K = \sigma - H. \quad (3)$$

Here $\sigma(\mathbf{r})$ is the total cross section, and H is the scattering operator defined by

$$(H\psi)(\mathbf{r}, \Omega) = \int_{(4\pi)} \sigma_s(\mathbf{r}, \Omega' \rightarrow \Omega) \psi(\mathbf{r}, \Omega') d\Omega', \quad (4)$$

where the differential cross section σ_s contains all angular information about the interactions.

The medium is assumed isotropic so that the scattering operator is invariant under rotation; therefore, H can be diagonalized using spherical harmonics,

$$H = \sum_{k>0} h_k Q_k. \quad (5)$$

Here Q_k is the orthogonal projection on the invariant subspace generated by the set of spherical harmonics $\{Y_k^l; |l| \leq k\}$, i.e.:

$$Q_k f(\Omega) = \sum_{|l| \leq k} Y_k^l(\Omega) \int_{(4\pi)} Y_k^{-l}(\Omega') f(\Omega') d\Omega'.$$

The constants h_k , in conjunction with σ , uniquely characterize the interaction of the monochromatic radiation with the medium. They can be written in terms of more familiar quantities as

$$h_k = \omega_k / (2k + 1),$$

where the ω_k 's are 4π times the coefficients of the expansion of σ_s on the Legendre polynomials of argument $\Omega \cdot \Omega'$.

The object of an inverse method is to provide a scheme for the computation of the coefficients σ and h_k in terms of angular intensities which are assumed known. One straightforward derivation of such a method follows by multiplying

transport Eq. (1) by $Y'_k(\Omega)$, and then integrating the resulting equation in $V \times (4\pi)$ over some volume V of homogeneous material²:

$$\sigma - h_k = \frac{\int_{(4\pi)} Y'_k(\Omega)(S_V(\Omega) - \Omega \cdot \psi_{\partial V}(\Omega)) d\Omega}{\int_{(4\pi)} Y'_k(\Omega)\psi_V(\Omega) d\Omega} \quad (6)$$

Thus only the spatially-averaged quantities

$$S_V(\Omega) = \int_V S(\mathbf{r}, \Omega) d\mathbf{r},$$

$$\psi_V(\Omega) = \int_V \psi(\mathbf{r}, \Omega) d\mathbf{r} \text{ and } \psi_{\partial V}(\Omega) = \int_{\partial V} \psi(\mathbf{r}, \Omega) dS$$

need to be known.

For an infinite medium containing a localized monodirectional plane source, Eq. (6) takes on an especially simple form.⁷ In plane geometry it is also possible to generalize the ideas leading to Eq. (6) by multiplying transport equation (1) by $z^n Y'_k(\Omega)$, where z is the spatial variable and n is a nonnegative integer. Such a procedure was used to obtain an inverse method for a finite slab without using information about the angular dependence of the distribution inside the medium,¹ and was used to obtain a set of inverse methods for the case of an infinite medium containing a localized, monodirectional plane source.⁸

All such inverse methods relying upon Eq. (6) imply that one must measure the intensity inside the volume V , a requirement that introduces the complications of locating a probe inside the medium and correcting for perturbations in $\psi(\mathbf{r}, \Omega)$. Therefore we will restrict our efforts to methods that use noninvasive measurements.

Typically the data for a noninvasive inverse method will be obtained by irradiating the surface of the body with some monochromatic intensity ψ_b , and by measuring the outgoing angular intensity. The effects of a possible volume source can be eliminated by subtracting the values of two measurements taken before and during irradiation, respectively, so we will limit our analysis to the case $S = 0$ in Eq. (1).

We proceed now to outline a general approach for the derivation of noninvasive inverse methods. We first define the scalar product

$$(f, g) = \int_{D \times (4\pi)} fg d\mathbf{r} d\Omega, \quad (7)$$

where f and g are arbitrary functions of \mathbf{r} and Ω . After integration by parts we obtain our basic equation,

$$(f, Bg) = (B^*f, g) + \langle f, g \rangle, \quad (8)$$

where B^* is the formal adjoint of the monochromatic transport operator,

$$B^* = -\Omega \cdot \nabla + K \quad (9)$$

and the surface contribution

$$\langle f, g \rangle = \int_{\Gamma \times (4\pi)} [fg] \Omega \cdot dS d\Omega \quad (10)$$

extends over the set Γ of surfaces of discontinuity of fg . In deriving Eq. (8) it has been assumed that f and g have at most

a finite number of discontinuities of the first kind, $[f, g]$ representing the jump at the discontinuity.

Notice that by appropriately selecting f and g as functions of the angular intensity ψ , Eq. (8) can be reduced to only the surface contribution, i.e., to an equation relating values of ψ at the boundary; this can potentially produce a noninvasive inverse method. With this in mind, we put

$$f = L\psi, \quad g = \psi \quad (11)$$

and choose the operator L , not necessarily linear, so that the volume contribution to Eq. (8) identically vanishes,

$$(B^*L\psi, \psi) = 0. \quad (12)$$

The resulting equation,

$$\langle L\psi, \psi \rangle = 0 \quad (13)$$

depends only on the values of ψ at the boundary. Thus, although, the operator L can act on the \mathbf{r} variable, $(L\psi)(\mathbf{r}, \Omega)$ evaluated at $\mathbf{r} \in \Gamma$ must depend only on values of $\psi(\mathbf{r}, \Omega)$ on Γ . Furthermore, since only values of ψ are accessible, L must not contain any spatial derivative except, of course, those along the boundary. Note that a simple, linear form of L is obtained by taking any integral operator in the Ω variable.

For each operator L satisfying the previous requirements, Eq. (13) will yield a noninvasive inverse method provided that L depends on the cross sections or, equivalently, on the operator K . It is obvious that when Γ comprises interior boundaries, not accessible to measurement, this inverse method will involve unknown intensity values; since ψ itself is continuous, the discontinuities of $fg = (L\psi)\psi$ originate from those of L and, ultimately, from the discontinuities of the cross sections. Thus, in the case of a body with internal discontinuities the inverse method so defined will not be complete, and supplementary equations will have to be added in order to estimate the values of the intensity at the discontinuities. In any case, the first problem posed by nonuniform bodies is the detection of the presence of discontinuities by noninvasive measurements. In the present work we will put aside this problem by dealing only with the case of a homogeneous medium.

We end this section by briefly discussing the method of solving Eq. (12). Since the angular intensity ψ is not known inside the body, we look only for operators L satisfying $B^*L\psi = 0$. In general, any solution L of this equation will contain spatial derivatives, but these spatial derivatives may be partially or totally eliminated by using the fact that ψ is the solution of the transport equation, i.e., $-\Omega \cdot \nabla = K$ when acting on ψ . This *a posteriori* elimination leads to a complicated procedure involving lengthy manipulations of the equations.³⁻⁵ Instead, we opt to perform such eliminations before calculating L .

The basic idea for satisfying $B^*L\psi = 0$ consists of permuting with L the spatial derivative component of B^* and then using the relations $\Omega \cdot \nabla = B - K$ and $B\psi = 0$ to get rid of the spatial derivative terms. In doing so, the most general equation one obtains is

$$B^*L + \tilde{L}B = 0, \quad (14)$$

where \tilde{L} is an arbitrary operator that is selected so that it minimizes the number of spatial derivative terms in Eq. (14).

Although a complete elimination of these terms is possible in the case of plane geometry, as is shown in the next section, it seems that a general solution does not exist for an arbitrary geometry. In the latter, the operator L will contain spatial derivatives and the experimental set-up will have to be modified to allow for measurements of derivatives of ψ in every surface layer of the body.

III. THE PLANE GEOMETRY CASE

In plane geometry the angular intensity $\psi(z, \mu, \theta)$ depends only on the spatial variable z , and on the angular variables μ and θ , where $\mu = \cos(\Omega \cdot \mathbf{e}_z)$ and θ is the azimuthal angle. Moreover, because of the symmetries of the transport operator it is possible to derive a set of independent transport equations for the Fourier components, $\psi^m(z, \mu)$, of the intensity with respect to the azimuthal angle θ . For instance, assuming that the external sources and the boundary conditions are even in θ , one can expand

$$\psi(z, \mu, \theta) = \sum_{m=0}^{\infty} (2 - \delta_{m0}) \cos m\theta \psi^m(z, \mu) \quad (15)$$

and obtain a transport equation for ψ^m by multiplying Eq. (1) by $\cos m\theta d\theta$ and then integrating over $(0, 2\pi)$. The equation is of the form of Eq. (1) but with ψ, S and ψ_b replaced by their Fourier components ψ^m, S^m and ψ_b^m . The corresponding transport operator is

$$B^m = \mu \partial_z + K^m, \quad (16)$$

with $K^m = \sigma - H^m$, and

$$H^m = \sum_{k>m} h_k Q_k^m, \quad (17)$$

where Q_k^m is the orthogonal projection

$$(Q_k^m \psi^m)(z, \mu) = \phi_k^m(\mu) \int_{-1}^1 \phi_k^m(\mu') \psi^m(z, \mu') d\mu'.$$

Here ϕ_k^m is the normalized associated Legendre function

$$\phi_k^m(\mu) = P_k^m(\mu) / N_k^m, \quad k \geq m, \quad (18)$$

and

$$(N_k^m)^2 = \frac{2}{2k+1} \frac{(k+m)!}{(k-m)!}.$$

To simplify our notation the superscript m will not be indicated in the following, except when necessary to avoid confusion.

In order to apply to the present case the ideas outlined in the preceding section, the scalar product of Eq. (7) is replaced by

$$(f, g) = \int_{D \times [-1, 1]} fg \, dz \, d\mu = \int_D f \cdot g \, dz, \quad (19)$$

where now $D = [z_-, z_+]$, and we have introduced the angular scalar product

$$f \cdot g = \int_{-1}^1 fg \, d\mu. \quad (20)$$

In particular, the surface integral (10) now reads as

$$\langle f, g \rangle = \sum_{\Gamma} [f(\mu g)],$$

where, for a continuous layer, $\Gamma = \{z_-, z_+\}$, so Eq. (13) becomes

$$\sum_{\Gamma} [L\psi \cdot (\mu\psi)] = 0. \quad (21)$$

We turn now to the solution of Eq. (14). In the case of a uniform slab it is possible to choose \tilde{L} so that the terms containing spatial derivatives cancel out. Indeed, since L does not depend on z , we can write

$-\Omega \cdot \nabla L + \tilde{L} \Omega \cdot \nabla = (-\mu L + \tilde{L} \mu) \partial_z$ that vanishes for $\tilde{L} = \mu L \mu^{-1}$. Replacing this value of \tilde{L} in Eq. (14) and multiplying on the left by μ^{-1} gives

$$(\mu^{-1} K) L + L (\mu^{-1} K) = 0, \quad (22)$$

which indicates that $\mu^{-1} K$ and L anticommute. The general solution of this equation is of the form

$$L \in \mathcal{A}(\mu^{-1} K), \quad (23)$$

where $\mathcal{A}(X)$ is the algebra of operators commuting with X , and R is a particular solution with a two-sided inverse R^{-1} such that $R^{-1} R = R R^{-1}$. If the medium is such that $\sigma \neq h_k$, K has a two-sided inverse; in any case, because of the diagonal structure of K , it is always possible to construct an inverse operator K^{-1} as any operator that acts as the inverse of K in Range(K), and is closed in Kernel(K).

An important part of $\mathcal{A}(\mu^{-1} K)$ is the abelian sub-algebra generated by polynomials of $\mu^{-1} K$ and its inverse $K^{-1} \mu$. In the following, we will restrict our attention to this particular class of linear solutions. Thus, we replace Eq. (23) by the more restrictive form

$$L = R \sum_{n=-\infty}^{\infty} \alpha_n (\sigma K^{-1} \mu)^n \quad (24)$$

for any convergent sum. Here the α_n 's are real numbers and the factor σ has been introduced for normalization purposes. As a particular solution we choose the reflection operator $R = R^{-1}$ defined by $(Rf)(\mu) = f(-\mu)$.

In principle, any operator of the form (24) will provide an inverse method when used in Eq. (13). This is true with the exception of those operators that satisfy Eq. (13) regardless of ψ . These spurious solutions verify the relation $\mu L + L^* \mu = 0$, where, from now on, the asterisk designates the adjoint with respect to the angular scalar product. Since R, K and μ are self-adjoint we obtain from Eq. (24),

$$\mu L + L^* \mu = \mu R \sum_{n=-\infty}^{\infty} \alpha_n (1 + (-)^n) (\sigma K^{-1} \mu)^n$$

so we can restrict the summation in Eq. (24) to only odd values of n .

At this point a few comments are appropriate. First, since the operators given by (24) are linear, it is sufficient to study only the operators L_l defined by $\alpha_n = \delta_{nl}$ and l odd:

$$L_l = R (\sigma K^{-1} \mu)^l. \quad (25)$$

On the other hand, since $\sigma^{-1} K = 1 - \sigma^{-1} H$ and $\sigma K^{-1} = 1 + H K^{-1}$, it is readily shown that the system of equations generated from (24) will be inhomogeneous. In particular, the source term originating from L_l is $\langle R \mu^l \psi, \psi \rangle$ and involves an angular integration with weight μ^{l+1} ; for negative powers of μ , this integral may be unbounded or may

produce inadmissible inaccuracies when using experimental data, so we will consider only the inverse methods generated for L_l 's with $l \geq -1$. Finally, observe that when the azimuthal number m is different from 0, it is necessary that the incident signal ψ_b be θ -dependent, otherwise the Fourier components ψ^m will be 0 for $m \neq 0$ and the corresponding equations will vanish identically.

The derivation from Eq. (21) of the inverse method corresponding to a given L_l is only a matter of straightforward algebraic manipulation, the details of which are indicated in Appendix A for $l = -1, 1$ and 3. Although the generation of inverse methods for higher values of l appears to be cumbersome, it requires only algebraic operations and, as shown in Appendix B, there is an iterative procedure that simplifies this derivation. To illustrate the advantage of the recursion relation the cases $l = 3$ and $l = 5$ are obtained in Appendix B.

The results are expressed in terms of the angular moments

$$f_k = \phi_k f = \int_{-1}^1 \phi_k(\mu') f(z, \mu') d\mu' \quad (26)$$

and the coefficients

$$\begin{aligned} d_k &= h_k / (\sigma - h_k) \\ \bar{d}_k &= 1 + d_k = \sigma / (\sigma - h_k) \\ \gamma_k &= \begin{cases} (k+1-m)N_{k+1} / ((2k+1)N_k), & k \geq m \\ 0 & k < m \end{cases} \quad (27) \end{aligned}$$

The inverse method generated by L_{-1} is defined by the system of equations,

$$\begin{aligned} \int_{-1}^1 \psi(z, \mu) \psi(z, -\mu) d\mu \Big|_{z_-}^{z_+} \\ = \sum_{k > m} (-)^{k-m} \frac{h_k}{\sigma} \psi_k^2 \Big|_{z_-}^{z_+}, \quad m = 0, 1, \dots, \end{aligned} \quad (28)$$

where, for a homogeneous medium

$$\int_F [f(z)] \equiv f(z) \Big|_{z_-}^{z_+} = \begin{cases} f(z_+) - f(z_-), & z_+ < \infty \\ -f(z_-), & z_+ = \infty. \end{cases}$$

For the inverse method generated by L_1 ,

$$\begin{aligned} \int_{-1}^1 \mu^2 \psi(z, \mu) \psi(z, -\mu) d\mu \Big|_{z_-}^{z_+} \\ = \sum_{k > m} (-)^{k-m} d_k (\mu \psi)_k^2 \Big|_{z_-}^{z_+}, \quad m = 0, 1, \dots \end{aligned} \quad (29)$$

Finally, from Appendix B the L_3 result is

$$\begin{aligned} \int_{-1}^1 \mu^4 \psi(z, \mu) \psi(z, -\mu) d\mu \Big|_{z_-}^{z_+} \\ = \sum_{k > m} (-)^{k-m} (\mu \psi)_k \{ 2(\bar{d}_k \bar{d}_{k+1} \bar{d}_{k+2} - 1) \gamma_k \gamma_{k+1} (\mu \psi)_{k+2} \\ + [\bar{d}_k^2 (\gamma_k^2 \bar{d}_{k+1} + \gamma_{k-1}^2 \bar{d}_{k-1}) - (\gamma_k^2 + \gamma_{k-1}^2)] \\ \times (\mu \psi)_k \} \Big|_{z_-}^{z_+}, \quad m = 0, 1, \dots \end{aligned} \quad (30)$$

Equations (28)–(30) are three independent sets of equations which can be used to calculate the coefficients h_k / σ for a finite or semiinfinite homogeneous slab. Because any real scatterer has a given order N of anisotropy beyond which the h_k 's, $k > N$, are negligible, in practice the system of equations

corresponding to any of these inverse methods is of finite order. Nevertheless, the order of anisotropy N constitutes another unknown, and the system of equations will have to be recursively solved, for increasing values of N , until the solution stabilizes.

Although the system of equations (29) is nonlinear in the cross sections, it is linear in the variables d_k . Due to the presence of off-diagonal moments, caused by the permutation between the operators μ and $H^{-1}K$, the system of equations (30) cannot be reduced to a linear form, and the same applies for any $l > 3$; the only linear inverse methods are those of Eqs. (28) and (29). Also, the systems of equations corresponding to these two methods have a triangular structure.

Inverse methods (28) and (29) have been previously obtained by McCormick.⁵ A particular form of inverse methods (28), (29), and (30), namely the one corresponding to $m = 0$ and $h_k = 0$ for $k > 2$, has been derived by Siewert.^{3,4} The recursion relation

$$(\mu^{n+1} \psi)_k = \gamma_k (\mu^n \psi)_{k+1} + \gamma_{k-1} (\mu^n \psi)_{k-1} \quad (31)$$

shows that there is not a unique way to write the moments in the inverse method equations. This helps explain the apparent differences between Eq. (30), Eq. (A12), and Siewert's result.⁴

We end this section by showing that all the results in plane geometry can be obtained without constructing the transport equations for the Fourier components of the intensity. Indeed, starting with the general transport equation for the total intensity, $\psi(z, \mu, \theta)$, we again obtain Eq. (22), but now the solutions of this equation will be operators acting on both μ and θ . In particular, assuming again that the angular intensity is an even function of θ , the orthogonal projection

$$M: \psi(z, \mu, \theta) \rightarrow \cos m \theta \psi^m(z, \mu) \quad (32)$$

commutes with the operator K , and it is easy to check that MR is a particular solution of Eq. (22). Thus the new solutions will be given by ML , with L from Eq. (24). Then the inverse method equation becomes

$$0 = \langle ML\psi, \psi \rangle = \langle ML\psi, M\psi \rangle = \langle L\psi, \psi \rangle_{\text{plane}},$$

where $\langle \cdot, \cdot \rangle_{\text{plane}}$ is the scalar product of Eq. (21).

IV. TIME-AND ENERGY-DEPENDENT CASES

A. Time-dependent case

Consider first the monochromatic time-dependent transport equation for a system without internal sources which begins to be irradiated at time $t = 0$,

$$\left. \begin{aligned} \left(\frac{1}{v} \partial_t + B \right) \psi &= 0, \text{ in } D \times (4\pi), & \text{for } t \geq 0 \\ \psi &= 0, \text{ in } D \times (4\pi), & \text{for } t \leq 0 \\ \psi &= \psi_b, \text{ on } D \times (2\pi)_-, & \text{for } t \geq 0 \end{aligned} \right\} \quad (33)$$

By Laplace transforming we obtain

$$\left. \begin{aligned} B_s \hat{\psi} &= 0, & \text{in } D \times (4\pi) \\ \hat{\psi} &= \hat{\psi}_b, & \text{on } D \times (2\pi)_- \end{aligned} \right\} \quad (34)$$

where

$$B_s = B + s/v = \Omega \cdot \nabla + \sigma_s - H \quad (35)$$

$$\sigma_s = \sigma + s/v \quad (36)$$

and

$$\hat{\psi}(\mathbf{r}, \Omega, s) = \int_0^\infty e^{-st} \psi(\mathbf{r}, \Omega, t) dt. \quad (37)$$

Notice that transport equation (34) for the Laplace transform of the intensity is of the form of Eq. (1) but with a modified cross section σ_s . Consequently, any inverse method derived for the stationary case will apply to this equation. The equations of the inverse method will depend on s through the modified cross section (36), and through the values of the Laplace transform of the intensity, Eq. (37). Because s is arbitrary, it is possible to generate different systems of equations from a single inverse method by simply changing the value of s , which gives flexibility for the treatment of the problem; selection of a large s enhances the importance of short-time measurements, whereas a small s enhances long-time measurements.

For the plane geometry case, both the unknowns and the matrix coefficients depend on s . In particular, the equations will involve the total cross section σ in a nonhomogeneous way because h_k/σ is now replaced by $h_k/(\sigma + s/v)$. Thus it will be necessary to add a supplementary equation to compute σ , which can be accomplished by taking any of the equations with a different value of s .

An interesting situation arises when the medium is irradiated with a periodic intensity starting early enough to ensure that the medium has reached its equilibrium by the time $t = 0$, i.e., that the intensity inside the medium oscillates with the same periodicity. For instance, assume a time behavior of the form $e^{-i\omega t}$; then the problem is similar to that of Eq. (33) but with an initial condition, $\psi = \psi_0(\mathbf{r}, \Omega)$, in $D \times (4\pi)$ at $t = 0$. By Laplace transforming we again obtain Eq. (34), but now

$$\sigma_s = \sigma - i\omega/v, \quad \hat{\psi}(\mathbf{r}, \Omega, s) = \psi_0/(s + i\omega). \quad (38)$$

In the present case ψ_0 is a complex function, and the inverse method equations will be complex. Notice that, since $\langle L\hat{\psi}, \hat{\psi} \rangle = \langle L\psi_0, \psi_0 \rangle (s + i\omega)^{-2}$, these equations will be independent of s . Nevertheless, the frequency ω gives us the same flexibility that we had with s in the previous case.

B. Energy-dependent case

The linear transport equation for energy-dependent interactions between neutrons and matter is still of the form of Eq. (1), but now sources, intensities and cross sections depend upon the energy variable E . The corresponding scattering operator is an integral operator in the variables Ω and E ,

$$(H\psi)(\mathbf{r}, \Omega, E) = \int_{(4\pi) \times (E)} \sigma_s(\mathbf{r}, \Omega', E' \rightarrow \Omega, E) \psi(\mathbf{r}, \Omega', E') d\Omega' dE', \quad (39)$$

where (E) designates the domain of integration in E . Again, for a rotationally invariant medium, this operator acts projection-wise in the angular variable.

$$H = \sum_{k>0} H_k Q_k, \quad (40)$$

except now H_k is an integral operator

$$(H_k f)(E) = \int_{(E)} h_k(\mathbf{r}, E' \rightarrow E) f(E') dE' \quad (41)$$

whose kernel describes the change of energy after collision for neutrons in the k th angular mode.

The technique outlined in Sec. II for the treatment of the inverse problem can be adapted to the present case by merely replacing the scalar product of Eq. (7) by

$$(f, g) = \int_{D \times (4\pi) \times (E)} fg M(E) d\mathbf{r} d\Omega dE, \quad (42)$$

where we have introduced a weight $M(E) > 0$. Then, with the appropriate modifications to account for the extra integration in the energy domain, Eqs. (8) and (10)–(14) apply to this case. The formal adjoint of the energy-dependent transport operator is now

$$B^* = -\Omega \cdot \nabla + K^* \quad (43)$$

because, in general, the operator K will not be self-adjoint.

Similarly, the treatment of the plane geometry case can be generalized to the energy-dependent case. Equation (22) is replaced by

$$\mu^{-1} K^* L + L \mu^{-1} K = 0, \quad (44)$$

which, with $T = \mu^{-1} K$ and $X = \mu L$, reduces to the more symmetric form

$$T^* X + XT = 0.$$

If X is a solution of this last equation, for any operator P commuting with T , then XP, P^*X and their adjoints are also solutions. Thus, given a particular solution of Eq. (44), we will be able to generate a set of solutions.

Because of the complicated structure of the operator H of Eq. (40), the problem of finding a solution to Eq. (44) is not a trivial one except, of course, when K is self-adjoint, which we have already considered. The operator K will be self-adjoint whenever the condition

$$M(E) \sigma_s(\mathbf{r}, \Omega', E' \rightarrow \Omega, E) = M(E') \sigma_s(\mathbf{r}, \Omega, E \rightarrow \Omega', E') \quad (45)$$

is satisfied for all $E, E' \in (E)$.

The operator K can be made self-adjoint if the scattering cross sections satisfy the detailed relation,⁹ which is Eq. (45) provided $M(E)$ is the Maxwellian for the equilibrium temperature T . If the signal ψ_b is zero for energies $E \gtrsim \kappa T$ (where κ is Boltzmann's constant), then we can constrain E to the thermal domain. Again Eq. (44) can be solved, with the technique previously developed, although some precautions have to be observed. For instance, $\sigma(E)$ now depends on the energy so it cannot be used as a normalization factor in Eq. (24); also the derivation of inverse methods involving K^{-1} will demand some approximation for the computation of $(\sigma - H_k)^{-1}$ in the energy domain.

Here we consider only the case $L_{-1} = R\mu^{-1}K$ for which, proceeding as in Appendix A, we get

$$\int_{(-1, 1) \times (E)} M(E) \sigma(E) \psi(z, \mu, E) \psi(z, -\mu, E) d\mu dE \Big|_z = \sum_{k>0} (-)^{k-m} \int_{(E)} M(E) \psi_k(E) dE$$

$$\times \int_{(E)} H_k(E' \rightarrow E) \psi_k(E') dE' \Big|_{z_-}^{z_+} \quad (46)$$

After defining appropriate spectrum-averaged cross sections, Eq. (46) can be written as

$$\int_{|-1,1| \times (E)} M(E) \psi(z, \mu, E) \psi(z, -\mu, E) d\mu dE \Big|_{z_-}^{z_+} \\ = \sum_{k \geq m} (-)^{k-m} \frac{\bar{h}_k}{\bar{\sigma}} \int_{(E)} M(E) \psi_k^2(E) dE \Big|_{z_-}^{z_+}, \quad m = 0, 1, \dots \quad (47)$$

which will yield the constants $\bar{h}_k/\bar{\sigma}$.

It is important to note that a correct definition of the spectrum-averaged cross sections, i.e., the one based on the spectrum created by the signal ψ_b , is possible only if the dependence of the flux on the energy can be factorized, or otherwise both $\bar{\sigma}$ and \bar{h}_k will depend on the azimuthal mode m and on the surface position, z_- or z_+ . In any case, for those averaged cross sections to be representative it is necessary not only that the factorization be a good approximation for the flux induced by ψ_b , but also that the spectrum so produced be close to the reference Maxwellian spectrum.

We consider now neutrons in the slowing-down energy range, for which there is no upscattering. We first consider the idealized case for which the scattering is elastic and both the total and the angle-integrated scattering cross sections are independent of energy. Then a Mellin transform in energy (which is equivalent to a Laplace transform in the lethargy variable, proportional to $-\ln E$) reduces the energy-dependent transport equation to a form nearly identical to the monoenergetic equation.^{10,11} Then applying the inverse methods of Sec. III one can obtain $\sigma^{-1} h_k(s)$, which are the transform of the coefficients $\sigma^{-1} h_k(E' \rightarrow E)$. If desired, the energy-dependent coefficients can be recovered by a numerical inverse transform.

The two cases just discussed were based on a continuous-energy description of the scattering. We turn now to the pragmatic case in which we introduce a multigroup representation by dividing the energy domain into a set of G disjoint groups, the g th group being defined by the range of energies $(E_g) = [E_g, E_{g+1}]$. In this approximation the flux is characterized by its averaged values on the groups, $\psi \sim \{\psi^g, g = 1, \dots, G\}$, and a spectral-averaged set of cross sections is used to describe group-to-group transfers. A natural basis in the energy domain is defined by the functions $\{\chi^g(E); g = 1, \dots, G\}$, where $\chi^g(E)$ is $(E_{g+1} - E_g)^{-1/2}$ times the characteristic function of (E_g) .

Consider now the multigroup formulation of the plane geometry transport equation for the m th Fourier component of the flux. Then, the set of functions of the form

$$\phi_k^g(\mu, E) = \phi_k(\mu) \chi^g(E), \quad k \geq 0, \quad g = 1, \dots, G, \quad (48)$$

with ϕ_k as defined in Eq. (18), constitutes a complete, orthonormal basis with respect to the scalar product

$$fg = \int_{|-1,1| \times (E)} fg d\mu dE. \quad (49)$$

In this basis

$$H = \sum_{k, g, g'} h_k^{gg'} Q_k^{gg'}, \quad (50)$$

where the summation is for $k \geq m$ and for $1 \leq g, g' \leq G$, and

$$(Q_k^{gg'} f)(\mu, E) = (\phi_k^g \cdot f) \phi_k^{g'}(\mu, E). \quad (51)$$

The constants $h_k^{gg'}$, related to averaged cross sections, characterize the transfer by scattering from group g' to group g for the angular mode k .

Let g be a group in the slowing-down domain, and suppose that the incident beam ψ_b is zero for energies $E > E_{g+1}$, and nonzero for $E \in (E_g)$. Then, $\psi^g = 0$ for $g' > g$ and the neutrons in group g behave as if they were monoenergetic so the inverse methods obtained in Sec. III apply to the present case. Varying the upper energy of the incident beam makes it possible to determine the in-group scattering coefficients h_k^{gg}/σ^g for all the groups in the slowing-down region.

This method, and that of Eq. (47), do not provide information about the scattering transfer cross sections between different energies. If such information is needed, one can utilize the inverse method defined in Eq. (6), but the penalty is that invasive measurements are required. In plane azimuthally-symmetric geometry, however, it is possible to obtain an inverse method that requires only the scalar flux inside the medium and not detailed angular measurements;¹ such a method has practical experimental advantages over one requiring angular data.

Proceeding as for Eq. (6), we apply ϕ_k^g to transport equation (1) and integrate the result in $D = [z_-, z_+]$ to obtain the multigroup version of Eq. (6) for the azimuthally-independent ($m = 0$) plane geometry case,

$$\sum_{g'} (\sigma^g \delta_g^{g'} - h_k^{gg'}) \int_D \psi_k^g(z) dz = -\phi_k^g \cdot (\mu \psi) \Big|_{z_-}^{z_+}, \\ k \geq 0, \quad 1 \leq g \leq G. \quad (52)$$

Here σ^g is the averaged total cross section in group g , $\delta_g^{g'}$ is the Kronecker delta, and $\psi_k^g = \phi_k^g \cdot \psi$.

With G linearly independent signals (such as $\psi_b = \chi^g(E) f(\mu)$ for some function f and $g = 1, \dots, G$), one can solve Eq. (52) for any angular mode k to obtain the components of the k th collision matrix $K_k^{gg'} = \sigma^g \delta_g^{g'} - h_k^{gg'}$. Besides the surface term on the RHS of Eq. (52), this calculation will require the volume contributions

$$\bar{\psi}_k^g = \int_D \psi_k^g(z) dz$$

that have to be determined by measuring the angular flux inside the medium. But, as shown in Appendix C, the coefficients $\bar{\psi}_k^g$ can be recursively calculated from surface contributions and from spatial moments of the scalar flux ψ_0^g .

Therefore, the inverse method of Eq. (52) will only require measurements of the surface angular flux, and measurements of the scalar flux inside the medium. In general, $\bar{\psi}_k^g$ depends upon the collision matrices K_l of order $l < k$, so inverse method (52) has to be solved for increasing values of k .

The practical validity of the calculated average cross section determined by such an inverse method will depend on the ability to obtain approximate factorizable spectra under experimental conditions.

V. CONCLUSIONS

A new derivation is presented for obtaining inverse methods for the linear transport equation. For homogeneous plane geometry a set of noninvasive inverse methods is obtained which encompasses previous ones.³⁻⁵ Time-dependent and neutron energy-dependent inverse problems also have been analyzed.

Although it is possible with plane geometry to keep developing independent sets of noninvasive method equations, they will become increasingly nonlinear. Only the two inverse methods corresponding to Eqs. (28) and (29) are linear, for which good numerical accuracy has been obtained⁶; these two methods require azimuthally-dependent measurements. Another alternative for developing an inverse method is to select equations corresponding to more than one L_l ; in particular, selecting only the $m = 0$ equations for $l = -1, 1, 3$, etc. yields an inverse method which depends only on azimuthally-symmetric measurements. Such a scheme will have the inherent numerical difficulties associated with nonlinear problems whenever the scattering is at least quadratically anisotropic.¹²

The use of time-dependent incident distributions makes possible any number of different calculations with a given noninvasive inverse method, since the variable s or ω can be selected arbitrarily. It should be noted, however, that measurements are still required for all times, but the value of s , for example, can be chosen to match the accuracy of the experimental information.

Noninvasive measurements of the surface distribution for neutrons in the thermal energy region will provide the Maxwellian-averaged energy-transfer cross sections. With a multigroup formulation, in-group cross sections can be obtained by noninvasive measurements using a set of different sources, one for each energy group. Measurements of the scalar flux inside the medium are needed if multigroup transfer cross sections are desired. (See note added in proof.)

In Sec. III only linear solutions for the operator L of Eq. (22) have been considered. The simplest nonlinear solution, $L = RT$, corresponding to a commutator of the form $T\psi = \psi_0$, with ψ_0 a constant, does not yield any inverse method. Possibly other nonlinear solutions should be investigated.

The question of the uniqueness of the inverse problem has not been addressed here. Nevertheless, numerical results obtained using Eqs. (28) and (29) seem to substantiate the uniqueness of the inverse solution for the homogeneous slab.⁶

Note added in proof: E. W. Larsen [J. Math. Phys. **22**, 158 (1981)] has constructed in plane geometry a noninvasive scheme for measuring multigroup transfer cross sections. His formulation utilizes a solution ψ^* to the adjoint transport equation, and requires that a supplementary set of noninvasive experiments be performed to determine the values of ψ^* on the slab surfaces. In our notation, Larsen's approach corresponds to selecting $L\psi = \psi^*$ as a solution of Eq. (12), in which case the operator L must contain information not available from the measurement of ψ . This suggests that our approach in Eqs. (8)–(13) could be generalized by including information from two or more experiments. For in-

stance, if we replace Eq. (11) by $f = L\psi_1$, $g = \psi_2$, then any L which depends upon the properties of the medium and satisfies $(B * L\psi_1, \psi_2) = 0$ would lead to an inverse method of the form $(L\psi_1, \psi_2) = 0$.

ACKNOWLEDGMENT

This work was done as a part of National Science Foundation Industry/University Cooperative Research Activity Grant ENG-7908377 to Mathematical Sciences Northwest, Inc.

APPENDIX A: DERIVATION OF THE INVERSE METHOD EQS. (28)–(30)

For the operator L_l of Eq. (25), the generic form of the inverse method equation (21) is

$$\mu\psi \cdot L_l \psi = \mu\psi \cdot R(\sigma K^{-1}\mu)\psi = 0, \quad (\text{A1})$$

where, for the sake of simplicity, the symbols Σ_T and $[\]$ have been omitted. We will perform our calculations using as a basis the functions ϕ_k defined in Eq. (18), which are orthonormal with respect to the scalar product (20). Thus,

$$f \cdot g = \sum_k f_k g_k, \quad (\text{A2})$$

where, from now on, the summation is for the integers $\geq m$, and f_k is the k th component of f , as defined in Eq. (26). In this basis, for any linear operator T , we can write $T = \sum_{kl} T_{kl} \phi_k \phi_l$.

with the components $T_{kl} = \phi_k \cdot T \phi_l$.

In particular,

$$(Tf)_k = \sum_l T_{kl} f_l. \quad (\text{A3})$$

The operators H , R , and $D = HK^{-1}$ are diagonal with components

$$H_{kk} = h_k, \quad R_{kk} = (-)^{k-m}, \quad D_{kk} = d_k,$$

where d_k is defined in Eq. (27).

Consider now the first operator $L_{-1} = R\mu^{-1}\sigma^{-1}K$.

Since R and μ anticommute and are self-adjoint,

$$\mu\psi \cdot L_{-1} \psi = -R\psi \cdot \sigma^{-1}K\psi = R\psi \cdot (\sigma^{-1}H - 1)\psi.$$

Using Eqs. (A2) and (A3) to express $R\psi \cdot \sigma^{-1}H\psi$ in components, the corresponding Eq. (A1) for $L = L_{-1}$ is

$$\psi \cdot R\psi = \sum_k (-)^{k-m} \frac{h_k}{\sigma} \psi_k^2. \quad (\text{A4})$$

Similarly, for $L_1 = R\sigma K^{-1}\mu$, we have

$\mu\psi \cdot L_1 \psi = R\mu\psi \cdot \sigma K^{-1}\mu\psi = R\mu\psi \cdot (1 + D)\mu\psi$ and, expressing $R\mu\psi \cdot D\mu\psi$ in components, we obtain

$$\mu^2\psi \cdot R\psi = \sum_k (-)^{k-m} d_k (\mu\psi)_k^2. \quad (\text{A5})$$

Finally, consider the case

$$\begin{aligned} L_3 &= R(\sigma K^{-1}\mu)^3 = R(\mu + D\mu)^3 \\ &= R[\mu^3 + \mu^2 D\mu + \mu D\mu^2 + D\mu^3 + \mu(D\mu)^2 \\ &\quad + D\mu^2 D\mu + (D\mu)^2 \mu + (D\mu)^3], \end{aligned} \quad (\text{A6})$$

for which we have to compute the contribution of every term to Eq. (A1). In order to reduce the number of computations, we observe that

$$\mu R A = B * R \mu \Rightarrow L A \psi \cdot \mu \psi = L B \psi \cdot \mu \psi \quad (\text{A7})$$

for any two operators A and B . Consequently, the fourth operator in Eq. (A6), $D\mu^3$, and the seventh, $(D\mu)^2 \mu$, will give the same contribution as the second, $\mu^2 D\mu$, and the fifth,

$\mu(D\mu)^2$, respectively. The first three terms of L_3 are readily dealt with:

$$\begin{aligned} \mu\psi \cdot R\mu^3\psi &= -\mu^4\psi \cdot R\psi, \\ \mu\psi \cdot R\mu^2 D\mu\psi &= R\mu^3\psi \cdot D\mu\psi = \sum_k (-)^{k-m} d_k (\mu\psi)_k (\mu^3\psi)_k, \\ \mu\psi \cdot R\mu D\mu^2\psi &= -R\mu^2\psi \cdot D\mu^2\psi \\ &= -\sum_k (-)^{k-m} d_k (\mu^2\psi)_k^2. \end{aligned} \quad (\text{A8})$$

For the remaining terms we need to explicitly calculate the operators μ , μ^2 , and $\mu D\mu$. Using the recursion relation for the associated Legendre functions, one obtains

$$\mu = \sum_k (\gamma_k \phi_k \phi_{k+1} + \gamma_{k-1} \phi_k \phi_{k-1}), \quad (\text{A9})$$

where the γ_k are given in Eq. (27). Then,

$$\begin{aligned} \mu^2 &= \sum_k [\gamma_k \gamma_{k+1} \phi_k \phi_{k+2} + (\gamma_k^2 + \gamma_{k-1}^2) \phi_k \phi_k + \gamma_{k-1} \gamma_{k-2} \phi_k \phi_{k-2}], \\ \mu D\mu &= \sum_k [\gamma_k \gamma_{k+1} d_{k+1} \phi_k \phi_{k+2} + (\gamma_k^2 d_{k+1} + \gamma_{k-1}^2 d_{k-1}) \phi_k \phi_k + \gamma_{k-1} \gamma_{k-2} d_{k-1} \phi_k \phi_{k-2}]. \end{aligned} \quad (\text{A10})$$

With the help of these formulas we obtain

$$\begin{aligned} \mu\psi \cdot R\mu(D\mu)^2\psi &= -DR\mu^2\psi \cdot \mu D\mu\psi = \sum_k (-)^{k-m} d_k (\mu^2\psi)_k [\gamma_k d_{k+1} (\mu\psi)_{k+1} + \gamma_{k-1} d_{k-1} (\mu\psi)_{k-1}], \\ \mu\psi \cdot R D\mu^2 D\mu\psi &= DR\mu\psi \cdot \mu^2 D\mu\psi = \sum_k (-)^{k-m} d_k (\mu\psi)_k [\gamma_k \gamma_{k+1} d_{k+2} (\mu\psi)_{k+2} \\ &\quad + (\gamma_k^2 + \gamma_{k-1}^2) d_k (\mu\psi)_k + \gamma_{k-1} \gamma_{k-2} d_{k-2} (\mu\psi)_{k-2}], \\ \mu\psi \cdot R(D\mu)^3\psi &= DR\mu\psi \cdot \mu D\mu D\mu\psi = \sum_k (-)^{k-m} d_k (\mu\psi)_k [\gamma_k \gamma_{k+1} d_{k+1} d_{k+2} (\mu\psi)_{k+2} \\ &\quad + (\gamma_k^2 d_{k+1} + \gamma_{k-1}^2 d_{k-1}) d_k (\mu\psi)_k + \gamma_{k-1} \gamma_{k-2} d_{k-1} d_{k-2} (\mu\psi)_{k-2}]. \end{aligned} \quad (\text{A11})$$

Thus, the inverse method equation from Eq. (A1) is, for $L = L_3$,

$$\begin{aligned} \mu^4\psi \cdot R\psi &= \sum_k (-)^{k-m} d_k \{(\mu\psi)_k [2(\mu^3\psi)_k + 2\gamma_k \gamma_{k+1} \tilde{d}_{k+1} d_{k+2} (\mu\psi)_{k+2} + (\gamma_k^2 \tilde{d}_{k+1} + \gamma_{k-1}^2 \tilde{d}_{k-1}) d_k (\mu\psi)_k] \\ &\quad - (\mu^2\psi)_k [2\gamma_k d_{k+1} (\mu\psi)_{k+1} + (\mu^2\psi)_k + 2\gamma_{k-1} d_{k-1} (\mu\psi)_{k-1}]\}, \end{aligned} \quad (\text{A12})$$

where $\tilde{d}_k = 1 + d_k$. Equation (A12) has been simplified by using the identity

$$\sum_k \gamma_{k-l} f_l = \sum_k \gamma_k f_{k+l}, \quad l \geq 0, \quad (\text{A13})$$

to transform the term containing γ_{k-2} .

APPENDIX B: A RECURSION RELATION FOR THE DERIVATION OF THE INVERSE METHOD EQUATION

Write Eq. (A1) in components,

$$\mu\psi \cdot L_{l+2}\psi = \sum_k (\mu\psi)_k (L_{l+2}\psi)_k = 0, \quad (\text{B1})$$

and observe that $L_{l+2} = (\sigma K^{-1}\mu)^2 L_l$. Then, with the help of Eq. (A3), we obtain a recursion relation for the $(L_{l+2}\psi)_k$:

$$(L_{l+2}\psi)_k = \sum_j ((\sigma K^{-1}\mu)^2)_{kj} (L_l\psi)_j. \quad (\text{B2})$$

To calculate the operator $(\sigma K^{-1}\mu)^2 = (1 + D)(\mu^2 + \mu D\mu)$ in Eq. (B2), we use Eq. (A10) to obtain

$$\begin{aligned} (\sigma K^{-1}\mu)^2 &= \sum_k \tilde{d}_k [\gamma_k \gamma_{k+1} \tilde{d}_{k+1} \phi_k \phi_{k+2} \\ &\quad + (\gamma_k^2 \tilde{d}_{k+1} + \gamma_{k-1}^2 \tilde{d}_{k-1}) \phi_k \phi_k \\ &\quad + \gamma_{k-1} \gamma_{k-2} \tilde{d}_{k-1} \phi_k \phi_{k-2}]. \end{aligned} \quad (\text{B3})$$

Let us first illustrate the use of the recursion relation (B2) by again deriving the inverse method for $l = 3$. We begin

by calculating the components for $l = 1$ from Eq. (A5),

$$(L_1\psi)_k = (-)^{k-m} (1 + d_k) (\mu\psi)_k = (-)^{k-m} \tilde{d}_k (\mu\psi)_k, \quad (\text{B4})$$

where the inhomogeneous part $(-)^{k-m} (\mu\psi)_k$ is the k th component of $-\mu^2\psi \cdot R\psi$. After use of Eqs. (B2)–(B4), we get

$$\begin{aligned} (L_3\psi)_k &= (-)^{k-m} \tilde{d}_k [\gamma_k \gamma_{k+1} \tilde{d}_{k+1} \tilde{d}_{k+2} (\mu\psi)_{k+2} \\ &\quad + \tilde{d}_k (\gamma_k^2 \tilde{d}_{k+1} + \gamma_{k-1}^2 \tilde{d}_{k-1}) (\mu\psi)_k \\ &\quad + \gamma_{k-1} \gamma_{k-2} \tilde{d}_{k-1} \tilde{d}_{k-2} (\mu\psi)_{k-2}]. \end{aligned} \quad (\text{B5})$$

Use of Eq. (B5) in Eq. (B1) will give the inverse method for L_3 . The new inhomogeneous term comes from the action of μ^2 on the components of the inhomogeneous term of $L_1\psi$. Since this new term can be identified as $-\mu^4\psi \cdot R\psi$, we obtain

$$\begin{aligned} \mu^4\psi \cdot R\psi &= \sum_k (-)^{k-m} (\mu\psi)_k \{2(\tilde{d}_k \tilde{d}_{k+1} \tilde{d}_{k+2} - 1) \gamma_k \\ &\quad \times \gamma_{k+1} (\mu\psi)_{k+2} + [\tilde{d}_k^2 (\gamma_k^2 \tilde{d}_{k+1} + \gamma_{k-1}^2 \tilde{d}_{k-1}) \\ &\quad - (\gamma_k^2 + \gamma_{k-1}^2)] (\mu\psi)_k\}, \end{aligned} \quad (\text{B6})$$

where we have used the identity (A13) to transform the term containing γ_{k-2} .

Similarly, using Eq. (B5), the L_5 result is found to be

$$\begin{aligned} \mu^6\psi \cdot R\psi &= \sum_k (-)^{k-m} (\mu\psi)_k [C_4 (\mu\psi)_{k+4} \\ &\quad + C_2 (\mu\psi)_{k+2} + C_0 (\mu\psi)_k], \end{aligned} \quad (\text{B7})$$

where

$$\begin{aligned}
 C_4 &= 2 \left(\prod_{j=0}^3 \gamma_{k+j} \right) \left(\prod_{j=0}^4 \bar{d}_{k+j} - 1 \right), \\
 C_2 &= 2\gamma_k \gamma_{k+1} \left\{ \prod_{j=0}^2 \bar{d}_{k+j} [\bar{d}_{k+2} (\gamma_{k+2}^2 \bar{d}_{k+3} + \gamma_{k+1}^2 \bar{d}_{k+1}) \right. \\
 &\quad \left. + \bar{d}_k (\gamma_k^2 \bar{d}_{k+1} + \gamma_{k-1}^2 \bar{d}_{k-1}) \right. \\
 &\quad \left. - (\gamma_{k+2}^2 + \gamma_{k+1}^2 + \gamma_k^2 + \gamma_{k-1}^2) \right\}, \\
 C_0 &= \bar{d}_k^2 [\gamma_k^2 \gamma_{k+1}^2 \bar{d}_{k+1}^2 \bar{d}_{k+2} \\
 &\quad + \bar{d}_k (\gamma_k^2 \bar{d}_{k+1} + \gamma_{k-1}^2 \bar{d}_{k-1})^2 \\
 &\quad + \gamma_{k-1}^2 \gamma_{k-2}^2 \bar{d}_{k-1}^2 \bar{d}_{k-2}] \\
 &\quad - [\gamma_k^2 \gamma_{k+1}^2 + (\gamma_k^2 + \gamma_{k-1}^2)^2 + \gamma_{k-1}^2 \gamma_{k-2}^2]. \quad (B8)
 \end{aligned}$$

APPENDIX C: ITERATIVE CALCULATION OF THE SPATIAL MOMENTS $\bar{\psi}_k^g$

Because the procedure to be developed is independent of the energy variable, we will suppress the group index and consider the moments

$$\bar{\psi}_k = \bar{D}\psi_k = \int_{z_-}^{z_+} \psi_k(z) dz, \quad (C1)$$

where we have implicitly defined the operator \bar{D} and have used angular scalar product of Eq. (20), $\psi_k = \phi_k \cdot \psi$.

Applying $\phi_k \cdot$ to the transport equation $(\mu \partial_z + K)\psi = 0$ we obtain, after having used recursion relation (31) for $n = 0$,

$$\partial_z (\gamma_k \psi_{k+1} + \gamma_{k-1} \psi_{k-1}) + K_k \psi_k = 0. \quad (C2)$$

Now define the operator D , an inverse of ∂_z , by

$$Df = \int_z^{z_+} f(z') dz' \Rightarrow D\partial_z = \delta_+ - 1, \quad (C3)$$

where

$$\delta_+ f = f(z_+). \quad (C4)$$

Then, applying D to Eq. (C2) we obtain the recursion relation

$$\psi_{k+1} = \delta_+ \psi_{k+1} + \gamma_k^{-1} [\gamma_{k-1} (\delta_+ - 1) \psi_{k-1} + DK_k \psi_k] \quad (C5)$$

and, by integrating in $[z_-, z_+]$:

$$\bar{\psi}_{k+1} = \beta_{k+1} + \gamma_k^{-1} K_k \bar{D}\bar{D}\psi_k, \quad (C6)$$

where β is defined as

$$\beta_{k+1} = a\psi_{k+1}(z_+) + \gamma_k^{-1} \gamma_{k-1} [a\psi_{k-1}(z_+) - \bar{\psi}_{k-1}]$$

and $a = z_+ - z_-$.

Although Eq. (C6) is not a closed recursion relation for the $\bar{\psi}_k$'s, it can be used in conjunction with Eq. (C5) to obtain, in an iterative manner, all the moments $\bar{\psi}_k$ for $k > 0$. The results will be expressed in terms of the surface contributions and the spatial moments of ψ_0 given by

$$\Psi_n = \frac{1}{n!} \int_z^{z_+} z^n \psi_0(z) dz. \quad (C7)$$

In practice, when using formulas (C5) and (C6) it is useful to take advantage of the fact that δ_+ and D commute with K_k , and that

$$\delta_+^2 = \delta_+, \quad \delta_+ D = 0, \quad \bar{D}D^n = D \frac{(z - z_-)^n}{n!}. \quad (C8)$$

As an example, we give the formula for the first five moments,

$$\begin{aligned}
 \bar{\psi}_0 &= \Psi_0, \\
 \bar{\psi}_1 &= \xi_1, \\
 \bar{\psi}_2 &= \beta_2 + \gamma_1^{-1} K_1 \xi_2, \\
 \bar{\psi}_3 &= \beta_3 + \gamma_2^{-1} K_2 \xi_2, \\
 \bar{\psi}_4 &= \beta_4 + \gamma_3^{-1} K_3 (a_2 \eta_2 - \gamma_2^{-1} \gamma_1 \xi_2 + \gamma_2^{-1} K_2 \xi_3), \\
 \bar{\psi}_5 &= \beta_5 + \gamma_4^{-1} K_4 [a_2 \eta_3 - \gamma_3^{-1} \gamma_2 \xi_2 + \gamma_3^{-1} K_3 \\
 &\quad \times (a_3 \eta_2 - \gamma_2^{-1} \gamma_1 \xi_3 + \gamma_2^{-1} K_2 \xi_4)], \quad (C9)
 \end{aligned}$$

where $a_n = a^n/n!$ and

$$\begin{aligned}
 \xi_n &= a_n \psi_1(z_+) + \gamma_0^{-1} K_0 \Psi_n, \\
 \eta_n &= \psi_{n+1}(z_+) + \gamma_n^{-1} \gamma_{n-1} \psi_{n-1}(z_+), \\
 \zeta_n &= a_n \eta_1 - \gamma_1^{-1} \gamma_0 \Psi_{n-1} + \gamma_1^{-1} K_1 \xi_{n+1}. \quad (C10)
 \end{aligned}$$

To apply these formulas to the multigroup case it is only necessary to replace ψ_k by the vector of components ψ_k^g , $g = 1, \dots, G$, Ψ_n by the vector of components Ψ_n^g , and K_k by the collision matrix of elements $K_k^{gg'}$. In such a case, formulas (C9) are equivalent to the multigroup results of Siewert.¹

Finally, observe that the formalism defined by Eq. (C5) and (C6) also can be applied to the m th Fourier form of the transport equation, with appropriate modifications, but in that case, the moment ψ_0 requires angular measurements.⁸

¹C. E. Siewert, Nucl. Sci. Eng. 67, 259 (1978).

²M. Kanal and J. A. Davies, Transp. Theory Stat. Phys. 8, 99 (1979).

³C. E. Siewert, Z. Angew. Math. Phys. (ZAMP) 30, 522 (1979).

⁴C. E. Siewert, J. Quant. Spectrosc. Radia. Transfer 22, 441 (1979).

⁵N. J. McCormick, J. Math. Phys. 20, 1504 (1979).

⁶N. J. McCormick and R. Sanchez, J. Math. Phys. Jan. 1981.

⁷N. J. McCormick and I. Kuščer, J. Math. Phys. 15, 926 (1974).

⁸N. J. McCormick and J. A. R. Veeder, J. Math. Phys. 19, 994 (1978); 20, 216 (1979).

⁹J. J. Duderstadt and W. R. Martin, Transport Theory (Wiley, New York, 1979).

¹⁰J. J. McInerney, Nucl. Sci. Eng. 22, 215 (1965).

¹¹D. G. Cacuci and H. Goldstein, J. Math. Phys. 18, 2436 (1977).

¹²W. L. Dunn and J. R. Maiorino, J. Quant. Spectrosc. Radia. Transfer 24, 203 (1980).

Linear transport in an exponential atmosphere ^{a)}

Edward W. Larsen

Theoretical Division, University of California, Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545

Thomas W. Mullikin

Department of Mathematics, Purdue University, West Lafayette, Indiana 47907

(Received 8 October 1980; accepted for publication 14 November 1980)

Two basic theoretical problems concerning linear transport in a subcritical half-space with an exponential scattering ratio are solved. First, the continuum eigenvolutions developed by Mullikin and Siewert are shown to be half-range complete. Second, the singular integral equation developed by Martin for the angular flux exiting the half-space is shown to possess a unique solution.

PACS numbers: 05.60. + w, 02.30. + g, 42.68.Db

I. INTRODUCTION

In this article we shall analyze some theoretical aspects of the following linear transport problem with an exponential scattering ratio $c(x) = c \exp(-x/s)$:

$$\mu \frac{\partial}{\partial x} \psi(x, \mu) + \psi(x, \mu) = \frac{c}{2} e^{-x/s} \int_{-1}^1 \psi(x, \mu') d\mu',$$

$$0 < x < \infty, \quad -1 \leq \mu \leq 1, \quad c > 0, \quad s > 0; \quad (1.1)$$

$$\psi(0, \mu) = f(\mu), \quad 0 \leq \mu \leq 1,$$

f is prescribed and Hölder-continuous; (1.2)

$$|\psi(x, \mu)| \leq M, \quad 0 \leq x < \infty, \quad -1 \leq \mu \leq 1. \quad (1.3)$$

This problem was originally proposed by Chamberlain and McElroy.¹ Later it was studied by Martin,² who derived a singular integral equation which the flux exiting the half-space, $\psi(0, -\mu)$ for $0 < \mu \leq 1$, must satisfy. Martin also showed that if the inequality

$$\frac{c}{2} \left[\frac{s}{s+1} \ln(1+2s) + \frac{\pi s}{s+1} + \left(\frac{s}{s+1} \right)^{1/2} \right] < 1 \quad (1.4)$$

is satisfied, then this singular integral equation uniquely determines the exiting flux.

The above inequality has been tested numerically, and has been shown to be conservative.^{3,4} That is, excellent numerical solutions of Martin's singular integral equation have been obtained for values of c and s which violate the inequality. Numerical difficulties have arisen, however, if the inequality is *sufficiently* violated. For example, if $c = 0.99$, Martin's inequality predicts unique solvability for $s < 0.570$, while numerical calculations using Siewert's collocation (i.e., " F_N ") method⁵ give excellent results for $s < 20$.^{3,4} For $s > 20$, Siewert's original method breaks down,⁴ although some recent modifications have apparently ameliorated this difficulty.⁶ Other numerical aspects of problem (1.1)–(1.3) have been considered by Mullikin.⁷

^{a)}Work by the first author (E. W. L.) was performed under the auspices of the U. S. Dept. of Energy. The second author (T. W. M.) expresses appreciation for the hospitality of the Sandia Corporation (Applied Mathematics Division 2646) where some of this work was performed.

Mullikin and Siewert³ have recently derived Martin's singular integral equation in a new way, by first constructing a set of continuum eigensolutions of the transport equation and then manipulating these eigensolutions. We shall sketch their derivation here.

The continuum eigensolutions are

$$\psi_\nu(x, \mu) = f_\nu(\mu) e^{-x/\nu} + g_\omega(\mu) e^{-x/\omega}, \quad 0 < \nu < 1, \quad (1.5)$$

where

$$f_\nu(\mu) = \delta(\nu - \mu), \quad (1.6)$$

$$g_\omega(\mu) = \frac{c\omega}{2} \left[\text{P.V.} \left(\frac{1}{\omega - \mu} \right) - \delta(\omega - \mu) \ln \left(\frac{1 + \omega}{1 - \omega} \right) \right], \quad (1.7)$$

(P.V. means principal value) and

$$\frac{1}{\omega} = \frac{1}{\nu} + \frac{1}{s}. \quad (1.8)$$

These solutions satisfy "full-range orthogonality," i.e.,

$$\int_{-1}^1 \mu \psi_\nu(0, -\mu) \psi_\nu(0, \mu) d\mu = 0, \quad 0 < \nu', \nu < 1. \quad (1.9)$$

To proceed, one *assumes* that the solution ψ of problem (1.1)–(1.3) can be written

$$\psi(x, \mu) = \int_0^1 a(\nu) \psi_\nu(x, \mu) d\nu. \quad (1.10)$$

This expansion is valid provided the boundary condition (1.2) is satisfied:

$$f(\mu) = \int_0^1 a(\nu) \psi_\nu(0, \mu) d\nu, \quad 0 < \mu \leq 1. \quad (1.11)$$

Eq. (1.11) can be rewritten as

$$f(\mu) = (I + \frac{1}{2}cL)a(\mu), \quad 0 < \mu \leq 1, \quad (1.12a)$$

where I is the identity operator, and L is defined by

$$(La)(\mu) = \text{P.V.} \int_0^{s/(s+1)} \omega \left(\frac{s}{s-\omega} \right)^2 a \left(\frac{\omega s}{s-\omega} \right) \frac{d\omega}{\omega - \mu} - \mu \left(\frac{s}{s-\mu} \right)^2 a \left(\frac{\mu s}{s-\mu} \right) \ln \left(\frac{1+\mu}{1-\mu} \right), \quad (1.12b)$$

where

$$a(\mu) \equiv 0 \quad \text{for } \mu > 1 \quad \text{and } \mu < 0. \quad (1.12c)$$

Formally, L can also be written as

$$(La)(\mu) = \int_0^{s/(s+1)} \omega \left(\frac{s}{s-\omega} \right)^2 a \left(\frac{\omega s}{s-\omega} \right) \times \left[\frac{\text{P.V.}}{\omega - \mu} - \delta(\omega - \mu) \ln \left(\frac{1+\omega}{1-\omega} \right) \right] d\omega. \quad (1.13)$$

Next one sets $x = 0$ in Eq. (1.10), operates by

$$\int_{-1}^1 \mu \psi_{\nu}(0, -\mu) \psi(0, -\mu) d\mu, \quad \text{uses Eq. (1.9), and rearranges to get}$$

$$\int_0^1 \mu \psi_{\nu}(0, \mu) \psi(0, -\mu) d\mu$$

$$= \int_0^1 \mu \psi_{\nu}(0, -\mu) f(\mu) d\mu, \quad 0 < \nu < 1, \quad (1.14)$$

which is Martin's singular integral equation. In this derivation one must assume that Eq. (1.12) holds, i.e., that the eigensolutions $\psi_{\nu}(0, \mu)$ are half-range complete. Larsen and Pomraning⁴ have recently proved this completeness if c and s satisfy Martin's inequality, Eq. (1.4).⁸

To summarize, it has been shown that if c and s satisfy the inequality (1.4), then:

- (a) The continuum eigensolutions are half-range complete (i.e., Eqs. (1.12) have a unique solution in $L_2[0, 1]$), and
- (b) Eq. (1.14) uniquely determines in $L_2[0, 1]$ the exiting flux.

Our goal in this article is to prove that if c and s are any positive constants such that the half-space $x > 0$ is subcritical, then the results (a) and (b) hold. [Unique solvability of Eq. (1.14) has recently been shown⁷ in a different Hilbert space.] In essence, we replace Martin's inequality by the much weaker condition that the half-space $x > 0$ be subcritical. [However, our proof requires the incident flux f to be Hölder-continuous, whereas the analyses in [3] and [4] only require $f \in L_2(0, 1)$. By a more technical analysis our proof should extend to $L_2(0, 1)$.] Interestingly, the expansion coefficients $a(\nu)$ which make the continuum eigensolutions half-range complete lie in $L_2(0, 1)$ but are generally not continuous on $(0, 1)$, no matter how smooth is the incident flux. In fact, $a(\nu)$ has a logarithmic singularity at each point $\nu_n = (1 + n/s)^{-1}$, $0 \leq n \leq \infty$, and is Hölder-continuous on every closed interval lying between any two consecutive such singular points. Thus the principal-value integral in Eq. (1.12b) must be interpreted in an L_2 sense. We shall discuss this in detail in Sec. II.

A summary of the remainder of this article follows. In Sec. II we establish half-range completeness of the continuum eigensolutions by proving that Eqs. (1.12) have a unique solution. The analysis in this section, which is based on the Laplace transform of problem (1.1)–(1.3), makes substantial use of analytic continuation arguments. In Sec. III we prove that Eq. (1.14) uniquely determines the exiting flux, and we conclude with a discussion in Sec. IV.

II. HALF-RANGE COMPLETENESS

The main purpose of this section is to prove that the continuum eigensolutions developed by Mullikin and

Siewert are half-range complete. From the discussion following Eq. (1.10) above, this result follows from:

Theorem 1: If Eqs. (1.1)–(1.3) have a unique solution, Eqs. (1.12) have a unique solution $a(\nu) \in L_2(0, 1)$. The solution $a(\nu)$ has a logarithmic singularity at each point

$$\nu_n = \frac{1}{1 + n/s}, \quad n = 0, 1, 2, \dots$$

and is Hölder-continuous on every closed subinterval which lies between any two consecutive such singular points.

The proof is contained in six subsections. In subsections (A) and (B), we derive an equation for $\hat{\phi}(z)$, the Laplace transform of the scalar flux $\phi(x)$, and we prove that $\hat{\phi}(z)$ is analytic everywhere in the complex plane except for the cut $(-\infty, -1]$. In subsections (C) and (D) we derive bounds on $|\hat{\phi}(z)|$ for $\text{Re}z < 0$ and $|z| \gg 1$, which we use in subsection (E) to deform the (inverse Laplace transform) contour integral representation for $\phi(x)$ around the cut. This generates a representation for $\phi(x)$ which contains a function $a(\nu) \in L_2(0, 1)$. Then in subsection (F) we show that this $a(\nu)$ is the unique solution of Eqs. (1.12).

A. Equation for $\hat{\phi}(z)$

We define

$$\alpha = 1/s > 0, \quad (2.1)$$

$$\phi(x) = \frac{1}{2} \int_{-1}^1 \psi(x, \mu) d\mu, \quad (2.2a)$$

$$\hat{\psi}(z, \mu) = \int_0^{\infty} e^{-zx} \psi(x, \mu) dx, \quad (2.2b)$$

and

$$\hat{\phi}(z) = \int_0^{\infty} e^{-zx} \phi(x) dx = \frac{1}{2} \int_{-1}^1 \hat{\psi}(z, \mu) d\mu. \quad (2.2c)$$

We note from Eqs. (1.3) and (2.2c) that for $\text{Re}(z) > 0$, $\hat{\phi}(z)$ is analytic and

$$|\hat{\phi}(z)| \leq M/2\text{Re}(z). \quad (2.3)$$

Now we compute the Laplace Transform of Eqs. (1.1) and (1.2) and rearrange to obtain

$$\hat{\psi}(z, \mu) = \frac{\mu \psi(0, \mu) + c \hat{\phi}(z + \alpha)}{1 + \mu z}. \quad (2.4)$$

Since $\hat{\psi}$ is analytic in z for $\text{Re}(z) > 0$, the above numerator must vanish for $z = -1/\mu$, $-1 \leq \mu < 0$. Hence, the exiting flux is given by

$$\psi(0, \mu) = -(c/\mu) \hat{\phi}(-1/\mu + \alpha), \quad -1 \leq \mu < 0. \quad (2.5)$$

Eqs. (2.5), (2.2), and (1.1)–(1.3) can be used to show that the exiting flux is a Hölder-continuous function of μ for $1 \leq \mu < 0$. We shall need this result below.

Now we integrate Eq. (2.4) over μ and use Eqs. (2.3) and (2.5) to obtain the difference equation

$$\hat{\phi}(z) = F(z) \hat{\phi}(z + \alpha) + G(z), \quad (2.6)$$

$$F(z) = \frac{c}{2} \int_{-1}^1 \frac{1}{1 + \mu z} d\mu, \quad (2.7)$$

$$G(z) = \frac{1}{2} \int_{-1}^1 \frac{\mu \psi(0, \mu)}{1 + \mu z} d\mu. \quad (2.8)$$

We note that F and G are analytic in z off the cuts $(-\infty, -1]$ and $[1, \infty)$.

B. Analytic continuation of $\hat{\phi}(z)$

Let z' be any point which lies neither in the right half-plane ($\text{Re} z' > 0$), nor on the cut $(-\infty, -1]$. Then for some smallest integer n and some $\epsilon > 0$, the ϵ -neighborhood of $z' + n\alpha$, $N_\epsilon(z' + n\alpha)$, lies in the right half-plane, and the sets $N_\epsilon(z' + k\alpha)$, for $k = 0, \dots, n-1$, do not intersect the cuts $(-\infty, -1]$ and $[1, \infty)$. Applying Eq. (2.6) recursively n times, we obtain

$$\hat{\phi}(z) = \left[\prod_{k=0}^{n-1} F(z + k\alpha) \right] \hat{\phi}(z + n\alpha) + \sum_{m=0}^{n-1} \left[\prod_{k=0}^{m-1} F(z + k\alpha) \right] G(z + m\alpha). \quad (2.9)$$

For z in $N_\epsilon(z')$ each term on the right-hand side is analytic, and we have proved

Lemma 1: $\hat{\phi}(z)$ is analytic everywhere in the complex plane except for the cut $(-\infty, -1]$.

C. Bounds on F and G

Let $\text{Im} z \neq 0$ and $\omega = z^{-1}$. Then Eq. (2.8) can be written as

$$G(z) = \frac{1}{2z} \int_{-1}^1 \psi(0, \mu) d\mu - \frac{1}{2z^2} \int_{-1}^1 \frac{\psi(0, \mu)}{\mu + \omega} d\mu.$$

Since $\psi(0, \mu)$ is Hölder-continuous for $0 \leq \mu \leq 1$ and for $-1 \leq \mu \leq 0$, and in general has a jump discontinuity at $\mu = 0$, then one-sided limits for G exist on $(-1, 0)$ and $(0, 1)$, and⁹

$$\int_{-1}^1 \frac{\psi(0, \mu)}{\mu + \omega} d\mu = O(\ln|\omega|), \quad |\omega| \ll 1.$$

This implies

Lemma 2: The function G satisfies

$$G(z) = \frac{1}{2z} \int_{-1}^1 \psi(0, \mu) d\mu + O\left(\frac{1}{|z|^2} \ln|z|\right), \quad |z| \gg 1,$$

and there exists a constant c_1 such that

$$|G(z)| \leq c_1/|z|, \quad |z| \geq 1 + \frac{1}{2}\alpha.$$

Similarly,

Lemma 3: There exists a constant c_2 such that

$$|F(z)| \leq c_2/|z|, \quad |z| \geq 1 + \frac{1}{2}\alpha.$$

D. Bounds on $\hat{\phi}$

We define the points

$$z_n = -1 - n\alpha, \quad n = 0, 1, 2, \dots \quad (2.10)$$

and for any $\epsilon < \min(1, \alpha/2)$ and $n \geq 0$, we define the set

$$S_{n, \epsilon} = \{z \mid |z_n - \text{Re} z| < \alpha/2, \quad \text{Im} z \neq 0, \quad \text{and } |z - z_n| \geq \epsilon\}. \quad (2.11)$$

Equations (2.9) and (2.3) and Lemmas 2 and 3 show that

for each n

$$u_{n, \epsilon} = \sup_{z \in S_{n, \epsilon}} |\hat{\phi}(z)| \quad (2.12)$$

is finite. Also Eqs. (2.6) and Lemmas 2 and 3 give for

$$z \in \bigcup_{n=1}^{\infty} S_{n, \epsilon}, \quad |\hat{\phi}(z)| \leq (1/|z|)[c_1 + c_2|\hat{\phi}(z + \alpha)|], \quad (2.13)$$

with c_1 and c_2 independent of ϵ , so that

$$u_{n, \epsilon} \leq \frac{1}{1 + (n - \frac{1}{2})\alpha} [c_1 + c_2 u_{n-1, \epsilon}], \quad n \geq 1. \quad (2.14)$$

There is then a constant $a > 0$, independent of ϵ , such that

$$u_{n, \epsilon} \leq (a/n)[1 + u_{n-1, \epsilon}], \quad n \geq 1. \quad (2.15)$$

This readily gives the estimate

$$u_{n+p, \epsilon} \leq \frac{a}{n+p} \left[\sum_{k=0}^{p-1} \left(\frac{a}{n}\right)^k + \left(\frac{a}{n}\right)^p u_{n, \epsilon} \right], \quad (2.16)$$

and shows, with $n > a$, that the sequence $\{u_{j, \epsilon}\}_{j=1}^{\infty}$ is bounded and has 0 as limit. This implies that $|\hat{\phi}|$ is bounded in

$\bigcup_{n=0}^{\infty} S_{n, \epsilon}$ and by Eq. (2.3) that there exists a constant c_3 , dependent on ϵ , such that

$$|\hat{\phi}(z)| \leq c_3/|z|, \quad z \in \bigcup_{n=0}^{\infty} S_{n, \epsilon}. \quad (2.17)$$

Next, since $\psi(0, \mu)$ is Hölder-continuous at $\mu = 1$, then $G(z)$, and $F(z)$, have logarithmic singularities at $z = -1$.

Hence, by Eq. (2.6), $\hat{\phi}(z)$ has in general a logarithmic singularity at $z = -1$ as well as at all the points z_n defined by Eq. (2.10). By a straightforward calculation, quite similar to the one which led to Eq. (2.17), we can obtain a constant c_4 , independent of n and ϵ , so that

$$|\hat{\phi}(z)| \leq \frac{c_4}{|z|} \ln \frac{1}{|z - z_n|}, \quad |z - z_n| < \epsilon. \quad (2.18)$$

Now let us define the two limit functions

$$\hat{\phi}^{\pm}(\xi) = \lim_{\epsilon \rightarrow 0^+} \hat{\phi}(\xi \pm i\epsilon).$$

Then Eq. (2.6) and the Hölder-continuity of $F^{\pm}(\xi)$ and $G^{\pm}(\xi)$ for $\xi \leq -1 - \delta$ (for any $\delta > 0$) imply that the $\hat{\phi}^{\pm}(\xi)$ are Hölder-continuous in every closed interval between any two of points z_n defined in Eq. (2.10).

Next, we combine Eqs. (2.6), (2.17), (2.18), and Lemmas 2 and 3 to obtain

$$\hat{\phi}(z) = \frac{1}{2z} \int_{-1}^1 \psi(0, \mu) d\mu + O\left(\frac{1}{|z|^2} \ln|z|\right) + O\left(\frac{1}{|z|^2} \ln \frac{1}{(\min_n |z - z_n|)}\right).$$

Hence for $\xi < -1$ and $\xi \notin \{z_n\}$,

$$\hat{\phi}^+(\xi) - \hat{\phi}^-(\xi) = b(\xi)/\xi^2, \quad (2.19)$$

where

$$|b(\xi)| = O(\ln|\xi|) + O\left(\ln \frac{1}{(\min_n |\xi - z_n|)}\right). \quad (2.20)$$

Thus $b(\xi)$ has logarithmic singularities at each z_n , is Hölder-continuous in every closed subinterval between any two consecutive z_n , and an elementary computation for $b(\xi)$ in Eq. (2.20) gives

$$\int_{-\infty}^{-1} \frac{|b(\xi)|^m}{\xi^2} d\xi < \infty$$

for $m = 1$ and 2 .

We now summarize our results in:

Lemma 4: Let z_n be defined by Eq. (2.10) and $S_{n,\epsilon}$ by Eq. (2.11). Then there exist constants c_3 and c_4 such that

$$|\hat{\phi}(z)| < \frac{c_3}{|z|}, \quad z \in \bigcup_{n=0}^{\infty} S_{n,\epsilon}$$

and

$$|\hat{\phi}(z)| < \frac{c_4}{|z|} \ln \frac{1}{|z - z_n|}, \quad |z - z_n| < \epsilon.$$

Functions $\hat{\phi}^{\pm}(\xi)$ and $b(\xi)$ related by

$$\hat{\phi}^+(\xi) - \hat{\phi}^-(\xi) = b(\xi)/\xi^2$$

are Hölder-continuous in every closed subinterval between any two consecutive z_n , and

$$\int_{-\infty}^{-1} \frac{|b(\xi)|^m}{\xi^2} d\xi < \infty$$

for $m = 1$ and 2 .

E. Representation of $\phi(x)$ via inverse Laplace transform

For $x > 0$, the inverse Laplace transform gives

$$\phi(x) = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} e^{zx} \hat{\phi}(z) dz, \quad \gamma > 0.$$

By Lemma 4, we can deform the contour of integration around the cut $(-\infty, -1]$ and obtain

$$\phi(x) = -\frac{1}{2\pi i} \int_{-\infty}^{-1} e^{\xi x} \frac{b(\xi)}{\xi^2} d\xi. \quad (2.21)$$

This integral converges for all $x \geq 0$, due to the final bound in Lemma 4. Introducing $\nu = -\xi^{-1}$ in Eq. (2.21) gives

$$\phi(x) = \frac{1}{2} \int_0^1 a(\nu) e^{-x/\nu} d\nu, \quad (2.22)$$

where

$$a(\nu) = -(1/\pi i) b(-1/\nu), \quad 0 < \nu < 1. \quad (2.23)$$

Using the properties of $b(\xi)$ stated in Lemma 4, we have

Lemma 5. There exists a function $a(\nu)$ such that

$$\phi(x) = \frac{1}{2} \int_0^1 a(\nu) e^{-x/\nu} d\nu.$$

Also, $a(\nu)$ (i) has a logarithmic singularity at each point

$$\nu_n = 1/(1 + n\alpha), \quad n = 0, 1, 2, \dots \quad (2.24)$$

(ii) is Hölder-continuous in every closed interval between any two consecutive ν_n , and (iii) satisfies

$$\int_0^1 |a(\nu)|^m d\nu < \infty,$$

for $m = 1$ and 2 . Thus ϕ in Eq. (2.2a) can be extended to complex x to be analytic for $\text{Re}(x) > 0$.

F. Proof of Theorem 1

We convert the transport problem (1.1)–(1.3) into the standard Peierl's integral equation for $\phi(x)$, by inverting the operator on the right side of Eq. (1.1) to solve explicitly for ψ , integrating over μ , and using the definition (2.2). Then we introduce the form (2.22) into this integral equation and perform some elementary operations to obtain

$$\begin{aligned} & \int_0^1 f(\nu) e^{-x/\nu} d\nu \\ &= \int_0^1 a(\nu) e^{-x/\nu} d\nu \\ &+ \frac{c}{2} \int_0^1 \left[\int_0^1 \frac{a(\mu)}{1 - \nu(1/\mu + 1/s)} d\mu \right] e^{-x/\nu} d\nu \\ &- \frac{c}{2} \int_0^1 \left[\int_{-1}^1 \frac{d\mu}{1 - \mu(1/\nu + 1/s)} \right] a(\nu) e^{-(1/\nu + 1/s)x} d\nu. \end{aligned} \quad (2.25)$$

In the last two integrals on the right side of this equation the inner integrals are in the Cauchy principal-value sense, the first of which exists in the L_2 sense because by Lemma 5, $a(\mu)$ is in L_2 .

The change of variables

$$1/\omega = 1/\nu + 1/s$$

now converts Eq. (2.25) into

$$\int_0^1 l(\nu) e^{-x/\nu} d\nu = 0, \quad x \geq 0, \quad (2.26)$$

where, with the operator L defined in Eq. (1.12b),

$$l(\nu) = f(\nu) - (I + \frac{1}{2}cL)a(\nu), \quad (2.27)$$

is in $L_2[0, 1]$, and hence in $L_1[0, 1]$. Since Eq. (2.26) implies that a function in $L_1[0, \infty)$ has a Laplace transform identically zero, it follows that $l(\nu) = 0$ almost everywhere. It follows from Eq. (2.27) that the function $a(\nu)$ in Lemma 5 satisfies Eqs. (1.12) almost everywhere. [It is easy to show that the points at which Eqs. (1.12) are not satisfied are exactly the points ν_n defined in Eq. (2.24).] Thus, there exists a solution of Eqs. (1.12). It remains to establish uniqueness.

Suppose there exist two L_2 solutions of Eqs. (1.12), $a_1(\nu)$ and $a_2(\nu)$. We define

$$\phi_j(x) = \int_0^1 a_j(\nu) e^{-x/\nu} d\nu, \quad j = 1, 2.$$

Since a_1 and a_2 satisfy Eqs. (1.12), then they satisfy Eq. (2.25), which implies that both ϕ_1 and ϕ_2 satisfy the Peierl's integral equation. But by assumption, the transport problem (1.1)–(1.3) has a unique solution, and this is also true for the Peierl's equation. Thus $\phi_1(x) = \phi_2(x)$ for $x \geq 0$, and $a_1(\nu) = a_2(\nu)$ almost everywhere, by uniqueness of the Laplace transform of L_1 functions. This completes the proof of Theorem 1. Q. E. D.

III. SOLUTION OF THE INTEGRAL EQUATION FOR THE EXITING FLUX

In this section we shall prove that Martin's integral equation (1.14) possesses a unique solution. This result is contained in:

Theorem 2: The singular integral equation for the exit-

ing flux,

$$\int_0^1 \mu \psi_\nu(0, \mu) \psi(0, -\mu) d\mu = \int_0^1 \mu \psi_\nu(0, -\mu) f(\mu) d\mu, \quad 0 < \nu \leq 1, \quad (3.1)$$

has a unique solution $\psi(0, -\mu) \in L_2(0, 1)$. Moreover, this solution is analytic in μ for $\text{Re}(\mu) > 0$.

Proof: Since $f(\mu)$ is Hölder-continuous, there exists a unique function $a(\nu)$ in $L^2[0, 1]$, satisfying the conditions of Theorem 1, such that

$$f(\mu) = \int_0^1 a(\nu) \psi_\nu(0, \mu) d\nu, \quad (3.2)$$

and such that

$$\phi(x) = \frac{1}{2} \int_0^1 a(\nu) e^{-x/\nu} d\nu, \quad x \geq 0 \quad (3.3)$$

for ϕ in Eq. (2.2a). From Eqs. (2.5) we have

$$\psi(0, -\mu) = \frac{c}{\mu} \int_0^\infty e^{-x(1/\mu) + 1/s} \phi(x) dx, \quad 0 < \mu \leq 1,$$

and from Eq. (3.3) the representation

$$\psi(0, -\mu) = \int_0^1 a(\nu) \psi_\nu(0, -\mu) d\nu. \quad (3.4)$$

Since Eq. (3.2) holds, then the analysis of Mullikin and Siewert, outlined in Sec. 1, shows that the function defined by Eq. (3.4) is a solution of Eq. (3.1). This solution is clearly analytic in $\text{Re}(\mu) > 0$. Uniqueness of the solution has been shown⁷ in a certain Hilbert space of analytic functions, but we proceed to show this in $L^2[0, 1]$.

If $D(\mu)$ is the difference of two solutions in $L^2[0, 1]$ to Eq. (3.1), then

$$\int_0^1 \mu \psi_\nu(0, \mu) D(\mu) d\mu = 0, \quad 0 < \nu \leq 1. \quad (3.5)$$

Let $b_n(\nu)$ in $L^2(0, 1)$ be the half-range coefficients of μ^n given by Theorem 1, i.e.,

$$\mu^n = \int_0^1 b_n(\nu) \psi_\nu(0, \mu) d\nu.$$

Then Eq. (3.5) gives

$$\begin{aligned} 0 &= \int_{-0}^1 b_n(\nu) \int_0^1 \mu \psi_\nu(0, \mu) D(\mu) d\mu d\nu \\ &= \int_0^1 \int_0^1 b_n(\nu) \psi_\nu(0, \mu) d\nu \mu D(\mu) d\mu \\ &= \int_0^1 \mu^n \mu D(\mu) d\mu, \quad n \geq 0. \end{aligned}$$

The interchange of integration for the principal-value integral part of $\psi_\nu(0, \mu)$ is permissible since $b_n(\nu)$ and $D(\mu)$ are in L_2 (cf. Ref. 13, p. 170). Hence $\mu D(\mu)$ is orthogonal to every polynomial in μ , so $D(\mu) = 0$ almost everywhere. This completes the proof of the theorem. Q. E. D.

IV. DISCUSSION

The results of Sec. II show that for $0 < s < \infty$, the continuum eigenfunctions of Mullikin and Siewert are half-range complete. However for $s = \infty$, it is well known that

the continuum solutions alone are not half-range complete; a discrete solution, linearly independent of the continuum solutions, is extant, and must be appended to the continuum solutions to have half-range completeness.¹⁰ Thus the transport problem (1.1)–(1.3) has a spectral discontinuity in passing from $s < \infty$ to $s = \infty$. (This is likely related to the fact that the Peierl's equation passes from discrete to continuous spectrum in passing from $s < \infty$ to $s = \infty$.)

On the other hand, from a physical point of view, the angular flux ψ should pass continuously to its value at $s = \infty$ in the limit as $s \rightarrow \infty$. (This continuity of ψ at $s = \infty$ is in fact straightforward to prove, but we shall not do this here.) Therefore, we should expect a nonuniform behavior in the expansion coefficients as $s \rightarrow \infty$. In fact, the nature of this nonuniform behavior is demonstrated in Sec. II, where it is shown that the half-range coefficients $a(\nu)$ for a general Hölder-continuous function $f(\mu)$ have logarithmic singularities at the infinite denumerable set of points

$$\nu_n = \frac{1}{1 + n/s}, \quad n = 0, 1, 2, \dots$$

These points become dense in the continuum $0 \leq \nu \leq 1$ as $s \rightarrow \infty$.

Thus, one can think of the spectral discontinuity at $s = \infty$ as being compensated by the increasingly pathological behavior of the half-range expansion coefficients as $s \rightarrow \infty$, so that the physical solution passes continuously as $s \rightarrow \infty$ to its $s = \infty$ value.

Since the half-range coefficients have the singular behavior described above, it appears hopeless to try to perform direct numerical computing of them in order to generate numerical solutions of boundary value problems. However, the development of these eigenfunctions and the proving of their half-range completeness does have the payoff of Theorem 2.

Finally, we note that the method used to prove Theorem 2 can be applied to transport problems in other types of nonhomogeneous media, provided a number of conditions are met. To illustrate, let us consider the problem

$$\mu \frac{\partial}{\partial x} \psi(x, \mu) + \psi(x, \mu) = \frac{c(x)}{2} \int_{-1}^1 \psi(x, \mu') d\mu', \quad 0 < x, \quad (4.1)$$

$$\psi(0, \mu) = f(\mu), \quad 0 < \mu < 1, \quad (4.2)$$

where f is prescribed and Hölder-continuous, and

$$|\psi(x, \mu)| < M, \quad 0 \leq x < \infty, \quad -1 \leq \mu \leq 1. \quad (4.3)$$

We assume this problem to have a unique solution (i.e., we require the half-space $x > 0$ to be subcritical).

Suppose that a family $\psi_\nu(z, \mu)$, $\nu \in \Sigma$, of solutions of Eqs. (4.1), (4.3) has been found. (We take Σ to consist of the continuum plus possibly a finite number of discrete points.

Also, we denote $\int_\Sigma (\cdot) d\nu$ as integration over the continuum plus summation over the discrete points.) Moreover, suppose that full-range orthogonality

$$\int_{-1}^1 \mu \psi_\nu(0, -\mu) \psi_\nu(0, \mu) d\mu = 0, \quad \nu', \nu \in \Sigma,$$

is satisfied, that for any Hölder-continuous f there exists a

unique function $a(\nu)$ such that

$$f(\mu) = \int_{\Sigma} a(\nu) \psi_{\nu}(0, \mu) d\nu, \quad 0 < \mu \leq 1 \quad \text{a.e.},$$

and that

$$\int_{\Sigma} a(\nu) \psi_{\nu}(0, -\mu) d\nu$$

is Hölder-continuous for $0 < \mu \leq 1$. Then, following the analysis of Mullikin and Siewert³ (see Sec. 1), the integral equation for the exiting flux is

$$\int_0^1 \mu \psi_{\nu}(0, \mu) \psi(0, -\mu) d\mu = \int_0^1 \mu \psi_{\nu}(0, -\mu) f(\mu) d\mu, \quad \nu \in \Sigma. \quad (4.4)$$

We can now repeat the analysis in Sec. III, and find that

$$\psi(0, -\mu) = \int_{\Sigma} a(\nu) \psi_{\nu}(0, -\mu) d\nu$$

is the unique solution of the integral equation (4.4).

The analysis outlined in the above paragraph applies not only to

$$c(z) = c_0 e^{-z/s}, \quad (4.5)$$

which is discussed earlier in this article, but also specifically to

$$c(z) = c_0 \quad (c_0 \leq 1)$$

and

$$c(z) = \frac{c_0 + kbe^{-z/s}}{1 + be^{-z/s}}. \quad (4.6)$$

The case $c(z) = c_0$ has been thoroughly studied,¹⁰ and a well-known set of “continuum plus one discrete” solutions satis-

fying all the conditions of the above paragraph is extant. We refer the reader to Ref. 10 for details. [Also, we note that the unique existence of a solution to Eq. (4.4) for $c(z) = c_0$ has been proved using other methods.¹¹] The function $c(z)$ defined by Eq. (4.6) has also been studied,¹² and again, “continuum plus one discrete” solutions satisfying all the conditions of the above paragraph for a restricted set of values of c_0, k, b , and s have been found. Thus for this case also, Eq. (4.4) has a unique solution given by Eq. (4.5).

¹J. B. Chamberlain and M. B. McElroy, *Astrophys. J.* **144**, 1148 (1966).

²B. J. Martin, *SIAM J. Appl. Math.* **20**, 703 (1971).

³T. W. Mullikin and C. E. Siewert, “Radiative Transfer in Inhomogeneous Atmospheres,” *Ann. Nucl. Energy* **7**, 205 (1980).

⁴E. W. Larsen, G. C. Pomraning, and V. C. Badham, *J. Math. Phys.* **21**, 2448 (1980).

⁵C. E. Siewert and P. Benoist, *Nucl. Sci. Eng.*, **69**, 156 (1979); P. Grandjean and C. E. Siewert, *Nucl. Sci. Eng.*, **69**, 161 (1979).

⁶R. D. M. Garcia and C. E. Siewert, “Radiative Transfer in Inhomogeneous Atmospheres—Numerical Results,” *J. Quant. Spectros. Radiat. Transfer* (to appear).

⁷T. W. Mullikin, “Some Singular Integral Equations in Linear Transport Theory,” Sandia National Laboratories Report SAND 80-1069 (1980). (A new version of this report, with error estimates, is in preparation by C. T. Kelley and T. W. Mullikin.)

⁸Actually, Larsen and Pomraning in Ref. 4 proved half-range completeness for a slightly sharper inequality than (1.4).

⁹N. I. Muskhelishvili, *Singular Integral Equations* (Noordhoff, Groningen, 1953).

¹⁰K. M. Case and P. F. Zweifel, *Linear Transport Theory* (Addison-Wesley, Reading, Mass., 1967).

¹¹T. W. Mullikin, *Trans. Amer. Math. Soc.*, **113**, 316 (1964).

¹²G. C. Pomraning and E. W. Larsen, *J. Math. Phys.*, **21**, 1603 (1980).

¹³F. G. Tricomi, *Integral Equations* (Interscience, New York, 1957).

Local gauge field theory formalism

R. J. McKellar

Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, N2L 3G1 Canada

(Received 10 December 1979; accepted for publication 18 March 1980)

In this paper we develop a local gauge field theory formalism which is designed for use in a concomitant approach to motivate the field equations of gauge field theories. Several interesting identities are derived which enable us to gain further insight into these theories and into previously used formalisms and techniques.

PACS numbers: 11.10.Np, 03.70. + k

1. INTRODUCTION

Ever since Yang and Mills¹ first introduced their well-known theory, gauge field theories have generated much interest. As in many other theories, it is sometimes difficult to motivate a particular set of field equations. In the area of relativity, much success has been met by using a concomitant approach.²⁻⁴ The techniques involved in such an approach rely heavily on the use of local coordinates. To deal with gauge field theories it is necessary to introduce local coordinates on both the usual underlying manifold of relativity X_n and the Lie group G . The formalism outlined in this paper shares features with the formalisms of both Utiyama⁵ and Kibble⁶ as well as those of Yang⁷ and Rund.⁸ Additional features are also present which give rise to identities that can be used to motivate gauge field theories from a slightly different viewpoint. In particular, the usual restriction to infinitesimal gauge transformations is avoided. This new formalism is not based upon a fiber bundle approach;⁹ however, certain aspects of fiber bundles do play a role in it.

On the manifold X_n with local coordinates x^i , $i = 1, \dots, n$, it has become common¹⁰ to represent the components of a tensor, a set of tensors, or similar geometric objects by a single symbol with a single upper case Latin letter as superscript, e.g., ρ^A , $A = 1, \dots, M$. Under a coordinate transformation

$$\bar{x}^i = \bar{x}^i(x^j), \quad (1.1)$$

with

$$J^i_j \equiv \frac{\partial x^i}{\partial \bar{x}^j}$$

and

$$J \equiv \det J^i_j \neq 0,$$

the transformation law of the ρ^A 's we are interested in can be expressed in the form

$$\bar{\rho}^A = C^A_B \rho^B. \quad (1.2)$$

The Einstein summation convention is also in effect for these new indices, so that repeated upper case Latin indices are summed from 1 to M . This law is such that

$$\text{with } \left. \begin{aligned} C^A_B &= C^A_B(J^i_j), \\ \det C^A_B &\neq 0, \end{aligned} \right\} \quad (1.3)$$

and furthermore, under a second coordinate transformation

$$\bar{\bar{x}}^i = \bar{\bar{x}}^i(\bar{x}^j)$$

with

$$K^i_j \equiv \frac{\partial \bar{x}^i}{\partial \bar{\bar{x}}^j}$$

and

$$K \equiv \det K^i_j \neq 0,$$

C^A_B satisfies the relation characteristic of a right action, viz.,

$$C^A_B(K^i_j)C^B_D(J^j_k) = C^A_D(J^i_h K^h_j). \quad (1.4)$$

A similar description can be developed when the quantity ρ^A is also subjected to a gauge transformation (see, e.g., Yang⁷) by an element $u = u(x^i)$ of an m -dimensional Lie group G . Relative to a canonical chart of the first kind¹¹ for G at its identity e , the coordinates (or components) of u (labeled with lower case Greek letters),

$$u^\alpha = u^\alpha(x^i), \quad \alpha = 1, \dots, m,$$

are such that $u^\alpha = 0$ for all α if and only if $u = e$ and the coordinates of u^{-1} are $-u^\alpha$. We shall be concerned with operations of the form $w: G \times G \rightarrow G$ such that

(i) w is analytic,

(ii) $w(e, u) = w(u, e) = u$,

and

(iii) $w(u, u^{-1}) = e$.

Operations which satisfy (i)–(iii) shall be referred to as *actions*. Two simple examples are the left action of u on v , i.e.,

$$w(u, v) = L_u v = uv,$$

and the right action of u on v , i.e.,

$$w(u, v) = R_u v = vu.$$

A nonassociative action is given by

$$w(u, v) = uv(vu)^{-1}uv.$$

It will be assumed that the quantity ρ^A transforms under a gauge transformation as

$$\bar{\rho}^A = T^A_B \rho^B, \quad (1.5)$$

where

$$\text{and } \left. \begin{aligned} T^A_B &= T^A_B(u^\alpha) \\ \det T^A_B &\neq 0. \end{aligned} \right\} \quad (1.6)$$

In addition, T^A_B satisfies the relation

$$T^A_B(v^\alpha)T^B_D(u^\alpha) = T^A_D[w^\alpha(u^\beta, v^\beta)], \quad (1.7)$$

where $w^\alpha(u^\beta, v^\beta)$ denotes the coordinates of $w(u, v)$ relative to the same chart expressed in terms of u^β and v^β , the coordinates of u and v respectively. T_B^A 's which satisfy (1.6) and (1.7) will be referred to as *Lie group representations of type w* of the Lie group G .

For the quantity ρ^A we would like to define a process of differentiation denoted by $\rho^A_{||a}$ which is such that under a coordinate transformation

$$\bar{\rho}^A_{||a} = C_B^A J_a^b \rho^B_{||b}, \quad (1.8)$$

while, under a gauge transformation

$$\dot{\rho}^A_{||a} = T_B^A \rho^B_{||a}. \quad (1.9)$$

This type of differentiation has been termed *double covariant differentiation* by Yang.⁷ Section 2 is devoted to obtaining identities from (1.3), (1.4), (1.6), and (1.7), which enable us to define a $\rho^A_{||a}$ which satisfies (1.8) and (1.9). We shall then assume that $\rho^A_{||a}$ takes the form

$$\rho^A_{||a} = \rho^A_{,a} - \bar{C}_{Br}^{As} \Gamma_s^r{}_a \rho^B - \bar{T}_{B\alpha}^A A_a^\alpha \rho^B, \quad (1.10)$$

where a comma denotes partial differentiation with respect to the local coordinates,

$$\bar{C}_{Br}^{As} \equiv \left. \frac{\partial C_B^A(J_j^a)}{\partial J_s^r} \right|_{J_j^j = \delta_j^j},$$

$$\bar{T}_{B\alpha}^A \equiv \left. \frac{\partial T_B^A(u^\beta)}{\partial u^\alpha} \right|_{u^\gamma = 0},$$

and $\Gamma_s^r{}_a$ and A_a^α are new quantities. In Sec. 3 we determine the transformation properties of $\Gamma_s^r{}_a$ and A_a^α under both kinds of transformations by demanding that (1.8) and (1.9) hold. The resulting properties lead us to call $\Gamma_s^r{}_a$ a linear connection and A_a^α a *gauge connection* (also referred to as a gauge potential and, in some cases, a gauge field).

Virtually any gauge field theory which is derivable from a variational principle employs a Lagrangian which is at most second order in the field variables ρ^A and first order in both the linear connection $\Gamma_b^a{}_c$ and the gauge connection A_a^α , i.e.,

$$L = L(\rho^A, \rho^A_{,a}; \rho^A_{,ab}; \Gamma_b^a{}_c; \Gamma_b^a{}_{c,d}; A_a^\alpha; A_{a,b}^\alpha). \quad (1.11)$$

Such Lagrangians are scalar densities under a coordinate transformation, i.e.,

$$\bar{L} = JL, \quad (1.12)$$

and scalars under a gauge transformation, i.e.,

$$\dot{L} = L. \quad (1.13)$$

By virtue of both transformation laws, (1.12) and (1.13), several invariance identities² which characterize L are obtained in Sec. 4.

2. TRANSFORMATION IDENTITIES

Since C_B^A is such that

$$\det C_B^A \neq 0,$$

C_B^A has an inverse which we denote by \hat{C}_B^A . By evaluating

$$C_B^A(K_j^i) C_D^B(J_j^i) = C_D^A(J_h^i K_j^h) \quad (2.1)$$

at $K_j^i = \delta_j^i$ and then $K_j^i = \hat{J}_j^i$, where \hat{J}_j^i is the inverse of J_j^i , it is possible to establish the properties

$$C_B^A(\delta_j^i) = \delta_B^A$$

and

$$C_B^A(\hat{J}_j^i) = \hat{C}_B^A(J_j^i).$$

At this point it should be noted that a quantity ρ^A whose C_B^A satisfies (2.1) need not be tensorial.¹²

In order to define a process of double covariant differentiation, it is useful to obtain three identities from (2.1). The derivative of (2.1) with respect to J_s^r , followed by evaluation at $J_b^a = \delta_b^a$, yields

$$C_B^A(K_j^i) \bar{C}_{Dr}^{Bs} = C_{Dr}^{At}(K_j^i) K_t^s, \quad (2.2)$$

where

$$C_{Dr}^{At}(K_j^i) \equiv \frac{\partial C_D^A(K_j^i)}{\partial K_t^r},$$

and we recall

$$\bar{C}_{Dr}^{Bs} \equiv C_{Dr}^{Bs}(\delta_j^i).$$

By taking the derivative of (2.1) with respect to K_b^a , and then evaluating at $K_b^a = \delta_b^a$, we obtain

$$\bar{C}_{Br}^{As} C_D^B(J_j^i) = C_{Dt}^{As}(J_j^i) J_t^r. \quad (2.3)$$

For our purposes we will express (2.2) and (2.3) in the forms

$$C_B^A \bar{C}_{Dr}^{Bs} = C_{Dr}^{At} J_t^s \quad (2.4)$$

and

$$\bar{C}_{Br}^{As} C_D^B = C_{Dt}^{As} J_t^r, \quad (2.5)$$

it being understood that the arguments of C_B^A and C_{Dr}^{At} are J_j^i . These are two of the three identities referred to. For the third identity we take the derivative of (2.4) with respect to J_b^a and then evaluate at $J_j^i = \delta_j^i$ to find

$$\bar{C}_{Ba}^{Ab} \bar{C}_{Dr}^{Bs} = \bar{C}_{Dra}^{Asb} + \bar{C}_{Dr}^{Ab} \delta_a^s,$$

where

$$\bar{C}_{Dra}^{Asb} \equiv \left. \frac{\partial^2 C_D^A}{\partial J_s^r \partial J_b^a} \right|_{J_j^j = \delta_j^j} = \bar{C}_{Dar}^{Asb}.$$

By making use of the above symmetry, we finally obtain the desired identity, viz.,

$$\bar{C}_{Ba}^{Ab} \bar{C}_{Dr}^{Bs} - \bar{C}_{Br}^{As} \bar{C}_{Da}^{Bb} = \bar{C}_{Dr}^{Ab} \delta_a^s - \bar{C}_{Da}^{As} \delta_r^b. \quad (2.6)$$

It is interesting to note that since \bar{C}_{Br}^{As} is constructed out of δ_j^i , it is numerically the same in all coordinate systems and so it is invariant under coordinate transformations, i.e.,

$$\bar{\bar{C}}_{Br}^{As} = \bar{C}_{Br}^{As}, \quad (2.7)$$

but, by combining (2.4), (2.5) and (2.7), we obtain

$$\bar{\bar{C}}_{Br}^{As} = C_D^A \hat{C}_B^E \hat{J}_a^s J_r^b \bar{C}_{Eb}^{Da}. \quad (2.8)$$

Thus the transformation law of $\bar{\bar{C}}_{Br}^{As}$ can also be expressed in the general form of (1.2), subject to (1.3) and (1.4). This situation is reminiscent of the Kronecker delta which is invariant under (1.1), and yet, is also a tensor of contravariant valency 1 and covariant valency 1.

We shall now follow a similar analysis for gauge transformations. In the process, we shall derive several interesting identities in the area of Lie groups. The global conditions (i)–(iii) which define an action can be expressed in terms of the local coordinates as:

(i) w^α can be expanded as a power series in both u^β and v^β about the origin,

$$(ii) w^\alpha(0, u^\beta) = w^\alpha(u^\beta, 0) = u^\alpha,$$

and

$$(iii) w^\alpha(u^\beta, -u^\beta) = 0.$$

When the local coordinates of the action w, w^α are expanded as a power series in u^β we obtain

$$w^\alpha = v^\alpha + A_{\beta}^{\alpha}(v^\mu)u^\beta + A_{\beta\gamma}^{\alpha}(v^\mu)u^\beta u^\gamma + \dots,$$

where

$$A_{\beta}^{\alpha} \equiv \left. \frac{\partial w^\alpha}{\partial u^\beta} \right|_{u^\gamma=0}.$$

From condition (ii) we see that

$$A_{\beta}^{\alpha}(0) = \delta_{\beta}^{\alpha},$$

and hence

$$\det A_{\beta}^{\alpha} \neq 0$$

in a neighborhood of the identity. Thus A_{β}^{α} has an inverse which we denote by \hat{A}_{β}^{α} . Expansion of A_{β}^{α} and $A_{\beta\gamma}^{\alpha}$ as a power series in v^μ leads to

$$w^\alpha = v^\alpha + u^\alpha + A_{\beta,\gamma}^{\alpha}(0)v^\gamma u^\beta + A_{\beta\gamma}^{\alpha}(0)u^\beta u^\gamma + \dots,$$

where

$$A_{\beta,\gamma}^{\alpha} \equiv \frac{\partial A_{\beta}^{\alpha}}{\partial v^\gamma}(v^\mu).$$

Conditions (ii) and (iii) imply

$$A_{\beta\gamma}^{\alpha}(0) = 0$$

and

$$A_{(\beta,\gamma)}^{\alpha}(0) = 0,$$

respectively, where parentheses around indices denote symmetrization. By expanding w^α first as a power series in v^β we find, in a similar manner,

$$w^\alpha = u^\alpha + v^\alpha + B_{\beta,\gamma}^{\alpha}(0)u^\gamma v^\beta + \dots,$$

where

$$B_{\beta}^{\alpha} \equiv \left. \frac{\partial w^\alpha}{\partial v^\beta} \right|_{v^\gamma=0},$$

$$B_{\beta,\gamma}^{\alpha} \equiv \frac{\partial B_{\beta}^{\alpha}}{\partial u^\gamma}(u^\mu),$$

$$B_{(\beta,\gamma)}^{\alpha}(0) = 0,$$

and B_{β}^{α} has an inverse \hat{B}_{β}^{α} in a neighborhood of the identity. Comparison of the two expansions leads to

$$w^\alpha = u^\alpha + v^\alpha + \frac{1}{2}C_{\beta}^{\alpha}{}_{\gamma} u^\beta v^\gamma + \dots,$$

where the $C_{\beta}^{\alpha}{}_{\gamma}$'s are constants given by

$$\begin{aligned} C_{\beta}^{\alpha}{}_{\gamma} &\equiv 2A_{\beta,\gamma}^{\alpha}(0) = -2B_{\beta,\gamma}^{\alpha}(0) \\ &= -C_{\gamma}^{\alpha}{}_{\beta}. \end{aligned}$$

One tends to think of the $C_{\beta}^{\alpha}{}_{\gamma}$'s as the structure constants of the group G ; however this is not necessarily the case.

By virtue of the fact that Lie group representations of type w T_B^A satisfy

$$\det T_B^A \neq 0,$$

T_B^A has an inverse which we denote by \hat{T}_B^A . When the relation

$$T_B^A(v^\alpha)T_D^B(u^\alpha) = T_D^A[w^\alpha(u^\beta, v^\beta)] \quad (2.9)$$

is evaluated at $v^\alpha = 0$ and then $v^\alpha = -u^\alpha$, we obtain

$$T_B^A(0) = \delta_B^A$$

and

$$T_B^A(-u^\alpha) = \hat{T}_B^A(u^\alpha),$$

respectively.

As in the case of the transformation law (2.1), we will now obtain three identities from (2.9) which are useful in defining a process of double covariant differentiation. We first take the derivative of (2.9) with respect to u^α and then evaluate at $u^\gamma = 0$ to obtain

$$T_B^A(v^\gamma)\tilde{T}_{D\alpha}^B = T_{D\beta}^A(v^\gamma)A_{\alpha}^{\beta}, \quad (2.10)$$

where

$$T_{D\beta}^A(v^\gamma) \equiv \frac{\partial T_D^A}{\partial v^\beta}(v^\gamma)$$

and recall

$$\tilde{T}_{D\alpha}^B \equiv T_{D\alpha}^B(0).$$

By taking the derivative of (2.9) with respect to v^α and then evaluating at $v^\gamma = 0$ we find that

$$\tilde{T}_{B\alpha}^A T_D^B(u^\gamma) = T_{D\beta}^A(u^\gamma)B_{\alpha}^{\beta}. \quad (2.11)$$

For our purposes we will express (2.10) and (2.11) in the forms

$$T_B^A \tilde{T}_{D\alpha}^B = T_{D\beta}^A A_{\alpha}^{\beta} \quad (2.12)$$

and

$$\tilde{T}_{B\alpha}^A T_D^B = T_{D\beta}^A B_{\alpha}^{\beta}, \quad (2.13)$$

it being understood that the arguments are u^γ . The third identity is obtained by first taking the derivative of (2.12) with respect to u^γ and then evaluating at $u^\gamma = 0$, which yields

$$\tilde{T}_{B\gamma}^A \tilde{T}_{D\alpha}^B = \tilde{T}_{D\alpha\gamma}^A + \frac{1}{2}\tilde{T}_{D\beta}^A C_{\alpha}^{\beta}{}_{\gamma}, \quad (2.14)$$

where

$$\tilde{T}_{D\alpha\gamma}^A \equiv \left. \frac{\partial^2 T_D^A(u^\beta)}{\partial u^\alpha \partial u^\gamma} \right|_{u^\nu=0} = \tilde{T}_{D\gamma\alpha}^A.$$

We then antisymmetrize (2.14) to arrive at the desired equation, viz.,

$$\tilde{T}_{B\alpha}^A \tilde{T}_{D\gamma}^B - \tilde{T}_{B\gamma}^A \tilde{T}_{D\alpha}^B = -C_{\alpha}^{\beta}{}_{\gamma} \tilde{T}_{D\beta}^A, \quad (2.15)$$

which is reminiscent of the well-known commutation law for the generators of a Lie algebra representation of a Lie group.

An extremely important Lie group representation of type w is the quantity

$$T_{\beta}^{\alpha} \equiv \hat{B}_{\gamma}^{\alpha} A_{\beta}^{\gamma}.$$

The invertibility of T_{β}^{α} follows from the invertibility of B_{β}^{α} and A_{β}^{α} . In order to illustrate that T_{β}^{α} satisfies (2.9), we suppose that there exists some Lie group representation T_B^A of type w . When (2.12) is evaluated at w^ν , we have

$$T_B^A(w^\nu)\tilde{T}_{D\alpha}^B = T_{D\beta}^A(w^\nu)A_{\alpha}^{\beta}(w^\nu). \quad (2.16)$$

By making use of (2.9) on the left-hand side and (2.13) on the

right-hand side of (2.16) we obtain

$$T_C^A(v^\nu)T_B^C(u^\nu)\tilde{T}_{D\alpha}^B = \tilde{T}_{B\gamma}^A T_D^B(w^\nu)\hat{B}_\beta^\gamma(w^\nu)A_\alpha^\beta(w^\nu).$$

Repeated applications of (2.12) and (2.13) evaluated at both u^ν and v^ν on the left-hand side of the above lead to

$$\begin{aligned} \tilde{T}_{B\tau}^A T_C^B(v^\nu)T_D^C(u^\nu)\hat{B}_\mu^\tau(v^\nu)A_\gamma^\mu(v^\nu)\hat{B}_\beta^\gamma(u^\nu)A_\alpha^\beta(u^\nu) \\ = \tilde{T}_{B\gamma}^A T_D^B(w^\nu)\hat{B}_\beta^\gamma(w^\nu)A_\alpha^\beta(w^\nu). \end{aligned}$$

By virtue of (2.9) and the invertibility of T_D^B we see that

$$\tilde{T}_{B\tau}^A T_\gamma^\tau(v^\nu)T_\alpha^\gamma(u^\nu) = \tilde{T}_{B\tau}^A T_\alpha^\tau(w^\nu).$$

Thus, provided G admits at least one Lie group representation of type w for which

$$g_{\beta\tau} \equiv \tilde{T}_{A\beta}^B \tilde{T}_{B\tau}^A$$

is nondengenerate, i.e.,

$$\det g_{\beta\tau} \neq 0, \quad (2.17)$$

we have that T_β^α satisfies (2.9), viz.,

$$T_\gamma^\tau(v^\nu)T_\alpha^\gamma(u^\nu) = T_\alpha^\tau(w^\nu).$$

A Lie group representation of type w that satisfies (2.17) will be called a semisimple Lie group representation of type w . Note that it is possible for G to admit a semisimple Lie group representation of type w without

$$h_{\beta\tau} \equiv C_{\gamma\beta}^\alpha C_{\alpha\tau}^\gamma$$

being nondegenerate.

For the most relevant case, viz., when w is the left action of u on v , i.e.,

$$w(u,v) = L_u v = uv,$$

the $C_{\beta\gamma}^\alpha$'s are the structure constants of the Lie group and (2.9) reduces to the condition that T_B^A is a right action. It can be shown that T_β^α is the inverse of the adjoint representation as follows. First of all, the associativity of multiplication is expressed as

$$w^\alpha[w^\beta(x^\gamma, y^\gamma), z^\beta] = w^\alpha[x^\beta, w^\beta(y^\gamma, z^\gamma)].$$

By setting $y^\gamma = -x^\gamma$ we obtain

$$z^\alpha = w^\alpha[x^\beta, w^\beta(-x^\gamma, z^\gamma)].$$

When the partial derivative of the above is taken with respect to z^μ , we find that evaluation at $z^\nu = 0$ yields

$$\delta_\mu^\alpha = \frac{\partial w^\alpha}{\partial v^\gamma}(x^\beta, -x^\beta)B_\mu^\gamma(-x^\beta),$$

and therefore,

$$\frac{\partial w^\alpha}{\partial v^\gamma}(x^\beta, -x^\beta) = \hat{B}_\gamma^\alpha(-x^\beta). \quad (2.18)$$

The adjoint representation expressed in local coordinates is

$$(d\tau_x)_\mu^\alpha \equiv \frac{\partial(\tau_x u)^\alpha}{\partial u^\mu} \Big|_{u^\nu=0},$$

where

$$\tau_x u = xux^{-1}.$$

Thus, by definition,

$$(d\tau_x)_\mu^\alpha = \frac{\partial w^\alpha}{\partial u^\mu}(x^\beta, w^\beta(u^\gamma, -x^\gamma)) \Big|_{u^\nu=0},$$

which reduces to

$$(d\tau_x)_\mu^\alpha = \frac{\partial w^\alpha}{\partial v^\gamma}(x^\beta, -x^\beta)A_\mu^\gamma(-x^\beta).$$

By virtue of (2.18) and the definition of T_β^α we then have

$$(d\tau_x)_\mu^\alpha = \hat{T}_\mu^\alpha(x^\beta),$$

as required. Note that if w had been the right action of u on v then T_B^A would be a left action and T_β^α itself would be the adjoint representation. In either case, since w is merely a multiplication, T_β^α is actually defined and invertible in the entire domain of the canonical chart of the first kind and not just in a neighborhood of the identity as guaranteed by our local formalism.

For any given action w , we have

$$\tilde{T}_{\beta\gamma}^\alpha \equiv \frac{\partial T_\beta^\alpha}{\partial u^\gamma} \Big|_{u^\nu=0} = C_{\beta\gamma}^\alpha.$$

Thus, when the Lie group representation of type w T_B^A is T_β^α , the identity (2.15) reduces to the Jacobi identity, viz.,

$$C_{\nu\alpha}^\mu C_{\tau\gamma}^\nu + C_{\nu\gamma}^\mu C_{\alpha\tau}^\nu + C_{\nu\tau}^\mu C_{\gamma\alpha}^\nu = 0.$$

Therefore, if G admits a semisimple Lie group representation of type w , the $C_{\beta\gamma}^\alpha$'s are the structure constants for some Lie group.¹³

As in the case of $\tilde{C}_{B\tau}^{As}$ under a coordinate transformation, $\tilde{T}_{B\alpha}^A$ has more than one transformation law under a gauge transformation. It is invariant, i.e.,

$$\tilde{T}_{B\alpha}^A = \tilde{T}_{B\alpha}^A, \quad (2.19)$$

and yet, by combining (2.12), (2.13), and (2.19), we obtain

$$\tilde{T}_{B\alpha}^A = T_C^A \hat{T}_B^D \hat{T}_\alpha^\beta \tilde{T}_{D\beta}^C. \quad (2.20)$$

Similarly, $g_{\alpha\beta}$ is invariant, but also satisfies

$$\dot{g}_{\alpha\beta} = \hat{T}_\alpha^\gamma \hat{T}_\beta^\tau g_{\gamma\tau}.$$

Therefore, \hat{T}_α^γ leaves $g_{\alpha\beta}$ invariant in the sense that

$$g_{\alpha\beta} = \hat{T}_\alpha^\gamma g_{\gamma\tau} \hat{T}_\beta^\tau,$$

and thus, provided

$$\det g_{\alpha\beta} \neq 0,$$

we have that

$$(\det T_\beta^\alpha)^2 = 1.$$

However, we also know that

$$T_\beta^\alpha(0) = \delta_\beta^\alpha$$

and hence

$$\det T_\beta^\alpha = 1.$$

It is also possible to show that

$$C_{\alpha\beta\gamma} \equiv C_{\alpha\gamma}^\mu g_{\mu\beta}$$

is totally antisymmetric. Several additional identities associated with Lie groups can be generalized using these techniques.

3. DOUBLE COVARIANT DIFFERENTIATION

In order to arrive at the conclusions outlined in the introduction we shall first assume that coordinate and gauge transformations commute in the sense that

$$C_B^A T_D^B = T_B^A C_D^B. \quad (3.1)$$

The derivatives of (3.1) with respect to J_s^r and u^α evaluated at $J_b^a = \delta_b^a$ and $u^\nu = 0$, respectively, yield

$$\tilde{C}_{Br}^{As} T_D^B = T_B^A \tilde{C}_{Dr}^{Bs} \quad (3.2)$$

and

$$C_B^A \tilde{T}_{D\alpha}^B = \tilde{T}_{B\alpha}^A C_D^B, \quad (3.3)$$

while the "mixed" second partial derivative of (3.1) evaluated at both identity transformations is

$$\tilde{C}_{Br}^{As} \tilde{T}_{D\alpha}^B = \tilde{T}_{B\alpha}^A \tilde{C}_{Dr}^{Bs}. \quad (3.4)$$

One of the consequences of (3.2) is that the quantity

$$g_{r\alpha}^s \equiv \tilde{C}_{Br}^{As} \tilde{T}_{A\alpha}^B,$$

by virtue of (2.20), satisfies

$$g_{r\alpha}^s = \hat{T}_{\alpha}^{\beta} g_{r\beta}^s.$$

When the partial derivative of the above is taken with respect to u^γ and then evaluated at $u^\nu = 0$, it is found that

$$0 = g_{r\beta}^s C_{\alpha}^{\beta \gamma}.$$

Thus, provided G admits a semisimple Lie group representation of type w ,

$$g_{r\beta}^s \equiv 0. \quad (3.5)$$

We then turn to the quantity

$$g_{sj}^{ri} \equiv \tilde{C}_{Bs}^{Ar} \tilde{C}_{Aj}^{Bi}.$$

By virtue of its transformation properties under coordinate and gauge transformations, g_{sj}^{ri} must be of the form³

$$g_{sj}^{ri} = \alpha \delta_s^r \delta_j^i + \beta \delta_j^r \delta_s^i,$$

where α and β are arbitrary constants. Therefore if we are given an equation of the form

$$g_{sj}^{ri} \psi_j^i = 0,$$

for some quantity ψ_j^i , it would be possible to conclude that

$$\psi_j^i = 0, \quad (3.6)$$

provided both

$$n\alpha + \beta \neq 0$$

and

$$\beta \neq 0,$$

which, in general, will hold since α and β are arbitrary for unspecified ρ^A . In our analysis we shall also meet equations of the form

$$\tilde{C}_{Bj}^{Ai} \psi_j^i + \tilde{T}_{B\alpha}^A \psi^\alpha = 0,$$

where ψ_j^i and ψ^α are some given quantities. By virtue of (3.5) and (3.6) and the assumption that G admits a semisimple Lie group representation of type w , we will be able to conclude that

$$\psi_j^i = 0$$

together with

$$\psi^\alpha = 0.$$

We are now ready to demand that

$$\bar{\rho}_{||a}^A = C_B^A J_a^b \rho_{||b}^B \quad (3.7)$$

should hold for $\rho^A_{||a}$ given by

$$\rho^A_{||a} = \rho^A_{,a} - \tilde{C}_{Br}^{As} \Gamma_s^r \rho^B - \tilde{T}_{B\alpha}^A A_a^\alpha \rho^B. \quad (3.8)$$

Expansion of the left-hand side of (3.7) yields

$$\begin{aligned} \bar{\rho}_{||a}^A &= C_B^A J_a^b \rho_{,b}^B + C_{Br}^{As} J_{sa}^r \rho^B \\ &\quad - \tilde{C}_{Br}^{As} \bar{\Gamma}_s^r C_D^B \rho^D - \tilde{T}_{B\alpha}^A \bar{A}_a^\alpha C_D^B \rho^D, \end{aligned}$$

where

$$J_{sa}^r \equiv \frac{\partial J_s^r}{\partial x^a}.$$

By making use of (2.4), (2.8), and (3.3), we obtain

$$\begin{aligned} \bar{\rho}_{||a}^A &= C_B^A J_a^b [\rho_{||b}^B - \tilde{C}_{Dj}^{Bj} (\hat{J}_i^s J_i^r \bar{\Gamma}_s^r \hat{J}_b^c \\ &\quad - \hat{J}_i^s \hat{J}_b^c J_{sc}^j - \Gamma_{ib}^j) \rho^D \\ &\quad - \tilde{T}_{D\alpha}^B (\bar{A}_c^\alpha \hat{J}_b^c - A_b^\alpha) \rho^D]. \end{aligned}$$

Therefore, for arbitrary ρ^A , (3.7) is satisfied if and only if

$$\begin{aligned} \tilde{C}_{Dj}^{Bj} (\bar{\Gamma}_s^r J_i^s \hat{J}_i^r \hat{J}_b^c - J_{sc}^j \hat{J}_i^s \hat{J}_b^c - \Gamma_{ib}^j) \\ + \tilde{T}_{D\alpha}^B (\bar{A}_c^\alpha \hat{J}_b^c - A_b^\alpha) = 0. \end{aligned}$$

Thus, provided G admits a semisimple Lie group representation of type w , we must have

$$\bar{\Gamma}_s^r = \hat{J}_j^r J_s^i J_c^b \Gamma_{ib}^j + \hat{J}_j^r J_{sc}^j$$

and

$$\bar{A}_c^\alpha = J_c^b A_b^\alpha,$$

which states that Γ_s^r behaves like a linear connection and A_c^α behaves like a covariant vector field under a coordinate transformation.

When we demand that under a gauge transformation

$$\dot{\rho}_{||a}^A = T_B^A \rho_{||a}^B, \quad (3.9)$$

expansion of the left-hand side of (3.9) yields

$$\begin{aligned} \dot{\rho}_{||a}^A &= T_B^A \rho_{,a}^B + T_{B\alpha}^A u^\alpha \rho^B \\ &\quad - \tilde{C}_{Br}^{As} \dot{\Gamma}_s^r T_D^B \rho^D - \tilde{T}_{B\alpha}^A \dot{A}_a^\alpha T_D^B \rho^D. \end{aligned}$$

Equations (2.10), (3.2), and (2.20) enable us to arrive at

$$\begin{aligned} \dot{\rho}_{||a}^A &= T_B^A [\rho_{||a}^B - \tilde{C}_{Dr}^{Bs} (\dot{\Gamma}_s^r - \Gamma_s^r) \rho^D \\ &\quad - \tilde{T}_{D\beta}^B (\hat{T}_\alpha^\beta \dot{A}_a^\alpha - \hat{A}_\alpha^\beta u^\alpha - A_a^\beta) \rho^D]. \end{aligned}$$

Thus, for arbitrary ρ^A , (3.9) holds if and only if

$$\tilde{C}_{Dr}^{Bs} (\dot{\Gamma}_s^r - \Gamma_s^r) + \tilde{T}_{D\beta}^B (\hat{T}_\alpha^\beta \dot{A}_a^\alpha - \hat{A}_\alpha^\beta u^\alpha - A_a^\beta) = 0.$$

Therefore, provided G admits a semisimple Lie group representation of type w ,

$$\dot{\Gamma}_s^r = \Gamma_s^r$$

and

$$\dot{A}_a^\alpha = T_\beta^\alpha A_a^\beta + \hat{B}_\beta^\alpha u^\beta,$$

i.e., Γ_s^r is invariant while A_a^α behaves like a gauge connection under a gauge transformation if w were the left action of u on v , i.e., $w = uv$.⁹

Whenever a new process of differentiation is defined one is always curious to see what the commutator of two successive derivatives is. By the definition (3.8) we have

$$\begin{aligned} \rho^A_{\parallel jk} &\equiv \rho^A_{\parallel jk} \\ &= \rho^A_{\parallel j,k} - (\tilde{C}_{Br}^{As} \delta_j^h + \delta_{Br}^A \delta_j^s \delta_r^h) \Gamma_{s,r}^k \rho^B \\ &\quad - \tilde{T}_{Ba}^A A_{jk}^{\alpha} \rho^B, \end{aligned}$$

which can be expanded as

$$\begin{aligned} \rho^A_{\parallel jk} &= \rho^A_{jk} - \tilde{C}_{Br}^{As} \Gamma_{s,r}^k \rho^B - \tilde{C}_{Br}^{As} \Gamma_{s,r}^j \rho^B_{,k} \\ &\quad - \tilde{T}_{Ba}^A A_{jk}^{\alpha} \rho^B - \tilde{T}_{Ba}^A A_{jk}^{\alpha} \rho^B_{,k} \\ &\quad - \tilde{C}_{Br}^{As} \Gamma_{s,r}^k (\rho^B_{,j} - \tilde{C}_{Da}^{Bb} \Gamma_{b,j}^a \rho^D \\ &\quad - \tilde{T}_{Da}^B A_{jk}^{\alpha} \rho^D) - \tilde{T}_{Ba}^A A_{jk}^{\alpha} (\rho^B_{,j} - \tilde{C}_{Da}^{Bb} \Gamma_{b,j}^a \rho^D \\ &\quad - \tilde{T}_{D\beta}^B A_{jk}^{\beta} \rho^D) - \Gamma_{jk}^h \rho^A_{\parallel h}. \end{aligned}$$

Antisymmetrization of the above with respect to j and k , together with (3.4), leads to

$$\begin{aligned} \rho^A_{\parallel jk} - \rho^A_{\parallel kj} &= -\tilde{C}_{Br}^{As} R_{s,r}^k \rho^B - \tilde{T}_{Ba}^A F_j^{\alpha} \rho^B \\ &\quad - \Delta_j^h \rho^A_{\parallel h} + (\tilde{C}_{Br}^{As} \tilde{C}_{Da}^{Bb} - \tilde{C}_{Ba}^A \tilde{C}_{Dr}^{Bs}) \\ &\quad + \tilde{C}_{Dr}^{Ab} \delta_a^s - \tilde{C}_{Da}^{As} \delta_r^b) \Gamma_{s,r}^k \Gamma_{b,j}^a \rho^D \\ &\quad + (\tilde{T}_{Ba}^A \tilde{T}_{D\beta}^B - \tilde{T}_{B\beta}^A \tilde{T}_{Da}^B - C_{B\gamma}^{\alpha} \tilde{T}_{D\gamma}^A) A_j^{\beta} A_k^{\alpha} \rho^D, \end{aligned}$$

where $R_{s,r}^k$ is the Riemann curvature tensor, i.e.,

$$R_{s,r}^k \equiv \Gamma_{s,jk}^r - \Gamma_{s,rj}^k + \Gamma_{s,j}^b \Gamma_{b,r}^k - \Gamma_{s,r}^b \Gamma_{b,j}^k,$$

F_j^{α} is the gauge curvature, i.e.,

$$F_j^{\alpha} \equiv A_{j,k}^{\alpha} - A_{k,j}^{\alpha} - C_{\gamma}^{\alpha} A_j^{\beta} A_k^{\gamma}$$

and Δ_j^h is the torsion tensor, i.e.,

$$\Delta_j^h \equiv \Gamma_{j,k}^h - \Gamma_{k,j}^h.$$

However, by virtue of (2.6) and (2.15), we see that the commutator simplifies to

$$\rho^A_{\parallel jk} - \rho^A_{\parallel kj} = -R_{Bjk}^A \rho^B - \Delta_j^h \rho^A_{\parallel h},$$

where

$$R_{Bjk}^A \equiv \tilde{C}_{Br}^{As} R_{s,r}^k + \tilde{T}_{Ba}^A F_j^{\alpha} \rho^B.$$

It is interesting to note that if G admits a semisimple Lie group representation of type w then

$$R_{Bjk}^A = 0$$

implies that both

$$R_{s,r}^k = 0$$

and

$$F_j^{\alpha} = 0.$$

We can also see that the gauge curvature's transformation laws are

$$\bar{F}_j^{\alpha} = J_j^a J_k^b F_a^{\alpha}$$

under a coordinate transformation, and

$$\hat{F}_j^{\alpha} = T_{\beta}^{\alpha} F_j^{\beta}$$

under a gauge transformation. It is a simple matter to show that the gauge curvature satisfies the cyclic identity

$$\begin{aligned} F_j^{\alpha} \rho^{\alpha}_{\parallel h} + F_h^{\alpha} \rho^{\alpha}_{\parallel j} + F_k^{\alpha} \rho^{\alpha}_{\parallel h} \\ = \Delta_j^i F_i^{\alpha} \rho^{\alpha}_{\parallel h} + \Delta_h^i F_i^{\alpha} \rho^{\alpha}_{\parallel j} + \Delta_k^i F_i^{\alpha} \rho^{\alpha}_{\parallel h}, \end{aligned}$$

which is an obvious counterpart of the Bianchi identity.

4. INVARIANCE IDENTITIES

Under a coordinate transformation the arguments of the Lagrangian

$$L = L(\rho^A, \rho^A_{,a}; \rho^A_{,ab}; \Gamma_{b,c}^a; \Gamma_{b,c,d}^a; A_{a,b}^{\alpha}; A_{a,b}^{\alpha}) \quad (4.1)$$

transform as

$$\begin{aligned} \bar{\rho}^A &= C_B^A \rho^B, \\ \bar{\rho}^A_{,a} &= C_B^A J_a^i \rho^B_{,i} + C_{Br}^{As} J_{sa}^r \rho^B, \\ \bar{\rho}^A_{,ab} &= C_B^A J_a^i J_b^j \rho^B_{,ij} + C_B^A J_{ab}^i \rho^B_{,i} \\ &\quad + C_{Br}^{As} J_{sb}^r J_a^i \rho^B_{,i} + C_{Br}^{As} J_{sa}^r J_b^i \rho^B_{,i} \\ &\quad + C_{Br}^{As} J_{sab}^r \rho^B + C_{Brt}^{Asu} J_{ub}^t J_{sa}^r \rho^B, \\ \bar{\Gamma}_{b,c}^a &= \hat{J}_r^a (J_b^s J_c^t \Gamma_{st}^r + J_{bc}^r), \\ \bar{\Gamma}_{b,c,d}^a &= -\hat{J}_i^a \hat{J}_r^i J_{jd}^r (J_b^s J_c^t \Gamma_{st}^r + J_{bc}^r) \\ &\quad + \hat{J}_r^a (J_{bd}^s J_c^t \Gamma_{st}^r + J_b^s J_{cd}^t \Gamma_{st}^r \\ &\quad + J_b^s J_c^t J_d^u \Gamma_{s,t,u}^r + J_{bcd}^r), \\ \bar{A}_a^{\alpha} &= J_a^i A_i^{\alpha}, \end{aligned}$$

and

$$\bar{A}_{a,b}^{\alpha} = J_a^i J_b^j A_{ij}^{\alpha} + J_{ab}^i A_i^{\alpha},$$

where

$$J_{sab}^r \equiv \frac{\partial J_{sa}^r}{\partial \bar{x}^b} = \frac{\partial^2 J_s^r}{\partial \bar{x}^a \partial \bar{x}^b}$$

and

$$C_{Brt}^{Asu} \equiv \frac{\partial C_{Br}^{As}}{\partial J_u^t} = \frac{\partial^2 C_B^A}{\partial J_s^t \partial J_u^t}.$$

Therefore, when the derivatives of the condition

$$\bar{L} = JL$$

are taken with respect to J_{stu}^r , J_{st}^r , and J_s^r , and then evaluated at the identity transformation $J_j^i = \delta_j^i$, we obtain the invariance identities

$$\begin{aligned} \frac{1}{3} \left(\frac{\partial L}{\partial \rho^A_{,st}} \tilde{C}_{Br}^{Au} \rho^B + \frac{\partial L}{\partial \rho^A_{,tu}} \tilde{C}_{Br}^{As} \rho^B + \frac{\partial L}{\partial \rho^A_{,us}} \tilde{C}_{Br}^{At} \rho^B \right) \\ + \frac{\partial L}{\partial \Gamma_{(s,t),u}^r} = 0, \end{aligned} \quad (4.2)$$

$$\begin{aligned} \frac{1}{2} \left(\frac{\partial L}{\partial \rho^A_{,s}} \tilde{C}_{Br}^{At} \rho^B + \frac{\partial L}{\partial \rho^A_{,t}} \tilde{C}_{Br}^{As} \rho^B \right) + \frac{\partial L}{\partial \rho^A_{,as}} \tilde{C}_{Br}^{At} \rho^B_{,a} \\ + \frac{\partial L}{\partial \rho^A_{,at}} \tilde{C}_{Br}^{As} \rho^B_{,a} + \frac{\partial L}{\partial \rho^A_{,st}} \rho^A_{,r} + \frac{\partial L}{\partial \Gamma_{(s,t)}^r} \\ + \frac{1}{2} \left(-\frac{\partial L}{\partial \Gamma_{b,c,s}^r} \Gamma_{b,c}^t - \frac{\partial L}{\partial \Gamma_{a,b,t}^r} \Gamma_{a,b}^s + \frac{\partial L}{\partial \Gamma_{s,c,t}^a} \Gamma_r^a \right. \\ \left. + \frac{\partial L}{\partial \Gamma_{t,c,s}^a} \Gamma_r^a \right) + \frac{\partial L}{\partial \Gamma_{b^a(s,t)}^r} \Gamma_{b^a,r} + \frac{\partial L}{\partial A_{(s,t)}^{\alpha}} A_r^{\alpha} = 0, \end{aligned} \quad (4.3)$$

and

$$\frac{\partial L}{\partial \rho^A} \tilde{C}_{Br}^{As} \rho^B + \frac{\partial L}{\partial \rho^A_{,a}} \tilde{C}_{Br}^{As} \rho^B_{,a} + \frac{\partial L}{\partial \rho^A_{,s}} \rho^A_{,r}$$

$$\begin{aligned}
& + \frac{\partial L}{\partial \rho^A_{,ab}} \tilde{C}_{Br}^{As} \rho^B_{,ab} + 2 \frac{\partial L}{\partial \rho^A_{,sb}} \rho^A_{,rb} - \frac{\partial L}{\partial \Gamma_{b'c}} \Gamma_{b'c}^s \\
& + \frac{\partial L}{\partial \Gamma_{s'c}^a} \Gamma_{r'c}^a + \frac{\partial L}{\partial \Gamma_{b's}^a} \Gamma_{b'a}^r - \frac{\partial L}{\partial \Gamma_{b'c,d}} \Gamma_{b'c,d}^s \\
& + \frac{\partial L}{\partial \Gamma_{s'c,d}^a} \Gamma_{r'c,d}^a + \frac{\partial L}{\partial \Gamma_{b's,d}^a} \Gamma_{b'a}^r + \frac{\partial L}{\partial \Gamma_{b'c,s}^a} \Gamma_{b'a}^r \\
& + \frac{\partial L}{\partial A_s^\alpha} A_r^\alpha + \frac{\partial L}{\partial A_{s,b}^\alpha} A_{r,b}^\alpha + \frac{\partial L}{\partial A_{a,s}^\alpha} A_{a,r}^\alpha = \delta_r^s L, \quad (4.4)
\end{aligned}$$

respectively.

When the arguments of the Lagrangian (4.1) undergo a gauge transformation, we find that

$$\begin{aligned}
\dot{\rho}^A &= T_B^A \rho^B, \\
\dot{\rho}^A_{,a} &= T_B^A \rho^B_{,a} + T_{B\alpha}^A u^\alpha_{,a} \rho^B, \\
\dot{\rho}^A_{,ab} &= T_B^A \rho^B_{,ab} + T_{B\alpha}^A u^\alpha_{,b} \rho^B_{,a} + T_{B\alpha}^A u^\alpha_{,a} \rho^B_{,b} \\
&\quad + T_{B\alpha}^A u^\alpha_{,ab} \rho^B + T_{B\alpha\beta}^A u^\beta_{,b} u^\alpha_{,a} \rho^B, \\
\dot{\Gamma}_{b'c}^a &= \Gamma_{b'c}^a, \\
\dot{\Gamma}_{b'c,d}^a &= \Gamma_{b'c,d}^a, \\
\dot{A}_a^\alpha &= T_{\beta a}^\alpha A_\beta^\alpha + \hat{B}_{\beta a}^\alpha u^\beta_{,a},
\end{aligned}$$

and

$$\begin{aligned}
\dot{A}_{a,b}^\alpha &= T_{\beta a}^\alpha A_{a,b}^\beta + T_{\beta\gamma}^\alpha u^\gamma_{,b} A_a^\beta \\
&\quad + \hat{B}_{\beta a}^\alpha u^\beta_{,ab} - \hat{B}_{\gamma a}^\alpha \hat{B}_{\beta b}^\mu B_{\mu\nu}^\gamma u^\nu_{,b} u^\beta_{,a},
\end{aligned}$$

where

$$T_{B\alpha\beta}^A \equiv \frac{\partial T_{B\alpha}^A}{\partial u^\beta} = \frac{\partial^2 T_B^A}{\partial u^\alpha \partial u^\beta}$$

and

$$B_{\mu\nu}^\gamma \equiv \frac{\partial B_{\mu\nu}^\gamma}{\partial u^\nu}.$$

Thus, the invariance identities obtained by taking the derivatives of the condition

$$\dot{L} = L$$

with respect to $u^{\beta}_{,rs}$, $u^{\beta}_{,r}$ and u^β , and then evaluating at the identity transformation $u^\nu = 0$, reduce to

$$\frac{\partial L}{\partial \rho^A_{,rs}} \tilde{T}_{B\beta}^A \rho^B + \frac{\partial L}{\partial A_{(r,s)}^\beta} = 0, \quad (4.5)$$

$$\begin{aligned}
& \frac{\partial L}{\partial \rho^A_{,r}} \tilde{T}_{B\beta}^A \rho^B + 2 \frac{\partial L}{\partial \rho^A_{,rb}} \tilde{T}_{B\beta}^A \rho^B_{,b} \\
& + \frac{\partial L}{\partial A_r^\beta} + \frac{\partial L}{\partial A_{a,r}^\alpha} C_{\gamma\beta}^\alpha A_a^\gamma = 0, \quad (4.6)
\end{aligned}$$

and

$$\begin{aligned}
& \frac{\partial L}{\partial \rho^A} \tilde{T}_{B\beta}^A \rho^B + \frac{\partial L}{\partial \rho^A_{,a}} \tilde{T}_{B\beta}^A \rho^B_{,a} + \frac{\partial L}{\partial \rho^A_{,ab}} \tilde{T}_{B\beta}^A \rho^B_{,ab} \quad (4.7) \\
& + \frac{\partial L}{\partial A_a^\alpha} C_{\gamma\beta}^\alpha A_a^\gamma + \frac{\partial L}{\partial A_{a,b}^\alpha} C_{\gamma\beta}^\alpha A_{a,b}^\gamma = 0,
\end{aligned}$$

respectively.

It is also possible to establish that the Euler–Lagrange expressions

$$\begin{aligned}
E_A &\equiv \frac{\partial L}{\partial \rho^A} - \frac{\partial}{\partial x^a} \left(\frac{\partial L}{\partial \rho^A_{,a}} \right) \\
&\quad + \frac{\partial^2}{\partial x^a \partial x^b} \left(\frac{\partial L}{\partial \rho^A_{,ab}} \right),
\end{aligned}$$

$$E_a^{bc} \equiv \frac{\partial L}{\partial \Gamma_{b'c}^a} - \frac{\partial}{\partial x^d} \left(\frac{\partial L}{\partial \Gamma_{b'c,d}^a} \right),$$

and

$$E_\alpha^a \equiv \frac{\partial L}{\partial A_a^\alpha} - \frac{\partial}{\partial x^b} \left(\frac{\partial L}{\partial A_{a,b}^\alpha} \right)$$

obey the transformation laws

$$\begin{aligned}
\bar{E}_A &= \hat{J} C_A^B E_B, \\
\bar{E}_A &= \hat{T}_A^B E_B, \\
\bar{E}_a^{bc} &= J \hat{J}_j^b \hat{J}_k^c J_a^i E^{jk}, \\
\bar{E}_a^{bc} &= E_a^{bc}, \\
\bar{E}_\alpha^a &= J \hat{J}_i^a E_\alpha^i,
\end{aligned}$$

and

$$\bar{E}_\alpha^a = \hat{T}_\alpha^\beta E_\beta^a.$$

Both sets of invariance identities yield “conservative laws” involving the various Euler–Lagrange expressions. When (4.5) and (4.6) are substituted into (4.7), (4.7) reduces to

$$E_A \tilde{T}_{B\beta}^A \rho^B - E_{\beta||a}^a + \Delta_a{}^b E_\beta^a = 0. \quad (4.8)$$

By taking the derivative of (4.4) with respect to x^s and substituting (4.2), (4.3), and (4.8), it is found that

$$\begin{aligned}
& (E_A \tilde{C}_{Br}^{As} \rho^B)_{||s} + \Delta_r{}^a E_A \tilde{C}_{Ba}^{As} \rho^B \\
& - E_A \rho^A_{||r} + E_\alpha^s F_{r's}^\alpha - E_r^{st}{}_{||ts} + E_a^{bs} \Delta_b{}^a \\
& + E_a^{bs} R_b{}^a{}_{rs} - \Delta_s{}^a (E_A \tilde{C}_{Br}^{As} \rho^B - 2 E_r^{(st)}{}_{||t}) \\
& + \Delta_t{}^b E_r^{st} + \Delta_d{}^c E_c^{ds} - \Delta_t{}^a{}_{||s} E_r^{st} = 0.
\end{aligned}$$

5. DISCUSSION

One of the features of the local gauge field theory formalism presented here is that we did not restrict ourselves to infinitesimal gauge transformations. It is felt that much more insight is gained with the use of full transformations, particularly as far as fiber bundles are concerned.

In the future, a concomitant approach using this formalism should be able to successfully motivate the field equations of specific theories. Attempts have been made¹⁴ along these lines; however, more exploitation of our formalism is needed.

ACKNOWLEDGMENTS

I would like to thank Dr. G. W. Horndeski for many fruitful discussions on the material in this paper and express my appreciation to the referee for his suggestions. I would also like to thank the Natural Sciences and Engineering Research Council of Canada for their financial support.

¹C. N. Yang and R. L. Mills, Phys. Rev. **96**, 191 (1954).

²H. Rund, Abh. math. Sem. Univ. Hamburg **29**, 243 (1966).

³D. Lovelock, Arch. Rat. Mech. Anal. **33**, 54 (1969).

⁴D. Lovelock, J. Aust. Math. Soc. **14**, 482 (1972); H. Rund, Tensor N. S. **18**, 239 (1967).

- ⁵R. Utiyama, *Phys. Rev.* **101**, 1597 (1956).
- ⁶T. W. B. Kibble, *J. Math. Phys.* **2**, 212 (1961).
- ⁷C. N. Yang, *Gauge Fields. Proc. Sixth Hawaii Topical Conf. Particle Physics*, edited by P. N. Dobson, Jr., S. Paksava, V. Z. Peterson, and S. F. Tuan, (University of Hawaii at Manoa, Honolulu, 1975).
- ⁸H. Rund, *Tensor N. S.* **33**, 97 (1979).
- ⁹W. Drechsler and M. E. Mayer, *Fiber Bundle Techniques in Gauge Theories* (Springer, New York, 1975).
- ¹⁰I. M. Anderson, *Gen. Rel. Grav.* **10**, 461 (1979); G. W. Horndeski, *Tensor N. S.* **28**, 303 (1974); D. Lovelock, *Aequationes Math.* **4**, 127 (1970); D. Lovelock, *Gen. Rel. Grav.* **5**, 399 (1974).
- ¹¹L. S. Pontryagin, *Topological Groups* (Gordon and Breach, New York, 1966), pp. 290, 418.
- ¹²R. J. McKellar, unpublished Ph.D. dissertation, University of Arizona (1978).
- ¹³L. S. Pontryagin, *Topological Groups* (Gordon and Breach, New York, 1966), p. 412.
- ¹⁴J. Anandan and R. Roskies, *J. Math. Phys.* **19**, 2614 (1978); Y. M. Cho, *Phys. Rev. D* **14**, 3335 (1976); N.C.T. Coote and A. J. Macfarlane, *Gen. Rel. Grav.* **9**, 621 (1978); R. Roskies, *Phys. Rev. D* **15**, 1722 (1977).

The role of noncompact Lie algebras in relativistic wave equations. I

A. Cant

Department of Theoretical Physics, University of St. Andrews, St. Andrews, Fife, Scotland

(Received 5 December 1979; accepted for publication 20 March 1980)

The role of real Lie algebras in the study of relativistic wave equations of the form $(\alpha^\mu \partial_\mu + i\kappa)\psi(x) = 0$ is considered. To a finite-dimensional equation there corresponds a Lie algebra S containing $\text{so}(4, \mathbb{C})$ and a vector operator $\{\alpha^\mu\}$. The importance of finding all possible real forms of S containing the Lorentz Lie algebra $\text{so}(3, 1)$ is discussed. This problem is solved in detail for certain "generic" cases, namely $S = \text{sp}(n, \mathbb{C})$, $\text{so}(n, \mathbb{C})$, and $\text{sl}(n, \mathbb{C})$. The exceptional algebras G_2 , F_4 , and E_6 are also considered.

PACS numbers: 11.10.Qr, 03.65.Fd, 02.20.Sv

1. INTRODUCTION

A previous paper (Cant and Hurst¹) considered some of the Lie algebraic properties of finite-dimensional Lorentz invariant wave equations of the form

$$\left(\alpha^\mu \frac{\partial}{\partial x^\mu} + i\kappa\right)\psi(x) = 0, \quad (1.1)$$

where α^μ ($\mu = 0, 1, 2, 3$) are $n \times n$ matrices and κ is a real nonzero constant. In this paper and the next, which are developments of Ref. 1, we examine the role played by real Lie algebras in the theory.

Before we state the problem, we need to recall¹ some of the basic properties of (1.1). Our notation is as in Ref. 1; for real Lie algebras we use the notation of Helgason.² All representations are *finite*-dimensional, unless otherwise stated.

If A belongs to the group \mathcal{L} of proper orthochronous Lorentz transformations, we have

$$\psi'(x') = \pi(A)\psi(x) \quad (x' = Ax), \quad (1.2)$$

and π is a representation of \mathcal{L} . The generators $I_{\mu\nu}$ of π satisfy

$$[I_{\mu\nu}, I_{\rho\sigma}] = g_{\nu\rho}I_{\mu\sigma} - g_{\mu\rho}I_{\nu\sigma} - g_{\nu\sigma}I_{\mu\rho} + g_{\mu\sigma}I_{\nu\rho}, \quad (1.3)$$

while the invariance condition is

$$[I_{\mu\nu}, \alpha_\rho] = g_{\nu\rho}\alpha_\mu - g_{\mu\rho}\alpha_\nu. \quad (1.4)$$

We say that $\{\alpha^\mu\}$ is a "vector operator" if (1.4) is satisfied.

Thus any wave equation of the form (1.1) is specified by giving a representation³ (π, V) of the Lorentz Lie algebra⁴ $\text{so}(3, 1) \cong \text{sl}(2, \mathbb{C})^R$ which admits a vector operator $\{\alpha^\mu\}$, and fixing such a vector operator. As is well-known, the representation π of $\text{sl}(2, \mathbb{C})^R$ extends to a unique representation (also denoted by π) of its complexification $\text{so}(4, \mathbb{C}) \cong \text{sl}(2, \mathbb{C}) \oplus \text{sl}(2, \mathbb{C})$ (or $D_2 = A_1 \oplus A_1$) which also acts on V . Let S denote the Lie algebra generated by $\pi(D_2)$ and the α^μ over \mathbb{C} . Then, if (ρ, W) is any representation of S , we can take the vector operator $\rho(\alpha^\mu)$, and the representation of D_2 [and thus $\text{sl}(2, \mathbb{C})^R$] obtained by finding the branching rules for the restriction of ρ to D_2 . This gives a new invariant wave equation. We therefore obtain a family of wave equations based on the initial equation by letting ρ go over all the irreducible representations of S .

This procedure was discussed in Ref. 1, and S was calculated for certain classes of equations. Our main point was

that the Bhabha⁵ case $S = \text{so}(5, \mathbb{C})$, corresponding to the situation $[\alpha^\mu, \alpha^\nu] = cI^{\mu\nu}$ ($c \in \mathbb{C}$, $c \neq 0$), is not the only one; in fact, S can be of arbitrarily large dimension. We considered, in particular, the Kursunoglu equation, for which $S = \text{sp}(12, \mathbb{C})$, and calculated some branching rules for $S \rightarrow D_2$, and the general form of the mass spectra.

The above procedure is valid whether or not there exists a real form S_0 of S which contains $\text{sl}(2, \mathbb{C})^R$, and it remains valid if ρ is allowed to be infinite-dimensional, provided that the representation of $\text{sl}(2, \mathbb{C})^R$ thus obtained is integrable to a representation of the group $\text{SL}(2, \mathbb{C})$.

Nevertheless, the problem of finding all such real forms is an important one. One finds that the existence of such a real form is closely related to the existence of operators corresponding to space reflection and charge conjugation; furthermore, invariance of (1.1) under these transformations leads, in most cases, to a distinguished real form. Also, if a real form S_0 does exist, then we have an embedding of the corresponding Lie groups: $\text{SL}(2, \mathbb{C}) \subset \mathcal{S}_0$. We can then directly consider representations ρ of \mathcal{S}_0 as providing new invariant wave equations; in this situation, since we have an embedding on the group level, the decomposition of ρ into irreducible representations of $\text{SL}(2, \mathbb{C})$ may be easier to find.

The present paper is devoted to finding those real Lie algebras, containing the Lorentz Lie algebra, which are of relevance for wave equations. In Sec. 2 we give notation and write down the necessary results of Ref. 1 in a general form. In Sec. 3 we find explicitly all the real forms S_0 of S for certain "generic" algebras S : namely, $S = \text{sp}(n, \mathbb{C})$, $\text{so}(n, \mathbb{C})$, $\text{sl}(n, \mathbb{C})$. We also discuss the exceptional Lie algebras G_2 , E_4 , and E_6 .

In a forthcoming paper we shall consider the connection with parity and charge conjugation, and briefly discuss the formation of infinite-dimensional wave equations.

2. PRELIMINARIES

We shall write the representation (π, V) of D_2 as¹

$$(\pi, V) = \left(\bigoplus_{r=1}^l \pi_r, \bigoplus_{r=1}^l V_r \right), \quad (2.1)$$

where π_r denotes the irreducible representation (k_r, l_r) of D_2 , with dimension $(2k_r + 1)(2l_r + 1)$. It is important to note that a given irreducible representation (k, l) can occur more

than once in π . Because of this, it turns out to be more useful to write

$$(\pi, V) = \left(\bigoplus_{j=1}^k \psi_j, \bigoplus_{j=1}^k Y_j \right), \quad (2.2)$$

where ψ_j is the direct sum of n_j copies of (k_j, l_j) . We shall use the labels $r, s = 1, \dots, t$ to refer to the *irreducible* subrepresentations of V , and $i, j = 1, \dots, k$ to refer to subrepresentations consisting of several copies of a single irreducible representation.

As discussed in Ref. 1, when V is an indecomposable S -module, π specifies an embedding of D_2 in the orthogonal algebra $\text{so}(V) \cong \text{so}(n, \mathbb{C})$ or the symplectic algebra $\text{sp}(V) \cong \text{sp}(n, \mathbb{C})$. We shall need the most general possibilities.

We always have the embeddings

$$D_2 \subset \bigoplus_{i=1}^k \text{so}(Y_i) \subset \text{so}(V) \quad (\rho = 1), \quad (2.3)$$

$$D_2 \subset \bigoplus_{i=1}^k \text{sp}(Y_i) \subset \text{sp}(V) \quad (\rho = -1),$$

relative to the bilinear form with matrix B given in terms of the decomposition (2.2) by⁶

$$B = \bigoplus_{i=1}^k (\Delta_i \otimes B_i), \quad (2.4)$$

where the $\Delta_i \in \text{GL}(n_i, \mathbb{C})$ satisfy $\Delta_i^T = \Delta_i$ ($i = 1, \dots, k$), and B_i is the matrix of the canonical form b_i defined on (k_i, l_i) which was introduced in Ref. 1. In (2.3) $\rho = +1$ (-1) according as the spin is integral (half-integral); it is clear that $B^T = \rho B$.

The embeddings (2.3) are an obvious generalization of (3.2) in Ref. 1.

However, if it happens that each irreducible representation π_r of D_2 occurs an *even* number of times, i.e., if n_i is even, for $i = 1, \dots, k$, then, as well as the above, we may choose the Δ_i such that $\Delta_i^T = -\Delta_i$ ($i = 1, \dots, k$). We then have $B^T = -\rho B$ and the embeddings

$$D_2 \subset \bigoplus_{i=1}^k \text{sp}(Y_i) \subset \text{sp}(V) \quad (\rho = 1), \quad (2.5)$$

$$D_2 \subset \bigoplus_{i=1}^k \text{so}(Y_i) \subset \text{so}(V) \quad (\rho = -1).$$

This possibility was noted in Ref. 1, but not pursued.

Given a vector operator $\{\alpha^\mu\}$, one can often choose the Δ_i in B such that $S \subseteq \text{so}(V)$ [$\text{sp}(V)$] [see Theorem 3.3 and the remark after (4.34) in Ref. 1]. We have also seen by means of examples in Sec. 3 of Ref. 1 that S is "almost always" equal to $\text{so}(V)$ [$\text{sp}(V)$] if such a B exists, and "almost always" equal to $\text{sl}(V)$ otherwise. We shall refer to these as generic cases, since the collection of families based on these algebras exhausts all finite-dimensional wave equations.

We shall keep the same explicit formulas for α^μ as in Ref. 1: α^μ splits via (2.2) into super matrix blocks $[i | \alpha^\mu | j]$, each involving the Kronecker product of a coupling matrix A_{ij} with a known combination of Dirac spinor matrices.

3. REAL FORMS

A. General results

We begin by making some observations on the general

problem of embeddings of real Lie algebras.

It is well known² that all possible real forms L_0 of a semisimple Lie algebra L over \mathbb{C} are obtained as follows. We find all involutive automorphisms s of the compact real form U of L [without distinguishing automorphisms conjugate within the group $\text{Aut}(U)$ of automorphisms of U]. Writing $U = K \oplus P$, where K and P are the eigenspaces of s corresponding to eigenvalues $+1$ and -1 , we take $L_0 = K \oplus iP$; L_0 is a (noncompact) real form of L , and this procedure gives every real form.

If L' , L are semisimple Lie algebras over \mathbb{C} , and U' , U are compact real forms, then Mal'cev⁷ has shown that L' can be embedded in L if and only if U' can be embedded in U . Suppose L'_0 and L_0 are real forms of L' and L , corresponding to the involutive automorphisms s' , s of U' , U . If U' can be embedded in U , then there is no guarantee that L'_0 can be embedded in L_0 . Also, if we have two embeddings of U' in U which are not conjugate within the group $\text{Int}(U)$ of inner automorphisms, then it is possible for L'_0 to be embedded in L_0 in one case, but not the other. The general problem of embeddings of real Lie algebras has been extensively discussed by Cornwell⁸⁻¹⁰ and Ekins and Cornwell.^{11,12} We shall use a slightly different version of the main theorem (34) of Ref. 8. We give a proof here, since the method of proof differs from that of Ref. 8.

Theorem 3.1: With the above notation, if U' is a subalgebra of U , then L'_0 is a subalgebra of L_0 if and only if s is an extension of s' , i.e.,

$$s(x') = s'(x') \quad (\forall x' \in U').$$

Proof: Write $L'_0 = K' \oplus iP'$, $L_0 = K \oplus iP$, where $U' = K' \oplus P'$ and $U = K \oplus P$. If σ and τ are the conjugations of L with respect to L_0 and U , then it is easy to see that σ commutes with τ . By the remark on p. 155 in Ref. 2, we have $K = L_0 \cap U$, $P = iL_0 \cap U$. Similarly $K' = L'_0 \cap U'$, $P' = iL'_0 \cap U'$.

Suppose $L'_0 \subset L_0$. Then

$$K' = L'_0 \cap U' \subset L_0 \cap U = K,$$

$$P' = iL'_0 \cap U' \subset iL_0 \cap U = P,$$

and so $s(x' + y') = x' - y' = s'(x' + y')$ ($x' \in K'$, $y' \in P'$). Thus s is an extension of s' . On the other hand, s extends s' means that $K' \subset K$, $P' \subset P$ and so $L'_0 \subset L_0$. \square

We shall make direct use of this theorem in the following situation. Suppose we are given a wave equation of the form (1.1), i.e., a representation (π, V) of D_2 which admits a vector operator α^μ . We know "how many" vector operators exist by Proposition 3.1 in Ref. 1. Fix α^μ , and let $S \subseteq \text{sl}(n, \mathbb{C})$ be the Lie algebra generated by $\pi(D_2)$ and the α^μ over \mathbb{C} . We shall assume¹ that S is irreducible, and thus semisimple. In order to find the real forms S_0 of S which contain $\text{sl}(2, \mathbb{C})^R$, we can apply Theorem 3.1.

We have the embedding

$$\text{so}(4, \mathbb{C}) \cong \text{sl}(2, \mathbb{C}) \oplus \text{sl}(2, \mathbb{C}) \subset S \subset \text{sl}(n, \mathbb{C}),$$

with a corresponding embedding of the compact real forms

$$\text{su}(2) \oplus \text{su}(2) \subset U \subset \text{su}(n),$$

where it is convenient to take $U = [S \cap \text{su}(n)]^R$ as the com-

compact real form of S . Let $S_0 = K \oplus iP$ be a real form of S , corresponding to the involutive automorphism s of U . It is well known² that $\mathfrak{sl}(2, \mathbb{C})^R$ arises from the involutive automorphism s' of $\mathfrak{su}(2) \oplus \mathfrak{su}(2)$ which sends (x, y) to (y, x) . By Theorem 3.1, $\mathfrak{sl}(2, \mathbb{C})^R$ is a subalgebra of S_0 if and only if s is an extension of s'

$$s(x, y) = (y, x) \quad [\forall (x, y) \in \mathfrak{su}(2) \oplus \mathfrak{su}(2)].$$

In practice, since $\text{Aut}(U)$ is known, we single out those automorphisms which are involutive extensions of s' [without distinguishing automorphisms conjugate within $\text{Aut}(U)$].

Clearly such a real form S_0 exists if and only if $V = \bigoplus_{r=1}^k V_r$, $[V_r = (k_r, l_r)]$ is isomorphic as a D_2 module to its conjugate $\bar{V} = \bigoplus_{r=1}^k \bar{V}_r$, $[\bar{V}_r = (l_r, k_r)]$. Thus the representations (k, l) and (l, k) occur with equal multiplicity in π .¹¹

What happens to the vector operator α^μ in the process of going to real forms? First of all, if α^0 is Hermitian

($\alpha^{0\dagger} = \alpha^{0T} = \alpha^0$), then since $\alpha^j = [\alpha^0, I^{0j}]$, it follows that $\alpha^{j\dagger} = -\alpha^j$, for $j = 1, 2, 3$. Thus $i\alpha^0, \alpha^j \in \mathfrak{su}(n)$ and hence $i\alpha^0, \alpha^j \in U$. Since $U = K \oplus P$ we can write (uniquely)

$$\alpha^\mu = k^\mu + p^\mu,$$

where $ik^0, k^j \in K; ip^0, p^j \in P$. Clearly we have

$$k^j = [p^0, I^{0j}], \quad p^j = [k^0, I^{0j}],$$

and S_0 contains the elements $i(\alpha^0)^\sim, (\alpha^j)^\sim$, where

$$(\alpha^\mu)^\sim = k^\mu + ip^\mu.$$

Thus, in general, we cannot expect S_0 to contain α^μ . However, this is of no importance, since in any representation ρ of S_0 , we can recover $\rho(\alpha^\mu)$ from $\rho[(\alpha^\mu)^\sim]$. The situation is simpler if $i\alpha^0 \in K$ ($i\alpha^0 \in P$), for then $\alpha^j \in P$ ($\alpha^j \in K$) and so $i\alpha^0 \in S_0$ ($\alpha^\mu \in S_0$).

If α^0 is not necessarily Hermitian, then since $S^R \cong U \oplus iU$, we can (uniquely) express α^μ in terms of two vector operators β^μ, γ^μ

$$\alpha^\mu = \beta^\mu + i\gamma^\mu,$$

where $i\beta^0, \beta^j, i\gamma^0, \gamma^j \in U$. We then apply the above procedure to β^μ and γ^μ separately; in any representation ρ of S_0 we can recover $\rho(\alpha^\mu)$ from $\rho[(\beta^\mu)^\sim]$ and $\rho[(\gamma^\mu)^\sim]$.

Since S is in general not known, and is hard to calculate, we cannot hope to solve this problem completely. However, using the above procedure, we can find the real forms explicitly for the generic cases $S = \mathfrak{sp}(n, \mathbb{C})$, $\mathfrak{so}(n, \mathbb{C})$, $\mathfrak{sl}(n, \mathbb{C})$ described in Sec. 2, and we do this in Secs. 3 B–3 D. We make continual use of the theorem, mentioned by Helgason² (p. 339), which asserts that a simple Lie algebra over \mathbb{R} is determined by its complexification and the structure of a maximal compact subalgebra.

For the exceptional Lie algebras G_2, F_4 , and E_6 , Ekins and Cornwell have explicitly described all the real forms containing $\mathfrak{sl}(2, \mathbb{C})^R$. In Sec. 3 E we shall indicate which cases correspond to wave equations by computing the number of linearly independent vector operators present.

B. real forms of $\mathfrak{sp}(n, \mathbb{C})$

(1) We consider first the embedding (2.3)

$$D_2 \subset \bigoplus_{i=1}^k \mathfrak{sp}(Y_i) \subset \mathfrak{sp}(V) \quad (\rho = -1), \quad (3.1)$$

relative to the antisymmetric bilinear form (2.4), which is assumed to be chosen such that the α^μ are skew relative to B . We take (π, V) to be of the form (2.2) with $n_i = n_{\bar{i}}, \forall i$,¹³ so that an extension of s' exists.

The compact real form of $\mathfrak{sp}(n, \mathbb{C})$ is $\mathfrak{usp}(n) = [\mathfrak{sp}(n, \mathbb{C}) \cap \mathfrak{su}(n)]^R$. It is well known that all the automorphisms s of $\mathfrak{usp}(n)$ are inner, i.e.,

$$s: \mathfrak{usp}(n) \rightarrow \mathfrak{usp}(n),$$

$$X \rightarrow MXM^{-1},$$

for some M in the corresponding Lie group $\text{USp}(n)$ (i.e., $M^\dagger M = I, M^T B M = B$). By Schur's lemma, s is involutive ($s^2 = I$) if and only if $M^2 = cI$ ($c \in \mathbb{C}$). We then have $(M^2)^T B M^2 = M^T (M^T B M) M = M^T B M = B$, and so $c^2 = 1$, i.e., $c = \pm 1$.

Let us write down explicitly all the involutive extensions s of s' . Clearly s is an extension of s' if and only if

$$MK_3 M^{-1} = L_3, \quad MK_\pm M^{-1} = L_\pm \quad (3.2)$$

(K_3, K_\pm, L_3, L_\pm are the canonical generators¹ of π). From (3.2) we find that the matrix blocks of M are of the form

$$[i|M|j] = \delta_{\bar{i}j} M(i) \otimes G_i, \quad (3.3)$$

where $M(i) \in \text{GL}(n_i, \mathbb{C})$, and G_i is the $d_i \times d_i$ matrix [$d_i = \dim(k_i, l_i) = (2k_i + 1)(2l_i + 1)$]

$$(G_i)_{m_i n_i; m_i' n_i'} = \delta_{m_i n_i} \delta_{n_i m_i'}, \quad (3.4)$$

$$G_i = G_{\bar{i}}, \quad G_i^2 = I, \quad G_i^T = G_i.$$

The conditions $M^2 = cI, M^\dagger M = I, M^T B M = B$ become

$$\left. \begin{aligned} M(i)M(\bar{i}) &= cI \\ M(i)^\dagger M(i) &= I \\ M(i)^T \Delta_i M(i) &= \Delta_{\bar{i}} \end{aligned} \right\}, \quad i = 1, \dots, k. \quad (3.5)$$

Since k and the d_i must be even, we write

$$\dim V = n = 4m = \sum_{i=1}^k n_i d_i.$$

We can now give the main result of this subsection.

Theorem 3.2: Keep the above notation. Suppose $s: X \rightarrow MXM^{-1}$ is an automorphism of $\mathfrak{usp}(4m)$ which is an involutive extension of s' . Then the corresponding real form of $\mathfrak{sp}(4m, \mathbb{C})$ is: (a) $\mathfrak{sp}(2m, 2m)$ if $M^2 = I$; (b) $\mathfrak{sp}(4m, \mathbb{R})$ if $M^2 = -I$.

Proof: First we restrict everything to the subspace $W_i = Y_i \oplus \bar{Y}_i$ of V , which is invariant under s . We write (abusing notation)

$$B = \begin{pmatrix} \Delta_i \otimes B_i & 0 \\ 0 & \Delta_{\bar{i}} \otimes B_i \end{pmatrix},$$

$$M = \begin{pmatrix} 0 & M(i) \otimes G_i \\ M(\bar{i}) \otimes G_{\bar{i}} & 0 \end{pmatrix}.$$

Put $\dim W_i = 4m_i = 2n_i d_i$.

(a) Suppose $M^2 = I$, so $M(\bar{i}) = M(i)^{-1}$.

Let

$$I_{2m_i, 2m_i} = \begin{pmatrix} -I_{2m_i} & 0 \\ 0 & I_{2m_i} \end{pmatrix} \in \text{USp}(4m_i).$$

Then it is easy to show that

$$U^{-1}MU = I_{2m,2m},$$

where

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} I \otimes I & M(i) \otimes G_i \\ -M(\bar{i}) \otimes I & I \otimes G_i \end{pmatrix} \in \text{USp}(4m_i).$$

Thus the two automorphisms $s: X \rightarrow MXM^{-1}$ and $\theta: X \rightarrow I_{2m,2m}XI_{2m,2m}$ [$X \in \text{usp}(4m_i)$] are conjugate within $\text{Aut}[\text{usp}(4m_i)]$, and so they give the same real form of $\text{sp}(4m_i, \mathbb{C})$. Now the maximal compact subalgebra K , consisting of matrices which are fixed by θ , is just

$$K = \left\{ \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \middle| X_1, X_2 \in \text{usp}(2m_i) \right\} \\ \cong \text{usp}(2m_i) \oplus \text{usp}(2m_i).$$

Thus the real form of $\text{sp}(4m_i, \mathbb{C})$ corresponding to θ (and to s) is $\text{sp}(2m_i, 2m_i)^2$.

The required result for all of V is obtained by combining the above for each i ; we obtain $\text{sp}(2m, 2m)$.

(b) Suppose $M^2 = -I$, so $M(\bar{i}) = -M(i)^{-1}$.

Then

$$U^{-1}MU = iI_{2m,2m} \in \text{USp}'(4m_i),$$

where

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} I \otimes I & I \otimes I \\ iM(\bar{i}) \otimes G_i & -iM(\bar{i}) \otimes G_i \end{pmatrix} \in U(4m_i),$$

and $\text{USp}'(4m_i)$ denotes the group of unitary matrices leaving invariant the antisymmetric form

$$B' = U^T B U = \begin{pmatrix} 0 & \Delta_i \otimes B_i \\ \Delta_i \otimes B_i & 0 \end{pmatrix}.$$

Let $\text{usp}'(4m_i)$ denote the Lie algebra of $\text{USp}'(4m_i)$. The map

$$\eta: \text{usp}(4m_i) \rightarrow \text{usp}'(4m_i), \\ X \rightarrow U^{-1}XU,$$

is an isomorphism, and s induces the automorphism

$s_1 = \eta \circ s_0 \eta^{-1}: X' \rightarrow I_{2m,2m} X' I_{2m,2m}$, of $\text{usp}'(4m_i)$. The set K' of matrices which are fixed by s_1 is

$$K' = \left\{ \begin{pmatrix} X'_1 & 0 \\ 0 & -(\Delta_i \otimes B_i)^{-1} X'_1 (\Delta_i \otimes B_i) \end{pmatrix} \middle| X'_1 \in \text{u}(2m_i) \right\}.$$

Clearly $K' \cong \text{u}(2m_i)$, and so, if we revert to $\text{usp}(4m_i)$, we see that the real form is $\text{sp}(4m_i, \mathbb{R})^2$, and the result for all of V follows: the real form is $\text{sp}(4m, \mathbb{R})$. \square

(2) If all the n_i are even, there is the embedding (2.5)

$$D_2 \subset \bigoplus_{i=1}^k \text{sp}(Y_i) \subset \text{sp}(V) \quad (\rho = 1). \quad (3.6)$$

It is clear that the analysis of $3B(1)$ is still more or less valid. However, we now have the possibility that V contains self-conjugate representations Y_i , i.e., $Y_i \cong \bar{Y}_i$. We shall introduce some notation, the purpose of which will become clear in the next theorem.

Suppose that $V = \bigoplus_{i=1}^k Y_i$, where $Y_1, \dots, Y_{k'}$ are self-conjugate, and $n_i = n_{\bar{i}}$ ($i = k' + 1, \dots, k$), so that an extension of s' exists. Put $4m = \sum_{i=k'+1}^k n_i d_i$. For $i = 1, \dots, k'$ write $d_i = (2k_i + 1)^2 = d'_i + d''_i$, where $d'_i = k_i(2k_i + 1)$

$< d''_i = (k_i + 1)(2k_i + 1)$. We also write

$$\dim Y_i = n_i d_i = p_i(n'_i) + q_i(n'_i) \quad (n'_i = 0, 2, 4, \dots, n_i),$$

where

$$p_i(n'_i) = n'_i d''_i + (n_i - n'_i) d'_i, \\ q_i(n'_i) = n'_i d'_i + (n_i - n'_i) d''_i. \quad (3.7)$$

Theorem 3.3: With embedding (3.6) we suppose that $s: X \rightarrow MXM^{-1}$ is an involutive automorphism of $\text{usp}(n)$ which is an extension of s' . Then

(a) if $M^2 = I$ we have the real form

$$\text{sp} \left(2m + \sum_{i=1}^{k'} p_i(n'_i), \quad 2m + \sum_{i=1}^{k'} q_i(n'_i) \right),$$

for each possible choice of the $n'_i \in \{0, 2, \dots, n_i\}$, where $i = 1, \dots, k'$;

(b) if $M^2 = -I$ we obtain the real form $\text{sp}(n, \mathbb{R})$.

Proof: First we observe that on the subspaces $Y_i \oplus \bar{Y}_i$ ($i = k' + 1, \dots, k$) of V , the proof of Theorem 3.2 is still valid. Consider, therefore, the subspace Y_i ($i = 1, \dots, k'$) and write, abusing the notation again,

$$B = \Delta_i \otimes B_i, \quad M = M(i) \otimes G_i.$$

Let us denote by a prime the standard realizations of the symplectic algebra and group [i.e., relative to $J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$]. We claim that there is an isomorphism

$$\eta: \text{usp}(n_i d_i) \rightarrow \text{usp}'(n_i d_i), \\ X \rightarrow O^{-1} X O \quad (O^T B O = J).$$

[Choose $O' \in \text{GL}(n_i d_i, \mathbb{C})$ such that $O'^T B O' = J$. Then there is an isomorphism

$$\eta': \text{sp}(n_i d_i, \mathbb{C}) \rightarrow \text{sp}'(n_i d_i, \mathbb{C}), \\ X \rightarrow O'^{-1} X O'.$$

Clearly, if $X^+ = -X$, then

$$[\eta'(X)]^+ = -Z \eta'(X) Z^{-1}, \quad \text{where } Z = O'^+ O'.$$

Thus η' [$\text{usp}(n_i d_i)$] is a set of matrices which are skew-Hermitian relative to the positive-definite Hermitian form Z ; so it is a compact real form of $\text{sp}'(n_i d_i, \mathbb{C})$. By corollary 7.3 in Ref. 2, there exists an automorphism σ of $\text{sp}'(n_i d_i, \mathbb{C})$ such that

$$\sigma: \eta' [\text{usp}(n_i d_i)] \rightarrow \text{usp}'(n_i d_i), \\ Y \rightarrow O''^{-1} Y O'' \quad (O''^T J O'' = J).$$

Clearly we can put $\eta = \sigma \circ \eta'$ (i.e., $O = O'' O'$).

We also denote by η the Lie group homomorphism

$$\eta: \text{USp}(n_i d_i) \rightarrow \text{USp}'(n_i d_i), \\ A \rightarrow O^{-1} A O.$$

(a) Suppose $M^2 = I$, so $M(i)^2 = I$ ($i = 1, \dots, k'$). Since $M(i) \in \text{USp}(n_i)$ we must have $\det M(i) = 1$. Thus $M(i)$ has n'_i eigenvalues equal to -1 and the remaining $n_i - n'_i$ eigenvalues equal to $+1$, say, where $n'_i \in \{0, 2, \dots, n_i\}$. On the other hand, G_i has $2k_i + 1$ entries $+1$ on the main diagonal, with the remaining $2k_i(2k_i + 1)$ entries $+1$ occurring in mirror image positions off the diagonal. Thus G_i has $d''_i = (2k_i + 1) + k_i(2k_i + 1)$ eigenvalues $+1$ and $d'_i = k_i(2k_i + 1)$ eigenvalues -1 . By the standard results on maximal tori in compact Lie groups (e.g., Theorem 4.21 in

Ref. 14) it is clear that $\eta(M)$ is conjugate within $\text{USp}'(n_i, d_i)$ to the matrix

$$K_{p_i/2, q_i/2} = \begin{pmatrix} I_{p_i/2, q_i/2} & 0 \\ 0 & I_{p_i/2, q_i/2} \end{pmatrix},$$

where $p_i = p_i(n'_i)$ and $q_i = q_i(n'_i)$ are given by (3.7). The involutive automorphism $\theta: X \rightarrow K_{p_i/2, q_i/2} X K_{p_i/2, q_i/2}$ of $\text{usp}'(n_i, d_i)$ has as its fixed set $\text{usp}'(p_i) \oplus \text{usp}'(q_i)^2$; the same is therefore true for the automorphism $s: X \rightarrow M X M^{-1}$ of our original realization of $\text{usp}(n_i, d_i)$. Thus the real form is $\text{sp}(p_i(n'_i), q_i(n'_i))$, and the theorem for all of V follows immediately.

(b) Suppose $M^2 = -I$, so $M(i)^2 = -I$ ($i = 1, \dots, k'$).

The argument is analogous to (a). This time $M(i)$ has eigenvalues $\pm i$ which must occur with equal multiplicity $\frac{1}{2}n_i$. Consequently, $\eta(M)$ is conjugate within $\text{USp}'(n_i, d_i)$ to $iI_{n_i, d_i/2, n_i, d_i/2}$. As is well known, the resulting real form is $\text{sp}(n_i, d_i, \mathbb{R})$, and the theorem for all of V follows. \square

C. Real forms of $\text{so}(n, \mathbb{C})$

The argument is similar to the case of $\text{sp}(n, \mathbb{C})$, so we shall give a briefer account.

(1) We start with the generally valid embedding (2.5)

$$D_2 \subset \bigoplus_{i=1}^k \text{so}(Y_i) \subset \text{so}(V) \quad (\rho = 1). \quad (3.8)$$

Put $V = \bigoplus_{i=1}^k Y_i$, with Y_1, \dots, Y_k self-conjugate, and $n_i = n_i$ ($i = k' + 1, \dots, k$), so that an extension of s' exists.

The automorphisms s of the compact form $\text{uso}(n) = [\text{so}(n, \mathbb{C}) \cap \text{su}(n)]^R$ are of the form

$$s: \text{uso}(n) \rightarrow \text{uso}(n),$$

$$X \rightarrow M X M^{-1},$$

for some M in the corresponding Lie group $\text{USp}(n)$ (i.e., $M^\dagger M = I$, $M^T B M = B$). By Schur's lemma, s is involutive ($s^2 = I$) if and only if $M^2 = cI$ ($c \in \mathbb{C}$). We then have $(M^2)^T B M^2 = M^T (M^T B M) M = M^T B M = B$, and so $c^2 = 1$, i.e., $c = \pm 1$.

The explicit form of s can be derived using the same arguments as in 3B(1); s is an extension of s' if and only if (3.2) holds. Clearly M is of the form (3.3), with the $M(i)$ satisfying (3.5). However, when $i = 1, \dots, k'$, we must have $M(i)^T \Delta_i M(i) = \Delta_i$ [by (3.5)]. Taking determinants, it is clear that $M^2 = -I$ is only possible when n_i is even, for $i = 1, \dots, k'$.

This time we put $2m = \sum_{i=k'+1}^k n_i d_i$. For $i = 1, \dots, k'$, we again write $d_i = d'_i + d''_i$ and $\dim Y_i = p_i(n'_i) + q_i(n'_i)$, where $n'_i = 0, 1, 2, \dots, n_i$, and $p_i(n'_i), q_i(n'_i)$ are given by (3.7).

Theorem 3.4: For embedding (3.8), we suppose that $s: X \rightarrow M X M^{-1}$ is an involutive automorphism of $\text{uso}(n)$ which is an extension of s' . Then (a) if $M^2 = I$, we obtain the real forms

$$\text{so}\left(m + \sum_{i=1}^{k'} p_i(n'_i), m + \sum_{i=1}^{k'} q_i(n'_i)\right)$$

for every fixed choice of the $n'_i \in \{0, 1, 2, \dots, n_i\}$, $i = 1, \dots, k'$; (b) if $M^2 = -I$ (with the n_i necessarily all even for $i = 1, \dots, k'$), we obtain the real form $\text{so}^*(n)$.

Proof: (a) Suppose $M^2 = I$. On the subspace $Y_i \oplus \bar{Y}_i$ of V with dimension $2m_i = 2n_i d_i$ ($i = k' + 1, \dots, k$), we write

$$B = \begin{pmatrix} \Delta_i \otimes B_i & 0 \\ 0 & \bar{\Delta}_i \otimes B_i \end{pmatrix},$$

$$M = \begin{pmatrix} 0 & M(i) \otimes G_i \\ M(\bar{i}) \otimes G_i & 0 \end{pmatrix}$$

Exactly as in the proof of Theorem (3.2), part (a), we have

$$U M U^{-1} = I_{m_i, m_i} \in \text{UO}(2m_i),$$

for the same U [which now belongs to $\text{UO}(2m_i)$], and the corresponding real form of $\text{so}(2m_i, \mathbb{C})$ is $\text{so}(m_i, m_i)$.

On Y_i ($i = 1, \dots, k'$), the argument goes like the proof of Theorem 3.3 part (a), except $\det M(i)$ can now be ± 1 , so n'_i may be chosen from the set $\{0, 1, 2, \dots, n_i\}$. The matrix $\eta(M)$ is conjugate within $\text{UO}(n_i, d_i)$ to the matrix $I_{p_i(n'_i), q_i(n'_i)}$, and so the corresponding real form is $\text{so}[p_i(n'_i), q_i(n'_i)]$. The theorem follows immediately.

(b) If $M^2 = -I$, the proof is similar to part (b) of Theorem 3.3; the real form is $\text{so}^*(n)$, with maximal compact subalgebra $u(n/2)$. \square

(2) If all the n_i are even we have the embedding (2.5)

$$D_2 \subset \bigoplus_{i=1}^k \text{so}(Y_i) \subset \text{so}(V) \quad (\rho = -1). \quad (3.9)$$

As in 3B(1), there can be no self-conjugate Y_i 's. In fact, the argument of Theorem 3.2 carries over to this case, if we replace $\text{usp}(4m)$ by $\text{uso}(4m)$, and $\text{USp}(4m)$ by $\text{UO}(4m)$, where we have again written $\dim V = 4m$.

Theorem 3.5: For embedding (3.9), we suppose that $s: X \rightarrow M X M^{-1}$ is an automorphism of $\text{uso}(4m)$ which is an involutive extension of s' . Then the real form of $\text{so}(4m, \mathbb{C})$ is (a) $\text{so}(2m, 2m)$ if $M^2 = I$; (b) $\text{so}^*(4m)$ if $M^2 = -I$.

D. Real forms of $\text{SL}(n, \mathbb{C})$

As we know, there may not be any bilinear form with respect to which the α^μ are skew. In such a case, we are interested in the real forms of $\text{sl}(n, \mathbb{C})$ which contain $\text{sl}(2, \mathbb{C})^R$. [The following results are still valid even if such a form does exist, but then S could not be all of $\text{sl}(n, \mathbb{C})$].

The compact form $\text{su}(n)$ has automorphisms of the form

$$s: \text{su}(n) \rightarrow \text{su}(n),$$

$$X \rightarrow M X M^{-1},$$

for some $M \in \text{U}(n)$ (i.e., $M^\dagger M = I$). Clearly $s^2 = I$ if and only if $M^2 = cI$ ($c \in \mathbb{C}$); this time c need not be ± 1 (but $|c| = 1$). The explicit form of such an s is given by (3.3) with

$$M(i) M(\bar{i}) = cI, \quad (3.10)$$

$$M(i)^\dagger M(i) = I.$$

However, there are also automorphisms of the form

$$s: \text{su}(n) \rightarrow \text{su}(n),$$

$$X \rightarrow N \bar{X} N^{-1} = -N X^T N^{-1}$$

[$N \in \text{U}(n)$]. We have $s^2 = I$ if and only if $N \bar{N} = cI$ ($c \in \mathbb{C}$). In this case we must have $c = \pm 1$. If s is to be an extension of s' , it is clear that

$$N K_3 N^{-1} = -L_3,$$

$$NK_{\pm} N^{-1} = -L_{\mp}. \quad (3.11)$$

From (3.11), it follows that N has matrix blocks

$$[i|N|j] = \delta_{ij} N(i) \otimes C_i, \quad (3.12)$$

where $N(i) \in \text{GL}(n_i, \mathbb{C})$, and C_i is the $d_i \times d_i$ matrix

$$(C_i)_{m_i n_i; m'_i n'_i} = (-1)^{k_i + l_i + m_i + n_i} \delta_{m_i, -n'_i} \delta_{n_i, -m'_i}, \quad (3.13)$$

$$C_{\bar{i}} = C_i, \quad C_i^2 = \rho I, \quad C_i^T = \rho C_i.$$

The conditions $N^{\dagger} N = I$ and $N\bar{N} = cI$ give

$$\begin{aligned} N(i)^{\dagger} N(i) &= I, \\ N(i) \overline{N(\bar{i})} &= c\rho I, \quad \forall i. \end{aligned} \quad (3.14)$$

First of all, let us suppose that $\rho = -1$ (half-integral spin)

$$D_2 \subset \text{sl}(n, \mathbb{C}) \quad (\rho = -1), \quad (3.15)$$

so that $V = \oplus_{i=1}^k Y_i$ with no Y_i self-conjugate. We take $n_i = n_{\bar{i}}, \forall i$, and write $\dim V = 4m$.

Theorem 3.6: With embedding (3.15), let s be an automorphism of $\text{su}(4m)$ which is an involutive extension of s' . Then (i) if s is of the form $X \rightarrow MXM^{-1}$, the corresponding real form of $\text{sl}(4m, \mathbb{C})$ is $\text{su}(2m, 2m)$; (ii) if s is of the form $X \rightarrow N\bar{X}N^{-1}$, the real form is (a) $\text{sl}(4m, \mathbb{R})$ (if $N\bar{N} = I$) and (b) $\text{su}^*(4m)$ (if $N\bar{N} = -I$).

Proof: (i) We have $M^2 = cI$, with $|c| = 1$. Put $c = e^{i\theta} (\theta \in \mathbb{R})$. On the subspace $Y_i \oplus \bar{Y}_i$, with dimension $4m_i$, we have

$$U^{-1} M U = e^{i\theta/2} \begin{pmatrix} -I \otimes I & 0 \\ 0 & I \otimes I \end{pmatrix} = e^{i\theta/2} I_{2m_i, 2m_i} \in \text{U}(4m_i),$$

where

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} I \otimes I & e^{-i\theta} M(i) \otimes I \\ -e^{-i\theta/2} M(\bar{i}) \otimes C_i & e^{-i\theta/2} I \otimes C_i \end{pmatrix} \in \text{U}(4m_i).$$

Thus the two automorphisms $s: X \rightarrow MXM^{-1}$ and $\theta: X \rightarrow I_{2m_i, 2m_i} X I_{2m_i, 2m_i}$ are conjugate within $\text{Aut}[\text{su}(4m_i)]$; an argument similar to that used in Theorem 3.2 then says that the required real form of $\text{sl}(4m_i, \mathbb{C})$ is $\text{su}(2m_i, 2m_i)$, and the result for all of V follows.

(ii) (a) If $N\bar{N} = I$, then on $Y_i \oplus \bar{Y}_i$ we see that

$$U^{-1} N \bar{U} = I,$$

where

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} I \otimes I & iI \otimes I \\ N(\bar{i}) \otimes C_i & -iN(\bar{i}) \otimes C_i \end{pmatrix} \in \text{U}(4m_i).$$

So s is conjugate within $\text{Aut}[\text{su}(4m_i)]$ to the automorphism $\theta: X \rightarrow \bar{X}$, which gives the real form $\text{sl}(4m_i, \mathbb{R})$, and the result follows.

(b) if $N\bar{N} = -I$ we have

$$U^{-1} N \bar{U} = J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \in \text{U}(4m_i),$$

where

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} I \otimes I & iI \otimes I \\ -iN(\bar{i}) \otimes C_i & iN(\bar{i}) \otimes C_i \end{pmatrix} \in \text{U}(4m_i).$$

Thus s is conjugate within $\text{Aut}[\text{su}(4m_i)]$ to the automorphism $\theta: X \rightarrow J\bar{X}J^{-1}$, which gives $\text{su}^*(4m_i)$. \square

If $\rho = 1$ (integral spin), we have

$$D_2 \subset \text{sl}(n, \mathbb{C}) \quad (\rho = 1), \quad (3.16)$$

with $Y_1, \dots, Y_{k'}$, self-conjugate. We assume that $n_i = n_{\bar{i}}$ ($i = k' + 1, \dots, k$); put $2m = \sum_{i=k'+1}^k n_i d_i$, and define $p_i(n'_i)$, $q_i(n'_i)$ as in (3.7), where $n'_i \in \{0, 1, 2, \dots, n_i\}$. We have the following theorem, the proof of which is obvious.

Theorem 3.7: With embedding (3.16), let s be an automorphism of $\text{su}(n)$ which is an involutive extension of s' . Then (i) if s is of the form $X \rightarrow MXM^{-1}$ the real forms of $\text{sl}(n, \mathbb{C})$ are

$$\text{su} \left(m + \sum_{i=1}^{k'} p_i(n'_i), \quad m + \sum_{i=1}^{k'} q_i(n'_i) \right),$$

for $n'_i \in \{0, 1, \dots, n_i\}$ ($i = 1, \dots, k'$); (ii) if s is of the form $X \rightarrow N\bar{X}N^{-1}$, the real form is (a) $\text{sl}(n, \mathbb{R})$ (if $N\bar{N} = I$) and (b) $\text{su}^*(n)$ (if $N\bar{N} = -I$), the n_i ($i = 1, \dots, k'$) being necessarily even in case (b).

E. Real forms of G_2, F_4, E_6

In this section we shall consider those embeddings of D_2 in the exceptional Lie algebras¹⁵ $S = G_2, F_4$ or E_6 , such that

(i) S contains a vector operator α^μ .

(ii) There is at least one real form S_0 of S containing $\text{sl}(2, \mathbb{C})^R$.

Problem (ii) has been considered by Ekins and Cornwall,¹² and they have given a complete list. It is easy to pick out those possibilities for which (i) is satisfied. We do this by finding the branching rules $S \rightarrow D_2$ for the adjoint representation of S , using the method described by Navon and Patera,¹⁶ based on Dynkin's¹⁷ theory (see also Ref. 1). This will give us the number of linearly independent vector operators belonging to S in each case.

We number the simple roots as in Humphreys¹⁸ (p. 58). Representations are denoted by their highest weights.¹

(1) $S = G_2$: There are no real forms¹² of G_2 containing $\text{sl}(2, \mathbb{C})^R$. Indeed, the only possible embeddings (specified by the reduction of the natural representation of G_2) are

$$G_2 \supset D_2,$$

$$(a) (1, 0) \rightarrow \pi = (1, 0) \oplus (\frac{1}{2}, \frac{1}{2}),$$

or

$$(b) (1, 0) \rightarrow \bar{\pi} = (0, 1) \oplus (\frac{1}{2}, \frac{1}{2}).$$

Although in each case there are two linearly independent vector operators in $\text{End} V = \text{Hom}(V, V)$, we find that the branching rule for the adjoint representation $(0, 1)$ of G_2 is

$$(a) (0, 1) \rightarrow \rho = (1, 0) \oplus (0, 1) \oplus (\frac{3}{2}, \frac{1}{2})$$

$$(b) (0, 1) \rightarrow \bar{\rho}.$$

This result is well known.¹⁹ It means that G_2 does not contain a vector operator. Thus, for G_2 , we conclude that neither (i) nor (ii) can be satisfied.

(2) $S = F_4$: The embeddings of D_2 in F_4 satisfying (ii) are given by specifying the reduction of the natural representation $V(\omega) = (0001)$ (dim. 26)

$$(a) (0001) \rightarrow (\frac{1}{2}, \frac{1}{2}) \oplus 4(\frac{1}{2}, 0) \oplus 4(0, \frac{1}{2}) \oplus 6(0, 0),$$

$$(b) (0001) \rightarrow 4(\frac{1}{2}, \frac{1}{2}) \oplus (0, 1) \oplus (1, 0) \oplus 4(0, 0),$$

TABLE I. The highest four weights of the F_4 -module $V(\omega)$ when regarded as a D_2 -module.

	(a)	(b)	(c)
μ_1	$(\frac{1}{2}, \frac{1}{2})$	(1,0)	$(\frac{3}{2}, \frac{1}{2})$
μ_2	$(\frac{1}{2}, 0)$	$(\frac{1}{2}, \frac{1}{2})$	$(\frac{3}{2}, -\frac{1}{2})$
μ_3	$(\frac{1}{2}, 0)$	$(\frac{1}{2}, \frac{3}{2})$	(1,1)
μ_4	$(\frac{1}{2}, 0)$	$(\frac{1}{2}, \frac{1}{2})$	(1,0)

$$(c) (0001) \rightarrow (\frac{3}{2}, \frac{1}{2}) \oplus (\frac{1}{2}, \frac{3}{2}) \oplus (1,1) \oplus (0,0).$$

In cases (a) and (b), there are embeddings of $sl(2, \mathbb{C})^R$ in both noncompact real forms of F_4 ; in case (c) $sl(2, \mathbb{C})^R$ can only be embedded in one such real form.¹²

The root system Φ for F_4 can be constructed in \mathbb{R}^4 , with $\{\epsilon_i \mid i = 1, \dots, 4\}$ being the usual orthonormal basis, as follows²⁰

$$\begin{aligned} &\pm \epsilon_i, \quad 1 \leq i \leq 4, \\ &\pm (\epsilon_i \pm \epsilon_j), \quad 1 \leq i < j \leq 4, \\ &\pm \frac{1}{2}(\epsilon_1 \pm \epsilon_2 \pm \epsilon_3 \pm \epsilon_4) \end{aligned}$$

(with all possible choices of sign). The simple roots are

$$\begin{aligned} \alpha_1 &= \epsilon_2 - \epsilon_3, \\ \alpha_2 &= \epsilon_3 - \epsilon_4, \\ \alpha_3 &= \epsilon_4, \\ \alpha_4 &= \frac{1}{2}(\epsilon_1 - \epsilon_2 - \epsilon_3 - \epsilon_4). \end{aligned}$$

If f denotes the embedding of D_2 in F_4 , then the map f^* , defined in Ref. 1, is specified by²¹

$$\begin{aligned} f^*(\alpha_1) &= -\mu_1 - \mu_2 + \mu_3 + 2\mu_4, \\ f^*(\alpha_2) &= \mu_3 - \mu_4, \\ f^*(\alpha_3) &= \mu_2 - \mu_3, \\ f^*(\alpha_4) &= \mu_1 - \mu_2, \end{aligned}$$

where $\mu_1, \mu_2, \mu_3, \mu_4$ are the highest four weights of $V(\omega)$ regarded as a D_2 -module. These weights are given in Table I. From these we can calculate the values of $f^*(\alpha_i)$; these are shown in Table II.

It is now possible to find $f^*(\alpha)$, for each $\alpha \in \Phi$, in cases (a), (b), and (c), giving the required branching rules for the adjoint representation

$$F_4 \rightarrow D_2$$

$$\begin{aligned} (a) (1000) &\rightarrow (1,0) \oplus (0,1) \oplus 5(\frac{1}{2}, \frac{1}{2}) \oplus 4(\frac{1}{2}, 0) \\ &\quad \oplus 4(0, \frac{1}{2}) \oplus 10(0,0), \\ (b) (1000) &\rightarrow (1,1) \oplus 4(1,0) \oplus 4(0,1) \oplus 4(\frac{1}{2}, \frac{1}{2}) \\ &\quad \oplus 3(0,0), \\ (c) (1000) &\rightarrow (2,1) \oplus (1,2) \oplus (\frac{3}{2}, \frac{1}{2}) \oplus (\frac{1}{2}, \frac{3}{2}) \\ &\quad \oplus (1,0) \oplus (0,1). \end{aligned}$$

Thus we see that the number of linearly independent vector operators in F_4 is five in case (a), four in case (b), and none in case (c). So (a) and (b) are the only ones satisfying (i), although, since (a) is an equation with mixed spins ($\rho = \pm 1$), there is no way that $\pi(D_2)$ and the α^μ can generate all of F_4 .

(3) $S = E_6$: The embeddings of D_2 in E_6 satisfying (ii) are given by specifying the reduction of the natural representation $V(\omega)$ (dim. 27)

$$\begin{aligned} (a) (100000) &\rightarrow (\frac{1}{2}, \frac{1}{2}) \oplus 4(\frac{1}{2}, 0) \oplus 4(0, \frac{1}{2}) \oplus 7(0,0), \\ (b) (100000) &\rightarrow 4(\frac{1}{2}, \frac{1}{2}) \oplus (0,1) \oplus (1,0) \oplus 5(0,0), \\ (c) (100000) &\rightarrow (\frac{1}{2}, \frac{3}{2}) \oplus (\frac{3}{2}, \frac{1}{2}) \oplus (1,1) \oplus 2(0,0), \\ (d) (100000) &\rightarrow 3(1,0) \oplus 3(0,1) \oplus (1,1), \\ (e) (100000) &\rightarrow (\frac{1}{2}, \frac{1}{2}) \oplus (1, \frac{1}{2}) \oplus (\frac{1}{2}, 1) \oplus (1,0) \\ &\quad \oplus (0,1) \oplus (\frac{1}{2}, 0) \oplus (0, \frac{1}{2}) \oplus (0,0), \\ (f) (100000) &\rightarrow (0,2) \oplus (2,0) \oplus (\frac{3}{2}, \frac{3}{2}) \oplus (0,0). \end{aligned}$$

In cases (a), (b), (d), $sl(2, \mathbb{C})^R$ can be embedded in all three noncompact real forms of E_6 ; in cases (c), (e), (f), $sl(2, \mathbb{C})^R$ can only be embedded in two real forms.¹²

The embeddings (a), (b), and (c) arise from

$$\begin{aligned} E_6 &\rightarrow F_4 \\ (\text{natural}) (100000) &\rightarrow (0001) \oplus (0000), \\ (\text{adjoint}) (010000) &\rightarrow (1000) \oplus (0001). \end{aligned}$$

Thus in (a) there are $5 + 1 = 6$ vector operators in E_6 ; but E_6 can still never be generated by $\pi(D_2)$ and the α^μ . In (b) there are $4 + 4 = 8$ vector operators in E_6 , and in (c) none.

We reject (d) and (f) immediately, since there are no vector operators at all. Thus we need only find directly the branching rule for the adjoint representation of E_6 according to embedding (e).

The root system Φ for E_6 is constructed in \mathbb{R}^8 as²⁰

$$\begin{aligned} &\pm (\epsilon_i \pm \epsilon_j) \quad 1 \leq i < j \leq 5 \quad \left[\begin{array}{l} \nu(i) = 0 \text{ or } 1 \\ \sum_i \nu(i) \text{ even} \end{array} \right] \\ &\frac{1}{2}(\epsilon_8 - \epsilon_7 - \epsilon_6 + \sum_{i=1}^5 (-1)^{\nu(i)} \epsilon_i) \end{aligned}$$

The simple roots being

$$\begin{aligned} \alpha_1 &= \frac{1}{2}[\epsilon_8 - \epsilon_7 - \epsilon_6 + (\epsilon_1 - \epsilon_2 - \epsilon_3 - \epsilon_4 - \epsilon_5)], \\ \alpha_2 &= \epsilon_1 + \epsilon_2, \\ \alpha_3 &= \epsilon_2 - \epsilon_1, \\ \alpha_4 &= \epsilon_3 - \epsilon_2, \\ \alpha_5 &= \epsilon_4 - \epsilon_3, \\ \alpha_6 &= \epsilon_5 - \epsilon_4. \end{aligned}$$

We have¹⁷

$$f^*(\alpha_1) = \mu_1 - \mu_2 = (0, \frac{1}{2}),$$

TABLE II. The values of $f^*(\alpha_i)$ for the embedding f of D_2 in F_4 .

	(a)	(b)	(c)
$f^*(\alpha_1)$	$(\frac{1}{2}, -\frac{1}{2})$	(0,1)	(0,1)
$f^*(\alpha_2)$	(0,0)	(0,0)	(0,1)
$f^*(\alpha_3)$	(0,0)	(0,0)	$(\frac{1}{2}, -\frac{3}{2})$
$f^*(\alpha_4)$	$(0, \frac{1}{2})$	$(\frac{1}{2}, -\frac{1}{2})$	(0,1)

$$f^*(\alpha_2) = \frac{1}{3}(-\mu_1 - \mu_2 - \mu_3 + 6\mu_4 - 4\mu_{25} + 2\mu_{26} + 2\mu_{27}) \\ = (0, 1),$$

$$f^*(\alpha_3) = \mu_2 - \mu_3 = (0, \frac{1}{2}),$$

$$f^*(\alpha_4) = \mu_3 - \mu_4 = (\frac{1}{2}, -\frac{3}{2}),$$

$$f^*(\alpha_5) = \mu_{25} - \mu_{26} = (0, \frac{1}{2}),$$

$$f^*(\alpha_6) = \mu_{26} - \mu_{27} = (0, \frac{1}{2}).$$

Thus we can find $f^*(\alpha)$, $\forall \alpha \in \Phi$, and the branching rule for the adjoint representation is

$$E_6 \rightarrow D_2,$$

$$(010000) \rightarrow (\frac{3}{2}, \frac{1}{2}) \oplus (\frac{1}{2}, \frac{3}{2}) \oplus (1, 1) \oplus 2(1, \frac{1}{2}) \oplus 2(\frac{1}{2}, 1) \\ \oplus 2(1, 0) \oplus 2(0, 1) \oplus 2(\frac{1}{2}, \frac{1}{2}) \oplus 2(\frac{1}{2}, 0) \\ \oplus 2(0, \frac{1}{2}) \oplus (0, 0).$$

We conclude that there are two vector operators in E_6 ; however, $\pi(D_2)$ and the α^i can never generate all of E_6 , since (e) corresponds to mixed spins ($\rho = \pm 1$).

ACKNOWLEDGMENT

The work described in this paper is part of the author's Ph.D. thesis (University of Adelaide, 1978). I should like to thank Professor C. A. Hurst for many stimulating conversations; I also thank Dr. A. L. Carey and Professor J. F. Cornwell for their helpful comments.

This work was supported by a Postgraduate Research Award from the Australian Government, and by a University of Adelaide Research Grant. I am also grateful to the Royal Commission for the Exhibition of 1851 for an Overseas Scholarship.

¹A. Cant and C. A. Hurst, *J. Aust. Math. Soc. B* **20**, 446 (1978).

²S. Helgason, *Differential Geometry and Symmetric Spaces* (Academic, New York, 1962).

³This notation means that π is a representation acting in the vector space V .

⁴As in Ref. 2, $\mathfrak{sl}(2, \mathbb{C})^{\mathbb{R}}$ denotes $\mathfrak{sl}(2, \mathbb{C})$ considered as a Lie algebra over \mathbb{R} .

⁵H. J. Bhabha, *Rev. Mod. Phys.* **17**, 200 (1945).

⁶The Kronecker product $A \otimes B$ of two matrices A, B , in this context, denotes the matrix with entries $(A_{pq}B)$, where A_{pq} are the matrix elements of A . This notation will prove to be very useful in dealing with repeated representations.

⁷A. I. Mal'cev, *Rec. Math. (Mat. Sbornik) N. S.* **16** (58), 163 (1945), **19** (61), 523 (1946).

⁸J. F. Cornwell, *Rep. Math. Phys.* **2**, 239 (1971).

⁹J. F. Cornwell, *Rep. Math. Phys.* **2**, 289 (1971).

¹⁰J. F. Cornwell, *Rep. Math. Phys.* **3**, 91 (1972).

¹¹J. M. Ekins and J. F. Cornwell, *Rep. Math. Phys.* **5**, 17 (1974).

¹²J. M. Ekins and J. F. Cornwell, *Rep. Math. Phys.* **7**, 167 (1975).

¹³The label \bar{l} refers to the conjugate (l_i, k_i) of (k_i, l_i) .

¹⁴J. F. Adams, *Lectures on Lie Groups* (Benjamin, New York, 1969).

¹⁵We omit the more complicated cases E_7, E_8 .

¹⁶A. Navon and J. Patera, *J. Math. Phys.* **8**, 489 (1967).

¹⁷E. B. Dynkin, *Am. Math. Soc. Trans. (Ser. 2)* **6**, 111 (1957).

¹⁸J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory* (Springer, Berlin, 1972).

¹⁹F. L. Bauer, *Sitz. Bay. Akad. Wiss.* **13**, 111 (1952).

²⁰N. Bourbaki, *Groupes et Algèbres de Lie* (Hermann, Paris, 1968), Chaps. 4-6.

²¹In Ref. 17, $f^*(\alpha_1)$ is incorrect. We can derive the correct formula by observing that the top four weights of the natural representation $V(\omega)$ of F_4 are

	f^*
λ_4	$\rightarrow \mu_1,$
$\lambda_4 - \alpha_4$	$\rightarrow \mu_2,$
$\lambda_4 - (\alpha_3 + \alpha_4)$	$\rightarrow \mu_3,$
$\lambda_4 - (\alpha_2 + \alpha_3 + \alpha_4)$	$\rightarrow \mu_4,$

and using the fact that $\lambda_4 = \alpha_1 + 2\alpha_2 + 3\alpha_3 + 2\alpha_4$.

The role of noncompact Lie algebras in relativistic wave equations. II

A. Cant

Department of Theoretical Physics, University of St. Andrews, St. Andrews, Fife, Scotland

(Received 5 December 1979; accepted for publication 20 March 1980)

This paper continues the study, begun in a previous paper, of the role of real Lie algebras in the theory of finite-dimensional Lorentz invariant wave equations of the form $(\alpha^\mu \partial_\mu + i\kappa)\psi(x) = 0$. The connection with the discrete transformations of space reflection and charge conjugation is established. The consequences for the formation of infinite-dimensional wave equations are briefly discussed.

PACS numbers: 11.10.Qr, 03.65.Fd, 02.20.Sv, 02.30.Jr

1. INTRODUCTION

In a previous paper,¹ hereafter referred to as I, we began to examine the role played by real Lie algebras in the theory of finite-dimensional Lorentz-invariant wave equations of the form

$$(\alpha^\mu \partial_\mu + i\kappa)\psi(x) = 0, \quad (1.1)$$

where the α^μ ($\mu = 0, 1, 2, 3$) are $n \times n$ matrices and κ is a real nonzero constant. Such a wave equation is specified by a representation π of the Lorentz Lie algebra $\mathfrak{sl}(2, \mathbb{C})^R$, acting in a space V , which admits a vector operator $\{\alpha^\mu\}$. We defined S to be the Lie algebra over \mathbb{C} generated by $\pi(D_2)$ and the α^μ , where D_2 denotes the complexification of $\mathfrak{sl}(2, \mathbb{C})^R$.

In I we considered the problem of finding all possible real forms S_0 of S that contain $\mathfrak{sl}(2, \mathbb{C})^R$, and solved it explicitly for the generic cases $S = \mathfrak{sp}(n, \mathbb{C})$, $\mathfrak{so}(n, \mathbb{C})$, and $\mathfrak{sl}(n, \mathbb{C})$.

In the present paper, which is a direct continuation of I, we shall show in Sec. 2 that the existence of the discrete transformations of space reflection and charge conjugation is closely related to the existence of a real form S_0 of S containing $\mathfrak{sl}(2, \mathbb{C})^R$. As mentioned in I, we show that for very many equations, invariance under these transformations leads to distinguished real forms. In Sec. 3 we briefly consider the formation of infinite-dimensional equations, and describe how one can predict the nature of the spectra of solutions corresponding to timelike or spacelike momenta.

Any unexplained notation is as in I. Some results of Ref. 2 will also be used. We write

$$(\pi, V) = \left(\bigoplus_{i=1}^k \psi_i, \bigoplus_{i=1}^k Y_i \right), \quad (1.2)$$

where Y_i is the direct sum of n_i copies of (k_i, l_i) .

2. THE RELATION BETWEEN SPACE REFLECTION, CHARGE CONJUGATION AND REAL FORMS

A. Space reflection

We restate the results of Ref. 2 in a more general form, allowing for the presence of repeated representations in π .

If

$$\psi^P(x') = P\psi(x), \quad (x^{0'} = x^0, \mathbf{x}' = -\mathbf{x}),$$

then P satisfies²

$$PK_3P^{-1} = L_3, \quad PK_\pm P^{-1} = L_\pm. \quad (2.1)$$

We also require

$$P^2 = cI, \quad (2.2)$$

where $c = 1$ for integral spin ($\rho = 1$) and $c = \pm 1$ for half-integral spin ($\rho = -1$). (When we consider charge conjugation in 2.2, we shall see that we have to take $c = -1$ for certain equations with half-integral spin.) The parity operator P exists if and only if the subrepresentations (k, l) and (l, k) of π always occur with the same multiplicity.

The wave equation (1.1) is invariant under space reflection if and only if²

$$P\alpha^0P^{-1} = \alpha^0. \quad (2.3)$$

The matrix blocks of P are, in terms of the decomposition (1.2),

$$[i|P|j] = \delta_{\bar{i}\bar{j}} \Pi(i) \otimes G_i. \quad (2.4)$$

where $\Pi(i) \in \text{GL}(n_i, \mathbb{C})$ and G_i is given by (3.4) in I. We have, from (2.2) and (2.3),

$$\Pi(i)\Pi(\bar{i}) = cI, \quad \forall i, \quad (2.5)$$

$$\Pi(i)A_{\bar{i}\bar{j}}\Pi(j)^{-1} = -A_{ij}, \quad \forall i, j. \quad (2.6)$$

If $i \equiv \bar{i}, j \equiv \bar{j}$, then (2.6) becomes

$$\Pi(i)A_{ij}\Pi(j)^{-1} = -A_{ij}, \quad (2.7)$$

which restricts the form of the matrices $\Pi(i)$ when $i \equiv \bar{i}$. For example, we have the following results.

Proposition 2.1: Suppose that π contains no repeated subrepresentations, and write $V = \bigoplus_{r=1}^l V_r$, V_r irreducible, with (say) V_1, \dots, V_l self-conjugate. Then invariance under space reflection is only possible if

$$\Pi(r) = -\Pi(s) = \pm c^{1/2} \quad \text{whenever}^3 \\ V_r \rightleftharpoons V_s \quad (1 \leq r, s \leq l).$$

Proof: This follows from (2.7) and the fact that $\Pi(i)^2 = cI$. \square

Proposition 2.2: Suppose S is irreducible (and thus semi-simple²), and that $V = \bigoplus_{i=1}^k Y_i$, where each Y_i is the direct sum of n_i copies of (k_i, k_i) . Then space-reflection invariance demands that

$$\Pi(i) = \pm (-1)^{i-1} I_{n_i}, \quad \forall i.$$

Proof: Let R be the matrix with blocks

$$[i|R|j] = \delta_{ij} (-1)^{i-1} \Pi(i) \otimes I_{d_i}.$$

Clearly $[R, \pi(x)] = 0, \forall x \in D_2$, and it follows from (2.7) that $[R, \alpha^0] = 0$. Thus R commutes with all the matrices in S , and so by Schur's Lemma R is a multiple of the identity. The result follows since V corresponds to integral spin ($c = 1$). \square

Clearly the map $\theta_\rho: X \rightarrow PXP^{-1}$ is very closely related to

the involutive automorphisms of the form $s: X \rightarrow MXM^{-1}$ ($X \in U$, the compact real form of S), which are extensions of s' , where s' is the automorphism of $\mathfrak{su}(2) \oplus \mathfrak{su}(2)$ that gives rise to the Lorentz Lie algebra $\mathfrak{sl}(2, \mathbb{C})^R$. Such automorphisms were discussed in detail in I. In fact P and M must both satisfy (2.1); they also have the property that $P^2, M^2 = \pm I$. However, we still have to check that θ_p is actually an automorphism of U . The situation is clarified by the following.

Proposition 2.3: Suppose S is irreducible, α^0 is Hermitian, and P is given ($P^2 = cI$) such that (1.1) is invariant under space reflection. Then $P \in U(n)$. If further $S \subseteq \mathfrak{sp}(n, \mathbb{C})$ [$\mathfrak{so}(n, \mathbb{C})$] relative to a bilinear form B , then $P' \in \text{USp}(n)$ [$\text{UO}(n)$], where $P' = \sigma^{1/2}P$ ($\sigma = \pm 1$).

Proof: We have from (2.1) and the Hermiticity properties of \mathbf{K} and \mathbf{L}

$$(P^\dagger P)\mathbf{K}(P^\dagger P)^{-1} = \mathbf{K}, \quad (P^\dagger P)\mathbf{L}(P^\dagger P)^{-1} = \mathbf{L},$$

and from (2.3) and the fact that α^0 is Hermitian

$$(P^\dagger P)\alpha^0(P^\dagger P)^{-1} = \alpha^0.$$

Therefore $P^\dagger P$ commutes with everything in S , and so by Schur's Lemma $P^\dagger P = kI$ ($k > 0$). Since $P^2 = cI$, we have $k^2 = 1$, whence $P^\dagger P = I$, so $P \in U(n)$.

If $BXB^{-1} = -X^T$ ($\forall X \in S$) and we put $Q = P^TBP$, then using (2.1) we obtain

$$\begin{aligned} Q\mathbf{K}Q^{-1} &= P^TBP\mathbf{K}P^{-1}B^{-1}(P^T)^{-1} = P^TB\mathbf{L}B^{-1}(P^T)^{-1} \\ &= -P^T\mathbf{L}^T(P^T)^{-1} = -(P^{-1}\mathbf{L}P)^T = -\mathbf{K}^T, \end{aligned}$$

and similarly

$$QLQ^{-1} = -\mathbf{L}^T.$$

We have, using (2.3),

$$Q\alpha^0Q^{-1} = -(\alpha^0)^T.$$

Thus $QXQ^{-1} = -X^T$, $\forall X \in S$, and so we must have⁴ $Q = P^TBP = \sigma B$, where $\sigma \in \mathbb{C}$. Using $P^2 = cI$, we find that $\sigma = \pm 1$, whence $P' = \sigma^{1/2}P \in \text{USp}(n)$ [$\text{UO}(n)$]. \square

We are mainly interested in the generic cases $S = \mathfrak{sp}(n, \mathbb{C})$, $\mathfrak{so}(n, \mathbb{C})$, and $\mathfrak{sl}(n, \mathbb{C})$, which were described in I, for which $U = \mathfrak{usp}(n)$, $\mathfrak{uso}(n)$, $\mathfrak{su}(n)$. The above result then says that $P' \in \text{USp}(n)$, $\text{UO}(n)$, $\text{U}(n)$, so that $\theta_p = \theta_{p'}$ is indeed an automorphism of U ; it is an involutive extension of s' , and so it gives rise to a real form S_0 that contains $\mathfrak{sl}(2, \mathbb{C})^R$. When $S = \mathfrak{sl}(n, \mathbb{C})$, S_0 is of the form $\mathfrak{su}(p, q)$ (see Theorems 3.6 and 3.7 of I). When $S = \mathfrak{sp}(n, \mathbb{C})$ or $\mathfrak{so}(n, \mathbb{C})$, the nature of the real form S_0 depends on whether $P'^2 = +I$ ($\sigma c = 1$) or $P'^2 = -I$ ($\sigma c = -1$) (see Theorems 3.2–3.5 of I).

The invariance condition (2.3) tells us that⁵

$$i\alpha^0 \in \mathbf{K} = \{X \in U \mid PXP^{-1} = X\},$$

and

$$\alpha^i \in \mathbf{P} = \{X \in U \mid PXP^{-1} = -X\}, \quad \text{so } i\alpha^i \in S.$$

\mathbf{K} is the maximal compact subalgebra of S_0 .

At this stage we know nothing about σ . However, as we shall see, it is a reasonable conjecture that $\rho\sigma c = 1$. In order to investigate this claim, we need to write down the relevant formulas in a way that allows them to be compared with each other.

Consider the D_2 submodule Y_i of V , with $i \neq \bar{i}$. Y_i is the direct sum of n_i copies of (k_i, l_i) . If S is irreducible, then

clearly in the graph² of V there exists a symmetric path Γ from i to \bar{i} :

$$i \rightleftharpoons i_1 \rightleftharpoons \dots \rightleftharpoons i_m \rightleftharpoons \bar{i},$$

where

$$i_m \equiv \bar{i}_1, \quad i_{m-1} \equiv \bar{i}_2, \dots$$

This is so because, by (2.6), $A_{ij} = 0 \Leftrightarrow A_{\bar{j}\bar{i}} = 0$. Put

$$A_{i\bar{i}}^{\Gamma} = A_{i_1, i_2, \dots, i_m, \bar{i}},$$

$$A_{\bar{i}i}^{\Gamma} = A_{\bar{i}_m, \bar{i}_{m-1}, \dots, \bar{i}_1, i}.$$

The condition (2.6) for invariance under space reflection then gives

$$\Pi(i)^{-1}A_{i\bar{i}}^{\Gamma}\Pi(\bar{i}) = (-1)^{m+1}A_{\bar{i}i}^{\Gamma} = \rho A_{\bar{i}i}^{\Gamma}, \quad (2.8)$$

since m is even (odd) when the spin is half integral (integral).

The condition $B\alpha^0B^{-1} = -(\alpha^0)^T$ leads to

$$\Delta_i A_{i\bar{i}}^{\Gamma} \Delta_{\bar{i}}^{-1} = (A_{\bar{i}i}^{\Gamma})^T, \quad (2.9)$$

where B is written in the form¹

$$B = \bigoplus_{i=1}^k (\Delta_i \otimes B_i). \quad (2.10)$$

We also have, from the fact that $P^TBP = \sigma B$ and $(\alpha^0)^\dagger = \alpha^0$,

$$\Pi(i)^T \Delta_i \Pi(i) = \sigma \Delta_{\bar{i}}, \quad \forall i, \quad (2.11)$$

$$(A_{i\bar{i}}^{\Gamma})^\dagger = A_{\bar{i}i}^{\Gamma}, \quad \forall i, \quad i \neq \bar{i}. \quad (2.12)$$

The conditions (2.5), (2.8), (2.9), (2.11), and (2.12), along with the assumed irreducibility of S , are what we need in order to see if $\rho\sigma c = 1$. We now consider a range of examples.

Proposition 2.4: With the above assumptions, suppose that there is some i ($i \neq \bar{i}$) for which n_i is odd, and that there is a symmetric path Γ from i to \bar{i} for which $\det(A_{i\bar{i}}^{\Gamma}) \neq 0$. Then $\rho\sigma c = 1$.

Proof: Since $\det(A_{i\bar{i}}^{\Gamma}) \neq 0$, and so $\det(A_{\bar{i}i}^{\Gamma}) \neq 0$, we have

$$\begin{aligned} \det[A_{i\bar{i}}^{\Gamma}(A_{\bar{i}i}^{\Gamma})^{-1}] &= \det[\Delta_{\bar{i}}\Delta_i^{-1}] \quad [\text{by (2.9)}] \\ &= \sigma^{n_i} \det[\Pi(i)^2] \quad [\text{by (2.11)}]. \end{aligned}$$

But it is also equal to

$$\begin{aligned} \rho^{n_i} \det[\Pi(i)\Pi(\bar{i})^{-1}] & \quad [\text{by (2.8)}] \\ &= (\rho c)^{n_i} \det[\Pi(i)^2] \quad [\text{by (2.5)}]. \end{aligned}$$

Thus $(\rho\sigma c)^{n_i} = 1$ and so $\rho\sigma c = 1$ since n_i is odd. \square

Proposition 2.5: Keep the assumptions of Proposition 2.2. Then $\rho\sigma c = 1$.

Proof: By Proposition 2.2, we find that

$$\Pi(i)^T \Delta_i \Pi(i) = \Delta_i. \quad \text{But by (2.11),}$$

$\Pi(i)^T \Delta_i \Pi(i) = \sigma \Delta_{\bar{i}} = \sigma \Delta_i$. Thus $\sigma = 1$, and $\rho\sigma c = 1$ because $\rho = c = 1$ in this case. \square

Proposition 2.6: If π contains no repeated subrepresentations, then $\rho\sigma c = 1$.

Proof: This follows from Propositions 2.4 and 2.5. \square

A deeper result is the following.

Proposition 2.7: Let $V = Y_i \oplus Y_{\bar{i}}$, with Y_i the direct sum of n_i copies of (k_i, l_i) , where $k_i < l_i$. Then $\rho\sigma c = 1$.

Proof: In this case $\rho = -1$ and the path from i to \bar{i} is trivial ($m = 0$). First of all we note that the matrices $A_{\bar{i}}$ and

$A_{\bar{i}}$ cannot be singular. For if $A_{\bar{i}}$ (and thus $A_{\bar{i}}$) are singular, let $\Gamma_{\bar{i}}$ and Γ_i be the projectors onto the kernels of $A_{\bar{i}}$ and $A_{\bar{i}}$, and put

$$\Gamma = \begin{pmatrix} \Gamma_i \otimes I & 0 \\ 0 & \Gamma_{\bar{i}} \otimes I \end{pmatrix}.$$

Then it is clear that $[\Gamma, \pi(x)] = 0, \forall x \in D_2, \Gamma^2 = \Gamma, \Gamma \neq 0$ or I , and $(1 - \Gamma)\alpha^0 \Gamma = 0$. But this is exactly the condition⁶ that S be reducible—which we have ruled out.

Put

$$R(i) = \Pi(i)A_{\bar{i}}^{-1}, \\ R(\bar{i}) = -\Pi(\bar{i})A_{\bar{i}}^{-1},$$

and write

$$R = \begin{pmatrix} R(i) \otimes I & 0 \\ 0 & R(\bar{i}) \otimes I \end{pmatrix}.$$

Then

$$R(i)A_{\bar{i}} = \Pi(i) = -A_{\bar{i}}\Pi(\bar{i})A_{\bar{i}}^{-1} \quad [\text{by (2.6)}] \\ = A_{\bar{i}}R(i).$$

Similarly, using (2.5) and (2.6), we have

$$R(\bar{i})A_{\bar{i}} = A_{\bar{i}}R(i),$$

and so $[R, \alpha^0] = 0$. Clearly R commutes with all of S , therefore $R = kI$ ($k \in \mathbb{C}, k \neq 0$), and

$$\Pi(i) = kA_{\bar{i}}, \quad \Pi(\bar{i}) = -kA_{\bar{i}}.$$

From (2.9) we have

$$k^{-1}\Delta_i \Pi(i) \Delta_{\bar{i}}^{-1} = -k^{-1}\Pi(\bar{i})^T,$$

i.e.,

$$\Pi(i)^T \Delta_i \Pi(i) = -c \Delta_{\bar{i}} \quad [\text{by (2.5)}].$$

Comparing this with (2.11), we have $\sigma = -c$, whence $\rho\sigma c = 1$. \square

In the general case we expect that irreducibility will force certain relations between $\Pi(i)$ and $A_{\bar{i}}$. If there are many couplings present, with many distinct symmetric paths Γ from i to \bar{i} , then these relations are difficult to find, but hopefully they imply that $\rho\sigma c = 1$. The fact that $\rho\sigma c = 1$ for such a large class of equations strongly suggests that it is true in general.

That $\rho\sigma c = 1$ has important consequences. We have $P^2 = \sigma c I = \rho I$. Thus for a given equation, with ρ fixed, and $S = \text{sp}(n, \mathbb{C})$ [$\text{so}(n, \mathbb{C})$], the real form S_0 of S determined by the parity operator P will be unique. The value of c is irrelevant. The possible real forms are as follows:

$$S = \text{sp}(n, \mathbb{C}) \begin{cases} \rho = -1, & \text{sp}(4m, \mathbb{R}) \\ \rho = +1, & \text{sp}(2p, 2q), \end{cases} \\ S = \text{so}(n, \mathbb{C}) \begin{cases} \rho = +1, & \text{so}(p, q), \\ \rho = -1, & \text{so}^*(4m), \end{cases}$$

where we have used Theorems 3.2–3.5 of I.

The real forms *not* listed above arise naturally if we consider, instead of P , the operator M ($M^2 = c'I, c' = \pm 1$) satisfying (2.1) and

$$M\alpha^0 M^{-1} = -\alpha^0. \quad (2.3')$$

As in Proposition 2.3, we find that $M^T B M = \sigma' B$ ($\sigma' = \pm 1$). The analysis of the properties of M can be de-

rived from those of P by the formal replacement $\rho \rightarrow \rho' = -\rho, \sigma \rightarrow \sigma', c \rightarrow c'$ in (2.5), (2.8), (2.9), (2.11), and (2.12). We conjecture that $\rho'\sigma'c' = 1$, and we have $M'^2 = -P'^2 = -\rho'I$, where $M' = \sigma'^{1/2}M$. The automorphism $\theta_{M'}$ of $\text{usp}(n)$ [$\text{uso}(n)$] leads to the following real forms:

$$\text{sp}(n, \mathbb{C}) \begin{cases} \rho = -1, & \text{sp}(2m, 2m), \\ \rho = +1, & \text{sp}(n, \mathbb{R}), \end{cases} \\ \text{so}(n, \mathbb{C}) \begin{cases} \rho = +1, & \text{so}^*(n), \\ \rho = -1, & \text{so}(2m, 2m), \end{cases}$$

again using Theorems 3.2–3.5 of I.

In such a case, (2.3') says that

$$i\alpha^0 \in P' = \{X \in U \mid M X M^{-1} = -X\},$$

and

$$\alpha^j \in K' = \{X \in U \mid M X M^{-1} = X\}, \quad \text{so } \alpha^\mu \in S_0.$$

K' is the maximal compact subalgebra.

It should be noted, however, that M (unlike P) does *not* arise in a physical way.

B. Charge conjugation

Again we restate more generally the results of Ref. 2.

We have

$$\psi^c(x) = C \overline{\psi(x)},$$

with²

$$CK_3 C^{-1} = -L_3, \quad CK_{\pm} C^{-1} = -L_{\mp}, \quad (2.13)$$

$$C\bar{C} = I, \quad (2.14)$$

$$C\bar{P} = PC. \quad (2.15)$$

A charge conjugation operator exists if and only if the subrepresentations (k, l) and (l, k) of π always occur with the same multiplicity.

The wave equation (1.1) is invariant under charge conjugation if and only if²

$$C \overline{\alpha^0} C^{-1} = -\alpha^0. \quad (2.16)$$

The matrix blocks of C are, in terms of (1.2),

$$[i|C|j] = \delta_{\bar{i}j} C(i) \otimes C_i, \quad (2.17)$$

where $C(i) \in \text{GL}(n_i, \mathbb{C})$, and C_i is given by (3.13) in I. From (2.14) and (2.15) we get

$$C(i) \overline{C(\bar{i})} = \rho I, \quad \forall i, \quad (2.18)$$

$$C(i) \overline{\Pi(\bar{i})} = \Pi(i) C(\bar{i}), \quad \forall i. \quad (2.19)$$

Note that if we combine (2.18) and (2.19), we obtain

$$\Pi(\bar{i}) C(i) \overline{\Pi(\bar{i})} C(i) = \Pi(\bar{i}) \Pi(i) C(\bar{i}) \overline{C(i)} \\ = c\rho I.$$

If $X = \Pi(\bar{i}) C(i)$, then $X\bar{X} = c\rho I$, and taking determinants gives $\det(X\bar{X}) = |\det(X)|^2 = (c\rho)^{n_i}$. This is consistent with the assumption $P^2 = I$ for integral spin ($\rho = 1$); with $P^2 = \pm I$ for half-integral spin ($\rho = -1$) if all the n_i are *even*; but only with $P^2 = -I$ for half-integral spin when the n_i are all *odd*.⁷ No mixture of even and odd n_i 's is allowed for half-integral spin.

The condition (2.16) becomes

$$C(i)\bar{A}_{ij}C(j)^{-1} = A_{ij}, \quad \forall ij. \quad (2.20)$$

Clearly the map $\eta_c: X \rightarrow -CX^TC^{-1} = C\bar{X}C^{-1}$ is very closely related to the involutive automorphisms of the form $s: X \rightarrow -NX^TN^{-1}$ ($X \in U, NN = I$), which are extensions of s' . An argument similar to that used in Proposition 2.3 shows that $C^+C = I$, i.e., $C \in U(n)$. Thus if $S = \mathfrak{sl}(n, \mathbb{C})$, it is clear that η_c is an automorphism of $\mathfrak{su}(n)$; it gives rise to the unique real form $\mathfrak{sl}(n, \mathbb{R})$ by Theorems 3.6 and 3.7 in I. The invariance condition (2.16) says that

$$\begin{aligned} i\alpha^0 \in K &= \{X \in \mathfrak{su}(n) \mid -CX^TC^{-1} = X\}, \\ \alpha^j \in P &= \{X \in \mathfrak{su}(n) \mid -CX^TC^{-1} = -X\}. \end{aligned}$$

$K \cong \mathfrak{uso}(n)$ is the maximal compact subalgebra of $\mathfrak{sl}(n, \mathbb{R})$.

3. INFINITE-DIMENSIONAL EQUATIONS

Suppose we are given a finite-dimensional wave equation, i.e., a representation π of D_2 (acting in V) with an embedding $D_2 \subset S$, where $\alpha^\mu \in S$, such that $\mathfrak{sl}(2, \mathbb{C})^R$ can be embedded in a real form S_0 of S . This situation was discussed in I. As in Sec. 3.1 of I, we have the Cartan decomposition $S_0 = K \oplus iP$; S_0 contains the elements $i(\alpha^0)^{\sim}$ and $(\alpha^j)^{\sim}$. On the group level we have $SL(2, \mathbb{C}) \subset \mathcal{S}_0$, with $\exp[i(\alpha^0)^{\sim}]^{\sim} \in \mathcal{S}_0$ and $\exp[(\alpha^j)^{\sim}]^{\sim} \in \mathcal{S}_0$.

We can generate an infinite-dimensional equation by considering an irreducible representation ρ of \mathcal{S}_0 . Its properties—most important being the $SL(2, \mathbb{C})$ content and the possible values of the momenta—are in principle obtainable from the representation theory of \mathcal{S}_0 .

It is mathematically convenient to assume that ρ is a *unitary* representation of \mathcal{S}_0 acting in a Hilbert $H(\rho)$. Also, it is physically appropriate, since it leads in perturbation theory to vertex functions with very well-behaved form factors.^{8,9}

Since ρ is K finite, the wavefunction $\psi(x)$ appears initially in a discrete “infinite-component” form, corresponding to the decomposition of ρ into irreducible finite-dimensional representations of K . It is straightforward to obtain the spin content from the branching rules for $K \rightarrow \mathfrak{so}(3, \mathbb{C})$, obtained, for example, by Dynkin’s method.² (We observe that in general ρ will *not* be $\mathfrak{so}(3, \mathbb{C})$ finite, so a particular spin could occur infinitely many times.) However, finding the branching rules for $\rho: \mathcal{S}_0 \rightarrow SL(2, \mathbb{C})$ is difficult because we typically have a direct integral of unitary irreducible representations, and the decomposition may not be at all obvious from the discrete form of the representation of the Lie algebra S_0 . Clearly Dynkin’s theory is of no use now, because the corresponding S module \mathcal{W} will have no weight spaces as a D_2 module. It may be better to attack the problem on the group level: if ρ is induced from a representation of some subgroup \mathcal{H} of \mathcal{S}_0 , then a method due to Mackey¹⁰ may enable us to find the $SL(2, \mathbb{C})$ decomposition by examining the double cosets of \mathcal{S}_0 with respect to \mathcal{H} and $SL(2, \mathbb{C})$.

It is well known¹¹ that the infinite-dimensional wave equation corresponding to ρ will in general possess a spectrum of solutions corresponding to spacelike momenta ($p^2 < 0$) as well as the more familiar solutions with timelike

momenta ($p^2 > 0$). All these solutions must be considered when quantization is carried out.⁸

It is important from both the technical and physical points of view to know whether these spectra are discrete or continuous. Clearly, if ρ is $\mathfrak{so}(3, \mathbb{C})$ finite, then since α^0 commutes with the generators of rotations it is clear that α^0 will have a discrete spectrum, and so there is a discrete spectrum of timelike solutions. In more general cases, we can use the results of Sec. 2.1. If we are given a finite-dimensional parity-invariant wave equation, then there is a distinguished real form $S_0 = K \oplus iP$ of S [when S is one of the generic algebras $\mathfrak{sp}(n, \mathbb{C}), \mathfrak{so}(n, \mathbb{C}), \mathfrak{sl}(n, \mathbb{C})$], such that $i\alpha^0 \in K, \alpha^j \in P$. Clearly $\rho(\alpha^0)$ and $\rho(\alpha^j)$ have discrete and continuous spectra, respectively. Thus there will be a discrete spectrum of timelike solutions and a continuous spectrum of spacelike solutions. On the other hand, the real form S_0 determined by the operator M discussed in 2.1 gives the reverse situation: $i\alpha^0 \in P', \alpha^j \in K'$, so there is a discrete spectrum of *spacelike* solutions and a continuous spectrum of *timelike* solutions.

We now give some examples.

(i) The most familiar example is the case $V = (\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$, where $S_0 = \mathfrak{sp}(4, \mathbb{R}) \cong \mathfrak{so}(3, 2)$, and ρ is the ladder representation of $\mathfrak{sp}(4, \mathbb{R})$, realized in terms of boson operators a_1, a_2, a_1^*, a_2^* . The initial finite equation is Dirac’s equation, which is parity invariant [so we can take $i\alpha^0 \in K \cong \mathfrak{u}(2)$], and the resulting infinite-dimensional equation consists of the two Majorana equations.⁸ The Lorentz and spin contents are just¹²

$$\begin{aligned} \mathfrak{sp}(4, \mathbb{R}) &\rightarrow \mathfrak{sl}(2, \mathbb{C})^R \rightarrow \mathfrak{so}(3) \\ \rho &\rightarrow \{\frac{1}{2}, 0\} \rightarrow (\frac{1}{2}) \oplus (\frac{3}{2}) \oplus (\frac{5}{2}) \oplus \dots, \\ &\oplus \{0, \frac{1}{2}\} \rightarrow (0) \oplus (1) \oplus (2) \oplus \dots \end{aligned}$$

There is a discrete spectrum of timelike solutions and a continuous spectrum of spacelike solutions.⁸

(ii) In general, we can consider the ladder representation ρ of $\mathfrak{sp}(n, \mathbb{R})$ (n even) for any of the embeddings 3.1 or 3.6 in I, where we can take $i\alpha^0 \in K \cong \mathfrak{u}(\frac{1}{2}n)$ if it is assumed that the finite equation is parity invariant. Such equations have been considered by many authors, in particular Palev¹³ and Takabayasi,^{14,15} who introduces certain local kinematical variables ξ_i , satisfying characteristic algebraic relations, which are interpreted as describing relativistic internal motion.

(a) Takabayasi¹⁴ considers the so-called “spinor model”, where $V = 2(\frac{1}{2}, 0) \oplus 2(0, \frac{1}{2})$; the ladder representation ρ of $\mathfrak{sp}(8, \mathbb{R})$ “contains” all the representations in the principal series of unitary irreducible representations of $SL(2, \mathbb{C})$.¹⁶

(b) Starting with the (parity invariant) Kursunoglu equation,² with $V = (1, \frac{1}{2}) \oplus (\frac{1}{2}, 1)$, one can consider the ladder representation of $\mathfrak{sp}(12, \mathbb{R})$. This example turns out to be very complicated: The Casimir operators for $SL(2, \mathbb{C})$ have a very messy form. The jump in complexity from the Majorana case (i) to this one is great, and it is certain that ρ is a direct integral of $SL(2, \mathbb{C})$ representations. The method of induced representations mentioned earlier may be helpful in this case.

(iii) Consider again the Dirac equation, $V = (\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$, and consider the embedding $\mathfrak{sl}(2, \mathbb{C})^R \subset \mathfrak{sp}(2, 2) \cong \mathfrak{so}(4, 1)$, with $\alpha^j \in K \cong \mathfrak{usp}(2) \oplus \mathfrak{usp}(2)$.

We let ρ be a member of the principal series of unitary irreducible representations of $\text{Sp}(2,2)$; then Ström¹⁰ has calculated the direct integral decomposition of ρ into representations of the principal series of $\text{SL}(2,\mathbb{C})$. The wave equation based on ρ will have a discrete spectrum of spacelike solutions, but a continuous spectrum of timelike solutions.

ACKNOWLEDGMENTS

The work described in this paper is part of the author's Ph.D. thesis (University of Adelaide, 1978). I should like to thank Professor C. A. Hurst for many stimulating conversations; I also thank Dr. A. L. Carey and Professor J. F. Cornwell for their helpful comments.

This work was supported by a Postgraduate Research Award from the Australian Government, and by a University of Adelaide Research Grant. I am also grateful to the Royal Commission for the Exhibition of 1851 for an Overseas Scholarship.

¹A. Cant, *J. Math. Phys.* **22**, 870 (1980).

²A. Cant and C. A. Hurst, *J. Aust. Math. Soc. B* **20**, 446 (1978).

³The notation $V_r \rightleftharpoons V_s$ means that there is a two-way coupling from V_r to V_s , i.e., α'' has nonzero matrix blocks $\langle r|\alpha''|s\rangle$ and $\langle s|\alpha''|r\rangle$. It was used in Ref. 2.

⁴H. Samelson, *Notes on Lie Algebras* (Van Nostrand-Reinhold, New York, 1969), p. 142.

⁵The same letter P is used to denote both a parity operator and the -1 eigenspace of an involutive automorphism s , but no confusion should arise in practice.

⁶E. C. G. Sudarshan, M. A. K. Khalil, and W. J. Hurley, *J. Math. Phys.* **18**, 855 (1977).

⁷This argument is a generalization of the well-known argument that we must take $P^2 = -I$ for the Dirac equation. See, for example, Y. Takahashi, *Introduction to Field Quantization* (Pergamon, New York, 1969), p. 177.

⁸D. Tz. Stoyanov and I. T. Todorov, *J. Math. Phys.* **9**, 2146 (1968).

⁹*Proceedings of the 1967 Conference on Particles and Fields*, edited by C. R. Hagen, G. Guralnik, and V. A. Mathur (Interscience, New York, 1967).

¹⁰S. Ström, in *Boulder Lectures in Theoretical Physics* (Colorado Association, University, 1971), Vol. XIII.

¹¹W. Rühl, *Commun. Math. Phys.* **6**, 312 (1967).

¹²We have used here the customary notation $\{l_0, l_1\}$ for the representations of $\text{sl}(2,\mathbb{C})^{\mathbb{R}}$. [See, for example, I. M. Gel'fand, R. A. Minlos, and Z. Ya. Shapiro, *Representations of the Rotation and Lorentz Groups and their Applications* (Pergamon, New York, 1963).]

¹³C. D. Palev, *Nuovo Cimento A* **62**, 585 (1969).

¹⁴T. Takabayasi, in *Proceedings of the 8th Nobel Symposium*, edited by N. Svartholm (Wiley-Interscience, New York, 1968).

¹⁵T. Takabayasi in Ref. 9.

¹⁶Also considered in Ref. 14 is Yukawa's "bilocal model", based on $V = 2(\frac{1}{2}, \frac{1}{2})$, with ρ the ladder representation of $\text{sp}(8, \mathbb{R})$; ρ is a direct integral of these representations $\{l_0, l_1\}$ for which $l_0 l_1 = 0$ and l_0 is integral. (Of course, no vector operator is present in this model!)

Sum rules in classical scattering^{a)}

D. Bollé^{b)}

Instituut voor Theoretische Fysica, Universiteit Leuven, B-3030 Leuven, Belgium

T. A. Osborn

Cyclotron Laboratory, Department of Physics, University of Manitoba, Winnipeg, Manitoba, R3T 2N2, Canada

(Received 6 August 1980; accepted for publication 26 November 1980)

This paper derives sum rules associated with the classical scattering of two particles. These sum rules are the analogs of Levinson's theorem in quantum mechanics which provides a relationship between the number of bound-state wavefunctions and the energy integral of the time delay of the scattering process. The associated classical relation is an identity involving classical time delay and an integral over the classical bound-state density. We show that equalities between the N th-order energy moment of the classical time delay and the N th-order energy moment of the classical bound-state density hold in both a local and a global form. Local sum rules involve the time delay defined on a finite but otherwise arbitrary coordinate space volume Σ and the bound-state density associated with this same region. Global sum rules are those that obtain when Σ is the whole coordinate space. Both the local and global sum rules are derived for potentials of arbitrary shape and for scattering in any space dimension. Finally the set of classical sum rules, together with the known quantum mechanical analogs, are shown to provide a unified method of obtaining the high-temperature expansion of the classical, respectively the quantum-mechanical, virial coefficients.

PACS numbers: 11.50.Li, 11.20.Dj, 03.80.+r

I. INTRODUCTION

This paper derives a class of sum rules associated with the classical scattering of two particles. The sum rules investigated here are the classical analogs of a set of similar rules that are known to exist in quantum scattering. Most widely known of these quantum results is Levinson's theorem for the partial wave phase shift $\delta_l(\epsilon)$. For a collision with angular momentum l and energy ϵ this theorem states¹

$$\delta_l(\infty) - \delta_l(0) = \int_0^\infty d\epsilon \frac{d\delta_l(\epsilon)}{d\epsilon} = -\pi n_l. \quad (1.1)$$

The symbol n_l is the integer number of distinct eigenfunctions of the radial Schrödinger equation. The integrand in Eq. (1.1) is proportional to the time delay of the collision characterized by l and ϵ . Thus Eq. (1.1) may be interpreted as a relationship between two physical properties of the scattering system, namely the time delay and the number of bound states. Stated in this manner the classical analog of Eq. (1.1) is suggested at once. One should seek a moment property like Eq. (1.1) that relates classical time delay to the classical bound-state density.

Recently the first step in this direction has been taken.² If the potential is central and short ranged, then for scattering in three dimensions the classical sum rule parallel to Eq. (1.1) has been obtained. In this paper we will extend these results in several different ways. First we will find the form of the sum rules for scattering in any space dimension. Second, the potential that causes the scattering will be allowed to

have an arbitrary shape and not just a form that conserves angular momentum. The final generalization is the proof that the classical sum rules hold for arbitrary neighborhoods in coordinate space. This local character of the sum rules is a feature peculiar to classical mechanics that is not shared by the known quantum sum rules. The quantum rules are global statements valid only after an integration is carried out over the whole coordinate space.³⁻⁶

The basic analytical technique employed here is one adopted from the study⁷ of the time evolution of quantum systems through arbitrary point sets Σ in coordinate space. In that way, a basic connection between the time evolution and the state density has been found. This result is called the spectral property of transit time. This property states that the sum of the transit times of all scattering orbits through a region Σ is equal to the density of all scattering states with energy ϵ and support on Σ . In the following we prove that the spectral property is valid for the classical time evolution provided that the state density is that implied by the classical phase space.

Section II describes the classical scattering theory we employ and provides a proof of the spectral property. Sections III and IV state and prove the set of sum rules in their local and global form. Finally the last section gives an application of the sum rules to the problem of understanding the high-temperature behavior of the second virial coefficient. It is shown that the knowledge of the set of classical sum rules implies, exactly as in the quantum case,⁵ the determination of the coefficients of this high-temperature expansion. Throughout this derivation, we clearly expose the unifying features of the classical and the quantum problem.

^{a)}Work supported in part by a grant from the Natural Sciences and Engineering Research Council Canada and by a NATO Research Grant.

^{b)}Bevoegdverklaard Navorsers N.F.W.O., Belgium.

II. CLASSICAL SCATTERING, TRANSIT TIMES, AND SPECTRAL PROPERTY

The objective of this section is to prove the spectral property of classical time delay. In order to do this, we first recall some key elements of classical scattering theory. Transit time of a scattering orbit through a point set Σ is defined in the representation of the orbit provided by the solutions of the Hamilton–Jacobi equations. Then by relating the definition of transit time to volume integrals in phase space we are able to prove the spectral property.

The nonrelativistic collision of two particles is equivalent to the problem of a point mass moving in a potential field, once the center-of-mass motion has been removed. The position of the mass point will be given by a vector r in an n -dimensional Euclidean space. The momentum is denoted by the vector p . This pair of vectors is represented by $Z = (r, p)$. If p_i is the component of p in direction i , then the classical Hamiltonian will be defined as

$$H(Z) = \frac{1}{2\mu} \sum_{i=1}^n p_i^2 + v(r), \quad (2.1)$$

where $v(r)$ is the potential and μ is the particle mass.

Phase space is defined as the set $\Gamma = \{Z: H(Z) < \infty\}$. A central feature of our derivation is that phase space can be decomposed into two nonintersecting parts: one for bound state motion, the other for scattering. Scattering theory is defined if the short range potential satisfies the following:

(A) $v(r)$ is bounded from below by $-v_- > -\infty$. For $M < \infty$, $v(r)$ is continuous with bounded derivatives up to order 2 on $\{r: v(r) < M\}$.

(B) $|\nabla v(r)| < \text{const}|r|^{-2-\delta}$, $\delta > 0$.

Condition (B) implies that the force goes to 0 when ever r is large. Under these circumstances Newton's equations of motion

$$\dot{Z} = (\dot{r}, \dot{p}) = (\mu^{-1}p, -\nabla v)$$

have unique solutions for any initial conditions $Z_0 = (r_0, p_0)$. The map

$$S_t: Z_0 \rightarrow Z_t$$

defines a one-parameter group of canonical transformations.

First we define the bound-state region of phase space, Γ_B . Let the norm of Z be

$$|Z|^2 = \sum_{i=1}^n (r_i^2 + p_i^2).$$

Take Γ_n to be

$$\Gamma_n = \{Z: |S_t Z| \leq n \text{ for } -\infty < t < \infty\}.$$

Then, the bound-state subset of Γ is given by

$$\Gamma_B = \bigcup_{n=1}^{\infty} \Gamma_n.$$

The set Γ_B is invariant under the action of S_t and is Lebesgue measurable. The scattering orbits are, up to a set of measure zero, the complement of Γ_B . The scattering phase space is defined

$$\Gamma_S = \{Z: |S_t Z| \rightarrow \infty \text{ as } t \rightarrow +\infty \text{ and } t \rightarrow -\infty\}.$$

Completeness in phase space is then the pair of statements

$$\Gamma_S \cap \Gamma_B = \emptyset, \quad (2.2)$$

$$\Gamma_S \cup \Gamma_B = \Gamma. \quad (2.3)$$

The sign \doteq in Eq. (2.3) denotes equality of the two sets only up to sets of measure zero. For potentials satisfying restrictions (A) and (B), Hunziker⁸ has given a rigorous proof of completeness. From now on we call $H(Z)$ a *scattering system* if completeness is valid.

A second general description of orbits is available from the solution of the Hamilton–Jacobi equations.⁹ Hamilton's characteristic function, $W(r, P)$, provides a canonical transformation to a new set of coordinates Q_i, P_i in which all but Q_1 are constants of motion. For a time-independent Hamiltonian, W satisfies the partial differential equation

$$H\left(r_i, \frac{\partial W}{\partial r_i}\right) = P_1 = \epsilon. \quad (2.4)$$

We call $H(Z)$ an *integrable* Hamiltonian if Eq. (2.4) has a solution $W(r, P)$ whose Jacobian satisfies

$$\det \frac{\partial^2 W(r, P)}{\partial r \partial P} \neq 0. \quad (2.5)$$

The generalized coordinates Q_i are obtained from $W(r, P)$ by

$$Q_i = \frac{\partial W(r, P)}{\partial P_i}. \quad (2.6)$$

If condition (2.5) is valid, then Eq. (2.6) may be inverted to find r_i as a function of Q_i, P_i . Further the p_i are given by

$$p_i = \frac{\partial W(r, P)}{\partial r_i} = p_i(Q, P). \quad (2.7)$$

Because W is a canonical transformation, it preserves phase space volume elements. So, we have

$$dZ = d^n r d^n p = d\epsilon dQ_1 d\alpha, \quad (2.8)$$

where

$$d\alpha = \prod_{i=2}^n dQ_i dP_i. \quad (2.9)$$

Finally we note that

$$Q_1 = t + \beta_1, \quad (2.10)$$

where β_1 is constant. From now on we drop the subscript on Q_1 .

Consider the definition of transit time of a scattering orbit through a space region Σ . Let m denote the Lebesgue measure on \mathbb{R}^n . Assume $\Sigma \subseteq \mathbb{R}^n$ and $m(\Sigma) < \infty$. Take Z to be a point in Γ_S having energy $H(Z) = \epsilon$. The values of $Z_t = S_t Z$, for all t , define a scattering trajectory. This trajectory is particularly simple in the representation (ϵ, Q, α) . Let

$$Z = (\epsilon, Q, \alpha) \in \Gamma_S;$$

then

$$Z_t = (\epsilon, Q + t, \alpha) \in \Gamma_S \quad (2.11)$$

Thus the parameters (ϵ, α) label a scattering trajectory. Take $P_\Sigma(Z)$ to be a projector onto Σ ,

$$P_\Sigma(Z) = P_\Sigma(r, p) = \begin{cases} 1, & \text{if } r \in \Sigma \\ 0, & \text{otherwise} \end{cases}$$

For a trajectory (ϵ, α) the transit time through Σ is given by the integral (possibly infinite)

$$T_{\Sigma}(\epsilon, \alpha) = \frac{1}{\hbar} \int_{-\infty}^{\infty} dt P_{\Sigma}(Z_t). \quad (2.12)$$

We have divided this integral by \hbar , so that $T_{\Sigma}(\epsilon, \alpha)$ has the same normalization as does its quantum equivalent.⁷ Let \tilde{P}_{Σ} be the (ϵ, Q, α) representation of $P_{\Sigma}(Z)$, then we can write the transit time as

$$\begin{aligned} T_{\Sigma}(\epsilon, \alpha) &= \frac{1}{\hbar} \int_{-\infty}^{\infty} dt \tilde{P}_{\Sigma}(\epsilon, Q + t, \alpha) \\ &= \frac{1}{\hbar} \int_{-\infty}^{\infty} dQ \tilde{P}_{\Sigma}(\epsilon, Q, \alpha). \end{aligned} \quad (2.13)$$

The first form of transit time, Eq. (2.12), is valid for any H that is a scattering system. The second form, Eq. (2.13), specifically assumes that H is integrable.

The next step is to define the sum over all orbits with energy ϵ . This sum is similar to the on-shell trace in quantum mechanics. Thus we denote the sum by the trace symbol tr:

$$\text{tr} T_{\Sigma}(\epsilon) = (1/h^{n-1}) \int d\alpha T_{\Sigma}(\epsilon, \alpha). \quad (2.14)$$

At this stage it is advantageous to consider the restricted phase space volumes determined by the conditions $H(Z) \leq \epsilon$ and $r \in \Sigma$. This restricted phase space is the integral

$$\Gamma(\epsilon, \Sigma) = \frac{1}{h^n} \int_{\Gamma} dZ \chi_I(H(Z)) P_{\Sigma}(Z), \quad (2.15)$$

where χ_I is the characteristic function for the interval $I = (-\infty, \epsilon)$. We represent the momentum vector in a spherical coordinate system

$$dZ = d^n p d^n r = p_0^{n-1} dp_0 d\hat{p}^n d^n r.$$

If $H(Z) = \epsilon$, then Eq. (2.1) implies that

$$p_0 = (2\mu)^{1/2} (\epsilon - v(r))_+^{1/2}, \quad (2.16)$$

where we employ the notation

$$(x)_+^{\nu} = \Theta(x) x^{\nu}, \quad \nu > -1, \quad x \in \mathbb{R}^1,$$

and where $\Theta(x)$ is +1 for $x > 0$ and 0 for $x < 0$. Integral (2.15) then assumes the form

$$\Gamma(\epsilon, \Sigma) = \gamma_n \int_{\Sigma} d^n r (\epsilon - v(r))_+^{n/2}, \quad (2.17)$$

where the constant γ_n is

$$\gamma_n = [(2\mu)^{n/2} / nh^n] \int d\hat{p}_n.$$

For now on we denote $s = n/2$. In the following it is convenient to isolate positive and negative parts of the potential, $v_{\pm}(r) = \Theta(\pm v(r)) v(r)$. Furthermore, it is useful to divide \mathbb{R}^n up into three disjoint pieces given by the sets (for $\epsilon > 0$)

$$\begin{aligned} R_+(\epsilon) &= \{r: v(r) > \epsilon/2, \quad r \in \mathbb{R}^n\}, \\ R_-(\epsilon) &= \{r: v(r) < -\epsilon/2, \quad r \in \mathbb{R}^n\}, \\ R_0(\epsilon) &= \{r: -\epsilon/2 \leq v(r) \leq \epsilon/2, \quad r \in \mathbb{R}^n\}. \end{aligned} \quad (2.18)$$

We now determine the conditions that ensure $\Gamma(\epsilon, \Sigma)$ is finite.

Lemma 1: Let $v_{\pm} \in L^s(\Sigma)$ and $m(\Sigma) < \infty$, then $\Gamma(\epsilon, \Sigma) < \infty$ for $\epsilon < \infty$.

Proof: Take $\Gamma(\epsilon, \Sigma) = \gamma_n J_n(\epsilon, \Sigma)$. It is apparent from integral (2.17) that $J_n(\epsilon, \Sigma)$ is a positive increasing function

of ϵ . Thus it is sufficient to show $J_n(\epsilon, \Sigma) < \infty$ for positive ϵ . In this case

$$\begin{aligned} J_n(\epsilon, \Sigma) &= \int_{\Sigma \cap R_+(2\epsilon)} d^n r \epsilon^s (1 - \epsilon^{-1} v(r))^s \\ &\quad + \int_{\Sigma \cap R_-(2\epsilon)} d^n r \epsilon^s (1 - \epsilon^{-1} v(r))^s \end{aligned}$$

In the first integrand we have, for all allowed r , $|1 - \epsilon^{-1} v(r)| \leq 2$. In the second integrand $|1 - \epsilon^{-1} v(r)| \leq 2\epsilon^{-1} |v(r)|$. These inequalities give us the bound

$$J_n(\epsilon, \Sigma) \leq (2\epsilon)^s m(\Sigma) + 2^s \int_{\Sigma \cap R_-(2\epsilon)} d^n r |v(r)|^s.$$

This completes the proof.

The finiteness of $\Gamma(\epsilon, \Sigma)$ leads at once to a proof of the spectral property. To begin with, note that $\Gamma(\epsilon, \Sigma)$ can be further decomposed into bound-state and scattering components. Completeness statement (2.3) implies

$$\Gamma(\epsilon, \Sigma) = \Gamma_S(\epsilon, \Sigma) + \Gamma_B(\epsilon, \Sigma), \quad (2.19)$$

where

$$\Gamma_S(\epsilon, \Sigma) = \frac{1}{h^n} \int_{\Gamma_S} dZ \chi_I(H(Z)) P_{\Sigma}(Z) \geq 0, \quad (2.20)$$

$$\Gamma_B(\epsilon, \Sigma) = \frac{1}{h^n} \int_{\Gamma_B} dZ \chi_I(H(Z)) P_{\Sigma}(Z) \geq 0. \quad (2.21)$$

The spectral property is then summarized by

Lemma 2: Let H be integrable and constitute a scattering system. Let $v_{\pm} \in L^s(\Sigma)$ and $m(\Sigma) < \infty$. Define the bound state density $n(\epsilon, \Sigma)$ by

$$n(\epsilon, \Sigma) = \frac{\partial}{\partial \epsilon} \Gamma_B(\epsilon, \Sigma). \quad (2.22)$$

Then $\text{tr} T_{\Sigma}(\epsilon)$ and $n(\epsilon, \Sigma)$ are positive L^1 functions on every energy interval $(-\infty, \epsilon)$, $\epsilon < \infty$. For almost all ϵ

$$\frac{\partial}{\partial \epsilon} \Gamma_S(\epsilon, \Sigma) = \frac{\Theta(\epsilon)}{2\pi} \text{tr} T_{\Sigma}(\epsilon), \quad (2.23)$$

$$\frac{\partial}{\partial \epsilon} \Gamma(\epsilon, \Sigma) = \frac{\Theta(\epsilon)}{2\pi} \text{tr} T_{\Sigma}(\epsilon) + n(\epsilon, \Sigma). \quad (2.24)$$

Proof: Lemma 1 and decomposition (2.19) give

$$\infty > \Gamma(\epsilon, \Sigma) \geq \Gamma_S(\epsilon, \Sigma), \quad \epsilon < \infty.$$

Let $I = [0, \epsilon)$. Then

$$\begin{aligned} \infty > \Gamma_S(\epsilon, \Sigma) - \Gamma_B(0, \Sigma) \\ &= \frac{1}{h^n} \int_{\Gamma_S} d\epsilon' dQ d\alpha \chi_I(\epsilon') \tilde{P}_{\Sigma}(\epsilon', Q, \alpha). \end{aligned}$$

This last integral is bounded and has a positive integrand for all ϵ', Q, α . By Fubini's theorem we can change the order of integration, giving us

$$\begin{aligned} \infty > \Gamma_S(\epsilon, \Sigma) - \Gamma_S(0, \Sigma) \\ &= \frac{1}{h^n} \int_0^{\epsilon} d\epsilon' \left[\int d\alpha dQ \tilde{P}_{\Sigma}(\epsilon', Q, \alpha) \right] \\ &= \int_0^{\epsilon} d\epsilon' \frac{\text{tr} T_{\Sigma}(\epsilon')}{2\pi}. \end{aligned} \quad (2.25)$$

Thus $\text{tr} T_{\Sigma}(\epsilon)$ is an L^1 function with respect to the measure $d\epsilon$ and relation (2.23) holds a.e. The step function $\Theta(\epsilon)$ reflects

the fact that for $Z \in \Gamma_s$, $H(Z) > 0$, there are no negative energy scattering states. The same derivation of Eq. (2.23) can also be applied to $\Gamma_B(\epsilon, \Sigma)$. We have

$$\begin{aligned} \infty > \Gamma_B(\epsilon, \Sigma) &= \frac{1}{h^n} \int_{\Gamma_n} d\epsilon' \chi_I(\epsilon') \\ &\times \int d\alpha dQ \tilde{P}_\Sigma(\epsilon', Q, \alpha), \end{aligned} \quad (2.26)$$

where $I = (-\infty, \epsilon)$. Thus $\Gamma_B(\epsilon, \Sigma)$ is absolutely continuous and has a positive derivative, which is the bound-state density $n(\epsilon, \Sigma)$, for almost all ϵ . This establishes that $n(\epsilon, \Sigma)$ is L^1 . Using the completeness sum (2.19) allows one to obtain Eq. (2.24).

The final task of this section is to construct time delay from the transit times for comparable free and exact systems. Transit times for the noninteracting system are obtained by setting $v(r) = 0$ in the analysis above. Denote by $H_0(Z)$ the Hamiltonian that results when $v(r) = 0$. In this case the scattering trajectories are all straight lines. There are no bound states. Free phase space is $\Gamma_0 = \{Z: H_0(Z) < \infty\}$. If (ϵ, Q, α_0) is the solution of the Hamilton-Jacobi equation with H_0 , then $S_t^0(\epsilon, Q, \alpha_0) = (\epsilon, Q + t, \alpha_0)$. With \tilde{P}_Σ^0 the (ϵ, Q, α_0) representation of $P_\Sigma(S_t^0, Z)$, the free transit time is defined by

$$\begin{aligned} T_\Sigma^0(\epsilon, \alpha_0) &= \frac{1}{\hbar} \int_{-\infty}^{\infty} dt P_\Sigma(S_t^0, Z) \\ &= \frac{1}{\hbar} \int_{-\infty}^{\infty} dQ \tilde{P}_\Sigma^0(\epsilon, Q, \alpha_0). \end{aligned} \quad (2.27)$$

The sum of transit times is, similarly,

$$\text{tr} T_\Sigma^0(\epsilon) = \frac{1}{h^{n-1}} \int d\alpha_0 T_\Sigma^0(\epsilon, \alpha_0), \quad (2.28)$$

and the restricted free phase space is

$$\Gamma_0(\epsilon, \Sigma) = \frac{1}{h^n} \int_{\Gamma_0} dZ \chi_I(H_0(Z)) P_\Sigma(Z). \quad (2.29)$$

The spectral relation (2.23) clearly remains valid with $\Gamma_s(\epsilon, \Sigma)$ and $\text{tr} T_\Sigma(\epsilon)$ replaced by $\Gamma_0(\epsilon, \Sigma)$ and $\text{tr} T_\Sigma^0(\epsilon)$ respectively. Since both $\text{tr} T_\Sigma(\epsilon)$ and $\text{tr} T_\Sigma^0(\epsilon)$ are L^1 functions, their difference, which gives the time delay, is defined as

$$\text{tr} q(\epsilon, \Sigma) = \text{tr} T_\Sigma(\epsilon) - \text{tr} T_\Sigma^0(\epsilon). \quad (2.30)$$

The shift of restricted phase space volume induced by the perturbation $v(r)$ is

$$\Delta\Gamma(\epsilon, \Sigma) = \Gamma(\epsilon, \Sigma) - \Gamma_0(\epsilon, \Sigma). \quad (2.31)$$

Summarizing, we have:

Lemma 3: Let H be given as in Lemma 2. Let $v \in L^s(\epsilon)$ and $m(\Sigma) < \infty$, then the time delay, $\text{tr} q(\epsilon, \Sigma)$ is an L^1 function of ϵ on every finite interval $(0, \epsilon)$. For almost all ϵ .

$$\frac{\partial}{\partial \epsilon} \Delta\Gamma(\epsilon, \Sigma) = \frac{\theta(\epsilon)}{2\pi} \text{tr} q(\epsilon, \Sigma) + n(\epsilon, \Sigma). \quad (2.32)$$

This is the spectral property for time delay. One advantage it has over the spectral property for transit times is that when $\Sigma \rightarrow \mathbb{R}^n$, it is plausible that $\text{tr} q(\epsilon, \Sigma)$ converges to a finite limit whereas $\text{tr} T_\Sigma(\epsilon)$ must diverge.

III. LOCAL SUM RULES

A local sum rule will be a relation involving time delay on an arbitrary region Σ that has finite volume, $m(\Sigma) < \infty$. Because the sum rules have somewhat different behavior in even and odd dimensional space, we will treat these two cases separately.

The first objective is to analyze the stability of the phase space volume difference $\Gamma(\epsilon, \Sigma) - \Gamma_0(\epsilon, \Sigma)$. The integral form of $\Delta\Gamma(\epsilon, \Sigma)$ is

$$\Delta\Gamma(\epsilon, \Sigma) = \gamma_n \int_\Sigma d^n r [(\epsilon - v(r))_+^s - (\epsilon)_+^s]. \quad (3.1)$$

The formal expansion of $(1 - \epsilon^{-1}v)^s$ is given by the generalized binomial expansion,

$$(1 - \epsilon^{-1}v)^s = \sum_{j=0}^{\infty} \binom{s}{j} (-\epsilon^{-1}v)^j. \quad (3.2)$$

The generalized binomial coefficient in Eq. (3.2) is just $s(s-1)\dots(s-j+1)(j!)^{-1}$. These coefficients are zero for $j > s$ if s is an integer. When s is an odd multiple of $1/2$, then the coefficients are nonzero for all j . For $s > -1$ the coefficients decrease in magnitude as j increases. The series (3.2) converges absolutely if $|\epsilon^{-1}v| < 1$ and $s > -1$.¹⁰

Consider now the case where the space dimension n is odd. Here it is useful to study a regularized version of $\Delta\Gamma(\epsilon, \Sigma)$ given by

$$\begin{aligned} \Delta\tilde{\Gamma}_N(\epsilon, \Sigma) &= \gamma_n \int_\Sigma d^n r [(\epsilon - v(r))_+^s - \Theta(\epsilon) \\ &\times \sum_{j=0}^{N+s-1/2} \binom{s}{j} (-v(r))^j \epsilon^{s-j}]. \end{aligned} \quad (3.3)$$

In this integral, N may be any nonnegative integer. Essentially, increasing N means that $\Delta\tilde{\Gamma}_N(\epsilon, \Sigma)$ decreases more rapidly in ϵ for large ϵ . Specifically we have that $(\partial/\partial\epsilon) \Delta\tilde{\Gamma}_N(\epsilon, \Sigma)$ satisfies the integral identity:

Lemma 4: Let n be odd ($s = n/2$). Take $u = N + s + 1/2$, where N is any nonnegative integer. If $v \in L^1(\Sigma) \cap L^u(\Sigma)$, then

$$\int_{-\infty}^{\infty} d\epsilon \epsilon^N \frac{\partial}{\partial \epsilon} \Delta\tilde{\Gamma}_N(\epsilon, \Sigma) = 0. \quad (3.4)$$

The set Σ may be either of finite measure or equal to \mathbb{R}^n .

Proof: If one integrates (3.4) by parts, it is seen that the two estimates

$$\epsilon^N \Delta\tilde{\Gamma}_N(\epsilon, \Sigma) = O(\epsilon^{-1/2}), \quad N \geq 0, \quad (3.5)$$

$$I(\epsilon_2) \equiv \int_{-\infty}^{\epsilon_2} d\epsilon \epsilon^{N-1} \Delta\tilde{\Gamma}_N(\epsilon, \Sigma) = O(\epsilon_2^{-1/2}), \quad N \geq 1, \quad (3.6)$$

are sufficient to show Eq. (3.4) is valid. Begin with statement (3.5). Decompose the integral (3.3) into a sum of three parts. For $\epsilon > 0$,

$$\begin{aligned} \int_\Sigma d^n r &= \int_{\Sigma \cap R_0(\epsilon)} d^n r + \int_{\Sigma \cap R_+(\epsilon)} d^n r \\ &+ \int_{\Sigma \cap R_-(\epsilon)} d^n r, \end{aligned} \quad (3.7)$$

where the sets $R_0(\epsilon)$ and $R_\pm(\epsilon)$ are defined in Eq. (2.18). Examine the integral with domain $\Sigma \cap R_0(\epsilon)$ first. $r \in R_0(\epsilon)$ im-

plies $|\epsilon^{-1}v(r)| \leq 1/2$. Since the $\binom{j}{l}$ are bound by a constant as $j \rightarrow \infty$, the integrand of the $R_0(\epsilon)$ integral is bounded by

$$\begin{aligned} \epsilon^{N+s} |1 - \epsilon^{-1}v(r)|^s - \sum_{j=0}^{u-1} \binom{s}{j} (-\epsilon^{-1}v(r))^j \\ \leq \text{const} \epsilon^{N+s} |\epsilon^{-1}v(r)|^u. \end{aligned}$$

Thus the $R_0(\epsilon)$ part of the integral has the bound

$$\text{const} \epsilon^{-1/2} \int_{\Sigma \cap R_0(\epsilon)} d^n r |v(r)|^u \leq \text{const} \epsilon^{-1/2} \|v^u\|_{\Sigma}.$$

Here $\| \cdot \|_{\Sigma}$ denotes the L^1 norm on Σ .

Now look at the integral with domain $\Sigma \cap R_+(\epsilon)$. For $r \in R_+(\epsilon)$, $|2v(r)| \geq \epsilon$. First note that every term in the sum over j is $O(\epsilon^{-1/2})$. A typical term is

$$\begin{aligned} \int_{\Sigma \cap R_+(\epsilon)} d^n r \epsilon^{s+N-j} |v(r)|^j \leq \frac{\text{const}}{\epsilon^{1/2}} \int_{\Sigma \cap R_+(\epsilon)} d^n r |v(r)|^u \\ \leq \frac{\text{const}}{\epsilon^{1/2}} \|v^u\|_{\Sigma}. \end{aligned} \quad (3.8)$$

The term $\epsilon^N (\epsilon - v(r))^s$ remains to be considered. Observe that $(\epsilon - v(r))^s \leq (2^{-1}\epsilon)^s$, so

$$\epsilon^{s+N} \int_{\Sigma \cap R_+(\epsilon)} d^n r (1 - \epsilon^{-1}v(r))_+^s \leq \epsilon^{s+N} 2^{-s} m(\Sigma \cap R_+(\epsilon)). \quad (3.9)$$

However, one also has, for any $u > 0$,

$$\begin{aligned} \|v^u\|_{\Sigma} = \int_{\Sigma} d^n r |v(r)|^u \geq \int_{\Sigma \cap R_+(\epsilon)} d^n r |v(r)|^u \\ \geq |2^{-1}\epsilon|^u m(\Sigma \cap R_+(\epsilon)). \end{aligned}$$

Thus

$$m(\Sigma \cap R_+(\epsilon)) \leq (2\epsilon^{-1})^u \|v^u\|_{\Sigma}, \quad (3.10)$$

and it follows that the $\epsilon^N (\epsilon - v(r))^s$ portion of the integral with domain $\Sigma \cap R_+(\epsilon)$ in (3.7) is $O(\epsilon^{-1/2})$. Similar arguments apply to the $\Sigma \cap R_-(\epsilon)$ integral of $\epsilon^N \Delta \tilde{I}_N(\epsilon, \Sigma)$. Thus estimate (3.5) is valid.

To complete the proof of Lemma 4, we have to establish that the integral $I(\epsilon_2)$ is $O(\epsilon_2^{-1/2})$. It is not difficult to justify interchanging the order of the $d\epsilon$ and $d^n r$ integrations in $I(\epsilon_2)$. Once this is done, the $d\epsilon$ integration, for fixed r , may be carried out exactly. The result is

$$I(\epsilon_2) = \gamma_n \int_{\Sigma} d^n r [A(r, \epsilon_2) - B(r, \epsilon_2)],$$

where

$$A(r, \epsilon_2) = N \sum_{l=0}^{N-1} \binom{N-1}{l} \frac{(\epsilon_2 - v(r))_+^{N+s-l} v(r)^l}{N+s-l}, \quad (3.11)$$

$$B(r, \epsilon_2) = N \sum_{j=0}^{N+s-1/2} \binom{s}{j} \frac{(-1)^j}{N+s-j} \epsilon_2^{N+s-j} v(r)^j. \quad (3.12)$$

Decompose the $d^n r$ integration into two parts, one with $|v(r)| > \epsilon_2$ and another with $|v(r)| \leq \epsilon_2$. Estimates of the type found in Eqs. (3.8)–(3.10) show that the contribution from the r domain with $|v(r)| > \epsilon_2$ is $O(\epsilon_2^{-1/2})$. Thus it is sufficient to consider $|v(r)| \leq \epsilon_2$. Then we can write $(\epsilon_2 - v)^\alpha = \epsilon_2^\alpha (1 - \epsilon_2^{-1}v)^\alpha$, $\alpha = N + s - l$, and the binomial expansion of $(1 - \epsilon_2^{-1}v)^\alpha$ is justified for $\alpha > 0$. With this expansion

the term $A(r, \epsilon_2)$ takes the form

$$\begin{aligned} A(r, \epsilon_2) = N \sum_{l=0}^{N-1} \binom{N-1}{l} \frac{v(r)^l}{N+s-l} \\ \times \sum_{k=0}^{\infty} \binom{N+s-l}{k} \epsilon_2^{N+s-l-k} (-v(r))^k. \end{aligned}$$

Let $j = l + k$; then the summation becomes

$$\begin{aligned} A(r, \epsilon_2) = \sum_{j=0}^{\infty} N \sum_{l=0}^{l_+} \binom{N-1}{l} \binom{N+s-l}{j-l} \frac{(-1)^{j-l}}{N+s-l} \\ \times v(r)^j \epsilon_2^{N+s-j}, \end{aligned}$$

where $l_+ = \min(N-1, j)$. The next step is to utilize the following binomial identity, valid for $j \leq N+s-1/2$:

$$\binom{s}{j} \frac{(-1)^j}{N+s-j} = \sum_{l=0}^{l_+} \binom{N-1}{l} \binom{N+s-l}{j-l} \frac{(-1)^{j-l}}{N+s-l}. \quad (3.13)$$

Thus $A(r, \epsilon_2)$ can be written

$$\begin{aligned} A(r, \epsilon_2) = N \sum_{j=0}^{N+s-1/2} \binom{s}{j} \frac{(-1)^j}{N+s-j} \epsilon_2^{N+s-j} v(r)^j \\ N \sum_{j=N+s+1/2}^{\infty} \sum_{l=0}^{l_+} \binom{N-1}{l} \binom{N+s-l}{j-l} \\ \times \frac{(-1)^{j-l}}{N+s-l} \epsilon_2^{N+s-j} v(r)^j. \end{aligned} \quad (3.14)$$

So it is seen that the first group of terms in $A(r, \epsilon_2)$ exactly cancel $B(r, \epsilon_2)$. Then second group of terms all have the estimate

$$\begin{aligned} \left| \int_{\Sigma \cap R_0(2\epsilon_2)} d^n r \epsilon_2^{N+s-j} v(r)^j \right| \\ \leq \int_{\Sigma \cap R_0(2\epsilon_2)} d^n r \epsilon_2^{N+s-j} \epsilon_2^{j-u} |v(r)|^u \leq \epsilon_2^{N+s-u} \|v^u\|_{\Sigma}, \end{aligned}$$

where $j \geq u$. Recalling that $u = N + s + 1/2$, we see that this term is of order $O(\epsilon_2^{-1/2})$ if $v \in L^u(\Sigma)$. This estimate is uniform in the variable j . If $s \geq 1/2$, the sum over j of the absolute value of the binomial coefficients in (3.14) converges. Thus statement (3.6) is valid. Throughout this analysis it is possible to take Σ to be \mathbb{R}^n .

A final remark concerns the identity (3.13). This is derived by making two series expansion of the integral

$$F(x, a) = \int_a^x dy y^{N-1} (y-a)^s,$$

where $a > 0$. One series is obtained by carrying out the binomial expansion of $(y-a)^s$ and then integrating term by term. The second series is found by changing variables in the integrand to $\xi = y-a$ and then expanding $(\xi+a)^{N-1}$ in a binomial series. Equating the coefficients of the same power in x gives one Eq. (3.13). This completes the proof of Lemma 4.

Identity (3.4) is a statement about the change in phase space associated with region Σ that occurs when a potential $v(r)$ is added to the free Hamiltonian H_0 . For $N=0$, identity (3.4) says that the volume of phase space associated with region Σ is invariant under the perturbation v , i.e. that $\Delta \tilde{I}_{N=0}(\infty, \Sigma) = 0$. For $N \geq 1$, the identity (3.4) says that ϵ^N is orthogonal to the regularized density $(\partial/\partial \epsilon) \Delta \tilde{I}_N(\epsilon, \Sigma)$.

The family of local sum rules for odd space dimension is obtained by combining the spectral property of time delay, Lemma 3, with the integral identity (3.4). The result is

Theorem 1: Let n be odd ($s = n/2$). Assume H is integrable and constitutes a scattering system. Set $\lambda = \min(s, 1)$ and $u = N + s + 1/2$. If $v \in L^\lambda(\mathcal{Z}) \cap L^u(\mathcal{Z})$ and $m(\mathcal{Z}) < \infty$, then, for each integer $N \geq 0$,

$$\int_0^\infty d\epsilon \epsilon^N \left\{ \text{tr} q(\epsilon, \mathcal{Z}) - 2\pi\gamma_n \sum_{j=1}^{N+s-1/2} \binom{s}{j} (s-j) \epsilon^{s-j-1} \right. \\ \left. \times \int_{\mathcal{Z}} d^n r (-v(r))^j \right\} \\ = -2\pi \int_{-\infty}^\infty d\epsilon \epsilon^N n(\epsilon, \mathcal{Z}). \quad (3.15)$$

In this form, the sum rules relate the energy integral of the trace of the classical time delay in region \mathcal{Z} to the integral over the bound-state density $n(\epsilon, \mathcal{Z})$ in the same region. The relation holds for every finite region \mathcal{Z} . This last property is the locality property of the sum rules and is a feature not shared by any of the known³⁻⁶ quantum equivalents to Eq. (3.15). Lemma 3 established only that $n(\epsilon, \mathcal{Z})$ is L^1 with respect to $d\epsilon$ for energy intervals $(-\infty, \epsilon)$. Thus for some systems it is possible that the integrals on both sides of the equality (3.15) are infinite. Sufficient conditions to ensure that these integrals are finite will be discussed in the next section.

We now turn to the structure of local sum rules in even space dimensions. In this case s is a positive integer and $(\epsilon - v(r))^s$ is a polynomial. This polynomial behavior is the reason that the even space dimensional sum rules differ from those in odd space dimensions. We now define the regularized shift in phase space volume by

$$\Delta \tilde{F}(\epsilon, \mathcal{Z}) = \gamma_n \int_{\mathcal{Z}} d^n r [(\epsilon - v(r))_+^s \\ - \theta(\epsilon) \sum_{j=0}^{s-1} \binom{s}{j} (-v(r))^j \epsilon^{s-j}]. \quad (3.16)$$

This regularized version of $\Delta \tilde{F}(\epsilon, \mathcal{Z})$ differs from $\Delta \tilde{F}_N(\epsilon, \mathcal{Z})$ defined by Eq. (3.3) in that there is no N dependence. The function $\Delta \tilde{F}(\epsilon, \mathcal{Z})$ satisfies

Lemma 5: Let n be even. If $v \in L^1(\mathcal{Z}) \cap L^{N+s+1}(\mathcal{Z})$ and N is a nonnegative integer, then

$$\int_{-\infty}^{a(\mathcal{Z})} d\epsilon \epsilon^N \frac{\partial}{\partial \epsilon} \Delta \tilde{F}(\epsilon, \mathcal{Z}) = \gamma_n \nu(N, s) \int_{\mathcal{Z}} d^n r v(r)^{N+s}. \quad (3.17)$$

The constant $\nu(N, s)$ and $a(\mathcal{Z})$ are given by

$$\nu(N, s) = \sum_{j=0}^s \binom{s}{j} \frac{(-1)^{s-j} N}{N+j}, \quad N \geq 1, \\ = (-1)^s, \quad N = 0, \quad (3.18)$$

$$a(\mathcal{Z}) = \sup_{r \in \mathcal{Z}} v_+(r). \quad (3.19)$$

The set \mathcal{Z} may be either of finite measure or equal to \mathbb{R}^n .

Proof: An integration by parts shows that relation (3.17) is implied by a pair of identities. The first of these gives the behavior of $\epsilon^N \Delta \tilde{F}(\epsilon, \mathcal{Z})$ at infinity

$$\epsilon^N \Delta \tilde{F}(\epsilon, \mathcal{Z}) = \gamma_n \epsilon^N \int_{\mathcal{Z}} d^n r (-v(r))^s + R_N(\epsilon), \quad (3.20)$$

where the remainder $R_N(\epsilon)$ is $O(\epsilon^{-1})$. If $a(\mathcal{Z}) < \infty$, then $R_N(\epsilon) = 0$ for $\epsilon > a(\mathcal{Z})$. The second identity is

$$\int_{-\infty}^{\epsilon_2} d\epsilon \epsilon^N \epsilon^{N-1} \Delta \tilde{F}(\epsilon, \mathcal{Z}) = \gamma_n \epsilon_2^N \int_{\mathcal{Z}} d^n r (-v(r))^s \\ - \gamma_n \nu(N, s) \int_{\mathcal{Z}} d^n r v(r)^{N+s}. \quad (3.21)$$

First consider Eq. (3.20). Definition (3.16) implies for $\epsilon > 0$

$$I \equiv \epsilon^N \left[\Delta \tilde{F}(\epsilon, \mathcal{Z}) - \gamma_n \int_{\mathcal{Z}} d^n r (-v(r))^s \right] \\ = \gamma_n \epsilon^N \sum_{j=0}^s \binom{s}{j} \epsilon^{s-j} \int_{\mathcal{Z} \cap R_+(2\epsilon)} d^n r (-v(r))^j.$$

Since $v(r) \geq \epsilon$ for $v \in R_+(2\epsilon)$, one has

$$|I| \leq \frac{\gamma_n}{\epsilon} \sum_{j=0}^s \binom{s}{j} \int_{\mathcal{Z} \cap R_+(2\epsilon)} d^n r |v(r)|^{N+s+1} \\ \leq \frac{\gamma_n}{\epsilon} \sum_{j=0}^s \binom{s}{j} \|v^{N+s+1}\|_{\mathcal{Z}}. \quad (3.22)$$

So $|I|$ is $O(\epsilon^{-1})$. If $a(\mathcal{Z}) < \infty$ and $\epsilon > a(\mathcal{Z})$, then $\mathcal{Z} \cap R_+(2\epsilon) = \emptyset$ and $I = 0$. This proves Eq. (3.20). Next examine integral (3.21). Interchange the $d\epsilon$ and the $d^n r$ integrations. One then finds the result (3.21) from direct calculations for any $\epsilon_2 > 0$. It is clear that throughout this analysis the set \mathcal{Z} may be equal to \mathbb{R}^n .

The local sum rules are found by introducing the spectral property of time delay into Lemma 5. We have

Theorem 2: Let n be even ($s = n/2$). Assume that H is integrable and constitute a scattering system. If $m(\mathcal{Z}) < \infty$ and $v \in L^1(\mathcal{Z}) \cap L^{N+s+1}(\mathcal{Z})$, then, for all nonnegative integers N ,

$$\int_0^{a(\mathcal{Z})} d\epsilon \epsilon^N \left[\text{tr} q(\epsilon, \mathcal{Z}) - 2\pi\gamma_n \sum_{j=1}^{s-1} \binom{s}{j} (s-j) \epsilon^{s-j-1} \right. \\ \left. \times \int_{\mathcal{Z}} d^n r (-v(r))^j \right] \\ = -2\pi \int_{-\infty}^{a(\mathcal{Z})} d\epsilon \epsilon^N n(\epsilon, \mathcal{Z}) \\ + 2\pi\gamma_n \nu(N, s) \int_{\mathcal{Z}} d^n r v(r)^{N+s}. \quad (3.23)$$

It is of interest to compare the local sum rules for even and odd space dimension. One difference is the appearance in even dimensions of the additional potential dependent term on the right-hand side of Eq. (3.23). The simplest case to discuss this difference is the $N = 0$ sum rule. If n is odd (e.g., $s = 1/2$), then estimate (3.5) gives us $\Delta \Gamma(\infty, \mathcal{Z}) = 0$. This means that the shift in phase space volume associated with \mathcal{Z} is zero when v is added to H_0 . By contrast for n even (e.g., $s = 1$) we have from Eq. (3.20).

$$\Delta \Gamma(\infty, \mathcal{Z}) = -\gamma_2 \int_{\mathcal{Z}} d^2 r v(r).$$

Thus there is a finite shift of the restricted phase space volume in this case. The existence of this finite shift comes from the fact that $(\epsilon - v(r))^s$ is a polynomial.

The second structural difference in the two types of sum

rules comes from the fact that when $a(\mathcal{Z}) < \infty$, then it is sufficient for n even to integrate $\epsilon^N \text{tr}q(\epsilon, \mathcal{Z})$ in $d\epsilon$ from 0 to $a(\mathcal{Z})$. In odd dimensions however one must always integrate from 0 to ∞ . The most dramatic example of this difference occurs for potentials $v(r)$ that are everywhere attractive. Then $a(\mathcal{Z}) = 0$, and the even dimensional sum rule Eq. (3.17) reduces to

$$\int_{-\infty}^0 d\epsilon \epsilon^N n(\epsilon, \mathcal{Z}) = \gamma_n \nu(N, s) \int_{\mathcal{Z}} d^n r v(r)^{N+s} \quad (3.24)$$

In this case this sum rule has collapsed to a statement predicting the integral over the energy moments of the bound-state density $n(\epsilon, \mathcal{Z})$ in terms of an integral over the potential.

IV. GLOBAL SUM RULES

Global sum rules are those rules which are valid for the entire space region \mathbb{R}^n . We know already that the basic integral identities involving the regularized phase space density $(\partial/\partial\epsilon) \Delta\tilde{\Gamma}(\epsilon, \mathcal{Z})$, Eqs. (3.4) and (3.17), are valid for $\mathcal{Z} = \mathbb{R}^n$. Thus the principal problem that must be studied in order to establish the global sum rules is the characterization of the convergence properties of $\Delta\Gamma(\epsilon, \mathcal{Z})$, $n(\epsilon, \mathcal{Z})$, and $\text{tr}q(\epsilon, \mathcal{Z})$ as the region \mathcal{Z} enlarges to become \mathbb{R}^n . Let $\{\mathcal{Z}_i\}$ denote a collection of increasing sets of finite measure in \mathbb{R}^n . Then we say \mathcal{Z}_i converges strongly to $\mathcal{Z} \subseteq \mathbb{R}^n$ if for every $f \in L^1(\mathbb{R}^n)$

$$\lim_{i \rightarrow \infty} \int_{\mathbb{R}^n} d^n r f(r) [\chi_{\mathcal{Z}_i}(r) - \chi_{\mathcal{Z}}(r)] = 0. \quad (4.1)$$

In Eq. (4.1), $\chi_{\mathcal{Z}_i}(r)$ is the characteristic function for the set \mathcal{Z}_i . We examine separately the convergence problem for the phase space shift $\Delta\Gamma(\epsilon, \mathcal{Z}_i)$, the bound-state densities $n(\epsilon, \mathcal{Z}_i)$ and the time delay $\text{tr}q(\epsilon, \mathcal{Z}_i)$.

We define a potential v to be *bound-state limited* if there exists an $\epsilon_+ < \infty$ such that there are no bound-state orbits with $\epsilon > \epsilon_+$ and $\Gamma_B(\epsilon_+, \mathbb{R}^n) < \infty$. We know that the function $\Gamma_B(\epsilon, \mathbb{R}^n)$ is an increasing function of ϵ . Thus, if v is bound-state limited, it follows that $\Gamma_B(0, \mathbb{R}^n) < \infty$. The phase space representation of $\Gamma_B(0, \mathbb{R}^n)$ is

$$\Gamma_B(0, \mathbb{R}^n) = \gamma_n \int_{\mathbb{R}^n} d^n r (-v_-(r))^s. \quad (4.2)$$

So a necessary requirement for v to be bound-state limited is that $v_- \in L^s(\mathbb{R}^n)$. If v is everywhere attractive, then there are no positive energy bound orbits so that the requirement $v_- \in L^s(\mathbb{R}^n)$ is also sufficient. The condition that H is not allowed to have bound orbits with energy greater than ϵ_+ is satisfied if $\epsilon_+ > a(\mathbb{R}^n)$ where $a(\mathbb{R}^n) = \sup_{r \in \mathbb{R}^n} v_+(r) < \infty$. This class of bound-state-limited potentials should include nearly all cases of physical interest.

We will first prove the convergence properties of the phase space shift.

Lemma 6: Let $v \in L^1(\mathbb{R}^n) \cap L^s(\mathbb{R}^n)$. Then $\Delta\Gamma(\epsilon, \mathcal{Z})$, given by the integral form (3.1), is absolutely convergent for all $\epsilon < \infty$, $s > 0$ and $\mathcal{Z} \subseteq \mathbb{R}^n$. Furthermore, suppose that $\{\mathcal{Z}_i\}$ is a sequence of finite measure sets converging strongly to $\mathcal{Z} \subseteq \mathbb{R}^n$ in the way indicated by Eq. (4.1), then

$$\lim_{i \rightarrow \infty} \Delta\Gamma(\epsilon, \mathcal{Z}_i) = \Delta\Gamma(\epsilon, \mathcal{Z}), \quad \text{a.e.} \quad (4.3)$$

Proof: Consider the boundedness first. Let $\epsilon \leq 0$; then $\epsilon - v(r) \leq -v(r)$ so that we get

$$\begin{aligned} \Delta\Gamma(\epsilon, \mathcal{Z}) &= \gamma_n \int_{\mathcal{Z}} d^n r (\epsilon - v(r))^s_+ \\ &\leq \gamma_n \int_{\mathcal{Z} \cap R_{(-2|\epsilon|)}} d^n r (-v(r))^s \leq \|v^s\|, \end{aligned}$$

where

$$\|v^s\| = \int_{\mathbb{R}^n} d^n r |v(r)|^s.$$

Take now $\epsilon > 0$; then

$$\Delta\Gamma(\epsilon, \mathcal{Z}) = \gamma_n \epsilon^s \int_{\mathcal{Z}} d^n r [(1 - \epsilon^{-1}v(r))^s_+ - 1]. \quad (4.4)$$

Decompose the domain \mathcal{Z} according to (2.18), viz.,

$$\mathcal{Z} = \mathcal{Z} \cap \{R_+(\epsilon) \cup R_-(\epsilon) \cup R_0(\epsilon)\}.$$

It is then straightforward to obtain the following bound from Eq. (4.4):

$$\begin{aligned} |\Delta\Gamma(\epsilon, \mathcal{Z})| &\leq \epsilon^s \{c_1 \int_{\mathcal{Z} \cap \{R_+(\epsilon) \cup R_-(\epsilon)\}} d^n r \\ &+ c_2 \int_{\mathcal{Z} \cap R_0(\epsilon)} d^n r |\epsilon^{-1}v(r)| \\ &+ c_3 \int_{\mathcal{Z} \cap R_-(\epsilon)} d^n r |\epsilon^{-1}v(r)|^s\}, \end{aligned} \quad (4.5)$$

where c_1, c_2 , and c_3 are finite constants depending on s . To bound the first integral in (4.5), we note

$$\begin{aligned} \infty > \|v\| &> \int_{\mathcal{Z} \cap \{R_+(\epsilon) \cup R_-(\epsilon)\}} d^n r |v(r)| \\ &\geq |\epsilon/2| m(\mathcal{Z} \cap \{R_+(\epsilon) \cup R_-(\epsilon)\}). \end{aligned}$$

So we get the final result

$$|\Delta\Gamma(\epsilon, \mathcal{Z})| \leq c'_1 \epsilon^{s-1} \|v\| + c'_2 \epsilon^{s-1} \|v\| + c'_3 \|v^s\| < \infty. \quad (4.6)$$

Now we consider the convergence problem (4.3). The symmetric difference of two sets \mathcal{Z}_i and \mathcal{Z} is denoted by $\mathcal{Z}_i \Delta \mathcal{Z} = (\mathcal{Z}_i \setminus \mathcal{Z}) \cup (\mathcal{Z} \setminus \mathcal{Z}_i)$. Take $\epsilon < 0$; then $\epsilon - v(r) \leq -v(r)$ so that

$$\begin{aligned} |\Delta\Gamma(\epsilon, \mathcal{Z}_i) - \Delta\Gamma(\epsilon, \mathcal{Z})| \\ = \int_{\mathcal{Z}_i \Delta \mathcal{Z}} d^n r (\epsilon - v(r))^s_+ &\leq \int_{\mathcal{Z}_i \Delta \mathcal{Z}} (-v(r))^s_+. \end{aligned}$$

Since $|v(r)|^s \in L^1(\mathbb{R}^n)$, the definition of strong convergence of $\{\mathcal{Z}_i\} \rightarrow \mathcal{Z}$, Eq. (4.1), implies

$$|\Delta\Gamma(\epsilon, \mathcal{Z}_i) - \Delta\Gamma(\epsilon, \mathcal{Z})| \rightarrow 0 \quad \text{as } i \rightarrow \infty. \quad (4.7)$$

For $\epsilon > 0$, estimates similar to (4.6) show that (4.7) remains valid. Note that the set \mathcal{Z} may be \mathbb{R}^n .

Next, we prove the convergence properties of the bound-state density. We have

Lemma 7: If v is bound-state limited, then $n(\epsilon, \mathbb{R}^n)$ is an L^1 function of ϵ on the interval $(-\infty, \epsilon_+)$. For \mathcal{Z}_i converging strongly to \mathbb{R}^n then

$$\lim_{i \rightarrow \infty} n(\epsilon, \mathcal{Z}_i) = n(\epsilon, \mathbb{R}^n), \quad \text{a.e.} \quad (4.8)$$

If, in addition, $v_- \in L^{N+s}(\mathbb{R}^n)$, then

$$\int_{-\infty}^{\epsilon_+} d\epsilon |\epsilon^N n(\epsilon, \mathbb{R}^n)| < \infty. \quad (4.9)$$

Proof: First observe that for fixed ϵ the sets $\Gamma_B(\epsilon, \Sigma_i)$ are nondecreasing for i increasing, and $\Gamma_B(\epsilon, \mathbb{R}) \geq \Gamma_B(\epsilon, \Sigma_i)$ for all i . One has by definition of the bound-state density

$$\int_{-\infty}^{\epsilon} [n(\epsilon', \mathbb{R}^n) - n(\epsilon', \Sigma_i)] d\epsilon' = \Gamma_B(\epsilon, \mathbb{R}^n) - \Gamma_B(\epsilon, \Sigma_i) \geq 0. \quad (4.10)$$

This implies $n(\epsilon', \mathbb{R}^n) \geq n(\epsilon', \Sigma_i) \geq 0$ for almost all (a.a.) $\epsilon' \in (-\infty, \epsilon_+)$. Since $\Gamma_B(\epsilon_+, \mathbb{R}^n) < \infty$, it follows that $n(\epsilon', \mathbb{R}^n)$ is L^1 on the interval $(-\infty, \epsilon_+)$. Furthermore, since $\Gamma_B(\epsilon, \mathbb{R}^n) \geq 0$ for all ϵ , one has that $n(\epsilon, \mathbb{R}^n) \geq 0$ for a.a. ϵ . The family of functions $n(\epsilon', \Sigma_i)$ are positive and uniformly bounded a.e. by a positive L^1 function $n(\epsilon', \mathbb{R}^n)$. So, for a.a. values of ϵ' , $n(\epsilon', \Sigma_i)$ must have a limit as $i \rightarrow \infty$. The Lebesgue dominated convergence theorem then applies to the $i \rightarrow \infty$ limit of Eq. (4.10):

$$\lim_{i \rightarrow \infty} [\Gamma_B(\epsilon, \mathbb{R}^n) - \Gamma_B(\epsilon, \Sigma_i)] = \int_{-\infty}^{\epsilon} [n(\epsilon', \mathbb{R}^n) - \lim_{i \rightarrow \infty} n(\epsilon', \Sigma_i)] d\epsilon'. \quad (4.11)$$

Consider again the left-hand side of Eq. (4.11) and write it, by using Eq. (2.21), as

$$\lim_{i \rightarrow \infty} [\Gamma_B(\epsilon, \mathbb{R}^n) - \Gamma_B(\epsilon, \Sigma_i)] = \lim_{i \rightarrow \infty} \frac{1}{h^n} \int_{\Gamma_n} dZ \chi_I(H(Z)) [1 - P_{\Sigma_i}(Z)], \quad (4.12)$$

where $I = (-\infty, \epsilon)$. Since v is bound-state limited, $\chi_I(H(Z))$ is an L^1 function of Z on Γ_B . Thus the integrand $\chi_I(H(Z)) [1 - P_{\Sigma_i}(Z)]$ is bounded by the L^1 function $\chi_I(H(Z))$ and has a limit 0 as $i \rightarrow \infty$ for each Z . Again we may apply the dominated convergence theorem to bring the limit $i \rightarrow \infty$ inside the integral (4.12) and get 0 for the whole expression. This result, together with Eq. (4.11) implies (4.8).

Finally we must demonstrate the bound (4.9). First calculate the phase integral of $|H(Z)|^N \chi_I(H(Z))$ for $I = (-\infty, 0)$:

$$\int_{-\infty}^0 d\epsilon |\epsilon|^N n(\epsilon, \mathbb{R}^n) = \gamma_n s \int_{\mathbb{R}^n} d^n r \int_{v(r)}^0 d\epsilon |\epsilon|^N (\epsilon - v(r))^{s-1}. \quad (4.13)$$

For ϵ values satisfying, $v(r) \leq \epsilon < 0$ one has the inequalities $|\epsilon|^N \leq (-v(r))^N$. Thus standard estimates, like the ones used in Lemmas 4 and 5, lead to

$$\int_{-\infty}^0 d\epsilon |\epsilon|^N n(\epsilon, \mathbb{R}^n) \leq \gamma_n s \int_{\mathbb{R}^n} d^n r |v_-(r)|^{N+s}. \quad (4.14)$$

This bound (4.14) controls the behavior of the N th moment of the negative energy boundstate density. It remains to consider the positive energy moments of $n(\epsilon, \mathbb{R}^n)$. We have, for $I = [0, \epsilon)$ and $0 < \epsilon < \epsilon_+$,

$$\begin{aligned} \int_{\Gamma_B} dZ H(Z)^N \chi_I(H(Z)) &= \int_0^{\epsilon} d\epsilon' (\epsilon')^N n(\epsilon', \mathbb{R}^n) \\ &\leq \epsilon_+^N \int_0^{\epsilon} d\epsilon' n(\epsilon', \mathbb{R}^n) \\ &= \epsilon_+^N [\Gamma_B(\epsilon, \mathbb{R}^n) - \Gamma_B(0, \mathbb{R}^n)]. \end{aligned} \quad (4.15)$$

If v is bound-state limited, then $\Gamma_B(\epsilon, \mathbb{R}^n) < \infty$; if $v_- \in L^s(\mathbb{R}^n)$, then $\Gamma_B(0, \mathbb{R}^n) < \infty$. Thus Eqs. (4.15) and (4.14) establish bound (4.9).

The convergence properties of the time delay function are summarized by

Lemma 8: Suppose H is integrable and forms a scattering system. Assume that v is bound-state limited and satisfies $v \in L^1(\mathbb{R}^n) \cap L^s(\mathbb{R}^n)$. If $\{\Sigma_i\}$ is a sequence of finite measure sets in \mathbb{R}^n that converge strongly to \mathbb{R}^n , then the functions $\text{tr}q(\epsilon, \Sigma_i)$ have a limit as $i \rightarrow \infty$ for almost all ϵ ,

$$\text{tr}q(\epsilon, \mathbb{R}^n) = \lim_{i \rightarrow \infty} \text{tr}q(\epsilon, \Sigma_i). \quad (4.16)$$

The function $\text{tr}q(\epsilon, \mathbb{R}^n)$ is an L^1 function of ϵ on the interval $(-\infty, \epsilon)$ and satisfies

$$\frac{\partial}{\partial \epsilon} \Delta \Gamma(\epsilon, \mathbb{R}^n) = \frac{\theta(\epsilon)}{2\pi} \text{tr}q(\epsilon, \mathbb{R}^n) + n(\epsilon, \mathbb{R}^n). \quad (4.17)$$

Proof: For $m(\Sigma_i) < \infty$ Lemma 3 gives us

$$\frac{\partial}{\partial \epsilon} \Delta \Gamma(\epsilon, \Sigma_i) - n(\epsilon, \Sigma_i) = \frac{\theta(\epsilon)}{2\pi} \text{tr}q(\epsilon, \Sigma_i).$$

So $\text{tr}q(\epsilon, \Sigma_i)$ has a limiting value as $i \rightarrow \infty$, if both $(\partial/\partial \epsilon) \Delta \Gamma(\epsilon, \Sigma_i)$ and $n(\epsilon, \Sigma_i)$ possess a limit. Lemma 7 shows that if v is bound-state limited, then $n(\epsilon, \Sigma_i)$ has a limit and this limit is L^1 in ϵ on the interval $(-\infty, \epsilon)$. It suffices therefore to show that $(\partial/\partial \epsilon) \Delta \Gamma(\epsilon, \Sigma_i)$ has a limit and that this limit is L^1 on $(-\infty, \epsilon)$ to prove the lemma.

Take $s \geq 1$ first. Divide the space \mathbb{R}^n into two disjoint sets:

$$\begin{aligned} S_+ &= \{r: r \in \mathbb{R}^n, v(r) \geq 0\}, \\ S_- &= \{r: r \in \mathbb{R}^n, v(r) < 0\}. \end{aligned} \quad (4.18)$$

Clearly $S_+ \cup S_- = \mathbb{R}^n$ and $S_+ \cap S_- = \emptyset$. The phase space shift $\Delta \Gamma(\epsilon, \Sigma)$ can be divided accordingly, $\Delta \Gamma = \Delta \Gamma^+ + \Delta \Gamma^-$. Recalling Eq. (3.1), we have

$$\begin{aligned} \Delta \Gamma^{\pm}(\epsilon, \Sigma) &= \int_{-\infty}^{\epsilon} d\epsilon' \int_{S_{\pm} \cap \Sigma} d^n r \gamma_n \\ &\quad \times s [(\epsilon' - v(r))_+^{s-1} - (\epsilon')_+^{s-1}], \end{aligned} \quad (4.19)$$

where $\Sigma \subseteq \mathbb{R}^n$. The integrand of the energy integral in Eq. (4.19) will be called $n_{\pm}^s(\epsilon', \Sigma)$. It is straightforward to check that the $n_+^s(\epsilon', \Sigma_i)$, respectively $n_-^s(\epsilon', \Sigma_i)$, are positive increasing, respectively negative decreasing, functions of Σ_i for increasing Σ_i and fixed ϵ' . From Lemma 6, we know furthermore that $\Delta \Gamma(\epsilon', \Sigma_i)$ and $\Delta \Gamma^{\pm}(\epsilon', \Sigma_i) < \infty$ for $\Sigma_i \subseteq \mathbb{R}^n$ if $v \in L^1(\mathbb{R}^n) \cap L^s(\mathbb{R}^n)$ and $\epsilon' < \infty$. So the sequences $\{n_{\pm}^s(\epsilon, \Sigma_i)\}$ must have a limit as $i \rightarrow \infty$. Then the monotone convergence theorem gives

$$\lim_{i \rightarrow \infty} \int_{-\infty}^{\epsilon} d\epsilon' n_{\pm}^s(\epsilon', \Sigma_i) = \int_{-\infty}^{\epsilon} d\epsilon \lim_{i \rightarrow \infty} n_{\pm}^s(\epsilon', \Sigma_i). \quad (4.20)$$

Using result (4.3) of Lemma 6, the left-hand side of Eq. (4.20) can be written as

$$\lim_{i \rightarrow \infty} \Delta \Gamma^{\pm}(\epsilon, \Sigma_i) = \Delta \Gamma^{\pm}(\epsilon, \mathbb{R}^n) = \int_{-\infty}^{\epsilon} d\epsilon' n_{\pm}^s(\epsilon', \mathbb{R}^n). \quad (4.21)$$

Comparing Eqs. (4.20) and (4.21), we get the results

$$\lim_{i \rightarrow \infty} n_{\pm}^s(\epsilon', \Sigma_i) = n_{\pm}^s(\epsilon', \mathbb{R}^n)$$

for almost all ϵ , and, of course, $n_{\pm}^s(\epsilon', \mathbb{R}^n)$ are L^1 functions of ϵ' on the interval $(-\infty, \epsilon)$. Thus summing up the n_{\pm}^s and n^s gives us immediately

$$\lim_{i \rightarrow \infty} \frac{\partial}{\partial \epsilon} \Delta \Gamma(\epsilon, \Sigma_i) = \frac{\partial}{\partial \epsilon} \Delta \Gamma(\epsilon, \mathbb{R}^n) \quad \text{a.a.} \epsilon, \quad (4.22)$$

and $(\partial/\partial \epsilon) \Delta \Gamma(\epsilon, \mathbb{R}^n)$ is L^1 on $(-\infty, \epsilon)$.

The case $s = \frac{1}{2}$ can be handled in the same way by dividing \mathbb{R} as follows

$$S_+(\epsilon) = \{r: r \in \mathbb{R}, \epsilon > v(r) > 0\},$$

$$S_-(\epsilon) = \mathbb{R} \setminus S_+(\epsilon).$$

The first set leads to a positive $n_+^s(\epsilon, \Sigma)$, the second one to a negative $n_-^s(\epsilon, \Sigma)$. The rest of the proof goes through unchanged.

Global sum rules result immediately by combining Eq. (4.17) of Lemma 8 with the two integral identities (3.4) and (3.17). The result is

Theorem 3: Assume that H is integrable and constitutes a scattering system, and that v is bound-state limited. Take $u = N + s + 1$ for n even or $u = N + s + 1/2$ for n odd and set $\lambda = \min(s, j)$. If $v \in L^\lambda(\mathbb{R}^n) \cap L^u(\mathbb{R}^n)$, then the sum rules (3.15) and (3.23) are valid for $\Sigma = \mathbb{R}^n$. Furthermore, the integrals over the bound-state density $n(\epsilon, \Sigma)$ that appear in the sum rules are finite.

V. HIGH-TEMPERATURE EXPANSION OF THE VIRIAL COEFFICIENT

This section discusses one application that results from the set of sum rules found in Sec. IV. It is shown that these sum rules determine completely the high-temperature expansion of the classical second virial coefficient, exactly as the quantum sum rules do for the quantum virial coefficient.⁵ Throughout the derivation, this unifying feature of the classical and the quantum problem is clearly exposed. We restrict our attention to the case of particles moving in three dimensions. As is well known, the quantum virial coefficient $a_2(\beta)$ has a closed form representation in terms of the derivative of the phase shift.¹¹ This Beth-Uhlenbeck form of $a_2(\beta)$ is really a time delay representation because the phase shift derivative is proportional to the quantum time delay. So, in general, the quantum virial coefficient may then be written¹²

$$a_2(\beta) = -2^{1/2} \lambda^3 \left\{ \sum_{i=1}^{\infty} d_i e^{-\beta \epsilon_i} + (1/2\pi) \int_0^{\infty} d\epsilon e^{-\beta \epsilon} \text{tr}q(\epsilon) \right\}. \quad (5.1)$$

Here $\beta = (kT)^{-1}$ and $\lambda = (\hbar^2 2\pi\beta/m)^{1/2}$. The variable λ is the thermal wavelength of a particle with mass m in a gas of temperature T . The discrete boundstate spectrum is $\{\epsilon_i\}$ and d_i denotes the degeneracy of the i th energy level. For a spherically symmetric potential the time delay sum $\text{tr}q(\epsilon)$ is

$$\text{tr}q(\epsilon) = 2 \sum_{l=0}^{\infty} (2l+1) \frac{d}{d\epsilon} \delta_l(\epsilon),$$

where $\delta_l(\epsilon)$ is the l th partial wave phase shift.

It is of interest to obtain the classical counterpart of Eq. (5.1) wherein the classical time delay of Sec. II is employed. The classical second virial coefficient $a_2^c(\beta)$ is the phase space integral¹³

$$a_2^c(\beta) = -2^{1/2} \lambda^3 h^{-3} \int d^3r d^3p e^{-\beta p^2/2\mu} (e^{-\beta v(r)} - 1),$$

where μ is the reduced mass $m/2$. Changing the variable p to the energy basis, one finds

$$a_2^c(\beta) = -2^{1/2} \lambda^3 (4\pi 2^{1/2} \mu^{3/2} h^{-3}) \times \int d\epsilon e^{-\beta \epsilon} \int d^3r [(\epsilon - v(r))_+^{1/2} - (\epsilon)_+^{1/2}].$$

However, the integrand in this expression is just the density of states on the classical phase space, discussed in Sec. II, and is thus proportional to $(\partial/\partial \epsilon) \Delta \Gamma(\epsilon, \mathbb{R}^3)$. Using the spectral property (4.17) gives us

$$a_2^c(\beta) = -2^{1/2} \lambda^3 \left\{ \int_{\epsilon_-}^{\epsilon_+} d\epsilon e^{-\beta \epsilon} n(\epsilon, \mathbb{R}^3) + (1/2\pi) \int_0^{\infty} d\epsilon e^{-\beta \epsilon} \text{tr}q(\epsilon, \mathbb{R}^3) \right\}. \quad (5.2)$$

The energies ϵ_+ and ϵ_- represent the largest and smallest values of the energy variable for which the bound-state density, $n(\epsilon, \mathbb{R}^3)$ is nonzero. That the classical virial must have a form like Eq. (5.2) was first realized by Bar-Gadda.¹⁴ He justified Eq. (5.2) on the basis that it is a semiclassical approximation to the quantum solution (5.1). However, as we have just demonstrated, result (5.2) for the classical virial coefficient is an exact result. Another exact expression for $a_2^c(\beta)$ in terms of the classical S matrix has been given by Bassetto, *et al.*¹⁵ Their result is consistent with our time delay solution (5.2).

By comparing Eqs. (5.1) and (5.2) it is seen that when one evaluates the virial coefficients in terms of time delay, the form of the solution is the same in both the classical and quantum cases. So both equations may be summarized by

$$a_v(\beta) = -2^{1/2} \lambda^3 \left\{ \int_{\epsilon_-}^{\epsilon_+} d\epsilon e^{-\beta \epsilon} n_v(\epsilon) + \int_0^{\infty} d\epsilon e^{-\beta \epsilon} \frac{\text{tr}q_v(\epsilon)}{2\pi} \right\} = -2^{1/2} \lambda^3 \{ a_v^b(\beta) + a_v^s(\beta) \}. \quad (5.3)$$

In Eq. (5.3) we have decomposed $a_v(\beta)$ into its bound state and scattering parts and taken out the common factor $-2^{1/2} \lambda^3$. The index v is either c or q , which indicates the classical or quantum case, respectively. In the quantum case, the bound-state density $n_v(\epsilon)$ is the series of δ functions

$$n_q(\epsilon) = \sum_i d_i \delta(\epsilon - \epsilon_i),$$

while classically

$$n_c(\epsilon) = n(\epsilon, \mathbb{R}^3).$$

The classical and quantum time delays $\text{tr}q_v(\epsilon)$ both satisfy a set of sum rules. Theorem 3 states the set in the classical case. The quantum analogue of this family has been obtained by Buslaev¹⁶ and Bollé.⁵ In both the quantum and classical case the sum rules assume a common form. The rule of order N is

$$\int_0^\infty d\epsilon \epsilon^N \left\{ \frac{\text{tr}q_v(\epsilon)}{2\pi} + \sum_{j=1}^{N+1} r_v(j) \epsilon^{1/2-j} \right\} = - \int_{\epsilon_1}^{\epsilon_2} d\epsilon \epsilon^N n_v(\epsilon). \quad (5.4)$$

The subtraction coefficients $r_v(j)$ are closely related in the two cases. One has for the first few terms

$$r_q(j) = r_c(j), \quad j = 1, 2,$$

$$r_q(3) = r_c(3) + (1/128\pi^2) (2\mu/\hbar^2)^{1/2} \int d^3r [\nabla v(r)]^2.$$

In fact the difference in value between $r_q(j)$ and $r_c(j)$ may be obtained from a simple recursion relation⁵ found by Perelomov¹⁷ in his study of the connection between the spectrum of the Schrödinger equation and Korteweg-de Vries type invariants. All formula for the difference $r_q(j) - r_c(j)$, $j \geq 3$, involve gradients of the potential and arise from the fact that v does not commute with the quantum kinetic energy operator.

Our principal observation is that the set of sum rules (5.4) determines the small β (high T) expansion of the general virial $a_v(\beta)$. To see how this comes about we use the method of Ref. 5. Note that $a_v^s(\beta)$ may be written

$$a_v^s(\beta) = -r_v(1) \sqrt{\pi} \beta^{-1/2} + \int_0^\infty d\epsilon \times e^{-\beta\epsilon} [\text{tr}q_v(\epsilon)/2\pi + r_v(1) \epsilon^{-1/2}].$$

The integral in this last equality can be expressed as

$$\begin{aligned} & - \int_0^\infty d\epsilon e^{-\beta\epsilon} \frac{d}{d\epsilon} \int_{\epsilon_1}^\infty d\epsilon_1 [\text{tr}q_v(\epsilon_1)/2\pi + r_v(1) \epsilon_1^{-1/2}] \\ & = - \int_{\epsilon_1}^{\epsilon_2} d\epsilon n_v(\epsilon) - \beta \int_0^\infty d\epsilon e^{-\beta\epsilon} d\epsilon \int_{\epsilon_1}^\infty d\epsilon_1 \\ & \quad \times [\text{tr}q_v(\epsilon_1)/2\pi + r_v(1) \epsilon_1^{-1/2}]. \end{aligned}$$

The right-hand side is obtained by an integration by parts. The $N = 0$ rule (5.4) has been used to evaluate the $\epsilon = 0$ portion of the surface term. This process may be repeated. For example, the last integral above is also

$$\beta \int_0^\infty d\epsilon e^{-\beta\epsilon} \frac{d}{d\epsilon} \int_{\epsilon_1}^\infty d\epsilon_2 \int_{\epsilon_2}^\infty d\epsilon_1 \times [\text{tr}q_v(\epsilon_1)/2\pi + r_v(1) \epsilon_1^{-1/2}].$$

By adding $r_v(2) \epsilon_1^{-3/2}$ to the square bracket term this integral can be computed with the help of an integration by parts and the use of the $N = 1$ rule to determine the nonvanishing part of the surface term. The entire series is

$$\begin{aligned} a_v^s(\beta) & = \sum_{j=1}^\infty \frac{2^{j-1} (-1)^j \sqrt{\pi}}{(2j-3)!!} r_v(j) \beta^{j-3/2} \\ & \quad - \sum_{j=0}^\infty \int_{\epsilon_1}^{\epsilon_2} d\epsilon n_v(\epsilon) \frac{(-\beta\epsilon)^j}{j!}, \end{aligned}$$

with the convention $(-1)!! = 1$. The second sum here is just the negative of $a_v^s(\beta)$; thus $a_v(\beta)$ has the expansion

$$a_v(\beta) = -2^{1/2} \lambda^3 \sum_{j=1}^\infty \frac{2^{j-1} (-1)^j \sqrt{\pi}}{(2j-3)!!} r_v(j) \beta^{j-3/2}. \quad (5.5)$$

This last form gives us the high-temperature expansion of the virial coefficient. The only unknown constants entering this expansion are the $r_v(j)$, the constants that appear in the sum rule family (5.4). Thus it is seen that the sum rule family completely determines the high-temperature expansion of $a_v(\beta)$, and this conclusion is valid in both the quantum and classical cases. Furthermore, in both cases we see an explicit cancellation between the bound state contributions and scattering contributions in this high-temperature expansion.

The sum rules for single channel scattering theory, discussed here, rely predominantly on the spectral property of time delay. This property is known to be valid in both classical and quantum scattering. It is reasonable to expect that similar sum rules are realized in multichannel few-particle collisions and that they will again provide a method for obtaining unified (classical and quantum) high-temperature expansions of the few particle virial coefficients.

¹N. Levinson, Kgl. Dan. Vidensk. Selsk. Mat. Fys. Medd. **25**, 1 (1949); R. G. Newton, *Scattering Theory of Waves and Particles* (McGraw-Hill, New York, 1966), Sec. 11; J. M. Jauch, *Helv. Phys. Acta* **30**, 143 (1957).

²T. A. Osborn, R. G. Froese, and S. F. Howes, *Phys. Rev. A* **22**, 101 (1980).

³V. S. Buslaev, *Sov. Phys. Dokl.* **7**, 295 (1962).

⁴T. A. Osborn and D. Bollé, *J. Math. Phys.* **18**, 432 (1977).

⁵D. Bollé and H. Smeesters, *Phys. Lett. A* **62**, 290 (1977); D. Bollé, *Ann. Phys. (N.Y.)* **121**, 131 (1979).

⁶R. G. Newton, *J. Math. Phys.* **18**, 1348 (1977); T. Dreyfus, *J. Phys. A* **9**, 1588 (1976); T. Dreyfus, *Helv. Phys. Acta* **51**, 321 (1978).

⁷L. W. MacMillan and T. A. Osborn, *Ann. Phys. (N.Y.)* **126**, 1 (1980).

⁸W. Hunziker, *Commun. Math. Phys.* **8**, 282 (1968); W. Hunziker, in *Scattering in Mathematical Physics*, edited by J. A. La Vita and J.-P. Marchand (Reidel, Boston, 1974), p. 79.

⁹H. Goldstein, *Classical Mechanics* (Addison-Wesley, Reading, Mass., 1959), Chap. 9.

¹⁰I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1972), p. 21.

¹¹G. E. Uhlenbeck and E. Beth, *Physica* **3**, 729 (1936); E. Beth and G. E. Uhlenbeck, *Physica* **4**, 915 (1937); L. Gropper, *Phys. Rev.* **50**, 963 (1936).

¹²T. A. Osborn and T. Y. Tsang, *Ann. Phys. (N.Y.)* **101**, 119 (1976).

¹³K. Huang, *Statistical Mechanics* (Wiley, New York, 1964), Chap. 14.

¹⁴U. Bar-Gadda, *Physica* **62**, 321 (1972).

¹⁵A. Bassetto, R. Soldati, M. Toller, and S. Zerbini, *Lett. Math. Phys.* **1**, 401 (1977).

¹⁶V. S. Buslaev, in *Topics in Mathematical Physics*, edited by M. Sh. Birman (Consultants Bureau, New York, 1967), Vol. 1, p. 69.

¹⁷A. M. Perelomov, *Ann. Inst. Henri Poincaré* **24**, 161 (1976).

Converging lower bounds to atomic binding energies ^{a)}

S. R. Singh

Department of Chemistry, York University, 4700 Keele Street, Downsview, Ontario, M3J 1P3, Canada

(Received 10 June 1980; accepted for publication 28 August 1980)

A method is proposed to obtain approximations converging from below to a finite number of the nonrelativistic binding energies of atomic systems. The method requires that the Hamiltonian be decomposable as a sum of an unperturbed part and a non-negative perturbation. The eigenvalues and the eigenvectors of the unperturbed part are assumed to be known. For computational purposes one needs the matrix elements of the square of the Hamiltonian, in addition to those of the Hamiltonian itself. These elements are used to construct a matrix valued function whose eigenvalues have the bounds as their fixed points. The elements of the matrix are obtained by solving a system of linear equations typical of variational methods. An iterative procedure is shown to yield converging lower bounds to the fixed points and thus to the binding energies.

PACS numbers: 31.15. + g, 03.65.Ge, 02.60. + y

1. INTRODUCTION

The interest in devising methods to compute the lower bounds to the atomic binding energies stems from the fact that while it is feasible to compute converging upper bounds, e.g., using the Rayleigh–Ritz method, their accuracy remains undetermined. Thus lower bounds would enable one to determine the binding energies with a known degree of accuracy. As a result several lower bound formulas have been derived.^{1–3} However, most of these formulas lack the desired convergence property,^{1,2} some of them are valid only for the ground state energy,³ and some require an unrealistic input, e.g., the exact binding energy.² Computationally, most of the methods are much more difficult to use than the methods for the upper bounds and yield poorer results.

In the present note we develop a method which enables one to compute converging lower bounds with realistic input. Because of the convergence property, the accuracy of the bounds can be improved to an arbitrary degree. A major requirement of the method is that the Hamiltonian be decomposable as a sum of an unperturbed part and a non-negative perturbation. The discrete spectrum of the unperturbed part, as well as that of the total Hamiltonian, is assumed to be contained in the negative real line. Also the eigenvalues and the eigenprojections of the unperturbed part are explicitly used and therefore are assumed to be known. In Sec. 2, we show that the binding energies are the unique fixed points of the eigenvalues of a matrix valued function. A function was obtained with the same property in Ref. 3 but only for the ground state energy. However the present result reduces to a different one even in this case. In Sec. 3, we show that a matrix can be constructed by solving a set of algebraic equations which approximates the previous matrix, and that the fixed points of its eigenvalues converge from below to the binding energies. Here we need the matrix elements of the square of the atomic Hamiltonian, in addition to those of the Hamiltonian itself. Most of the methods known require this input.^{1,2} In Sec. 4, we show that an iterative procedure can be

used to approximate the lower bounds of Sec. 3 from below, thus yielding sequences of lower bounds that converge to the binding energies themselves. It is indicated in Sec. 5, that the present method also yields an approximate eigenvector.

2. SOME PRELIMINARY RESULTS

Let $H = H_0 + V$, $V \geq 0$ be the Hamiltonian representing a nonrelativistic atomic system with the center of mass part removed, and \mathcal{H} , the underlying Hilbert space equipped with the scalar product (\cdot, \cdot) . We assume that the eigenvalues E_0^j of H_0 and the corresponding normalized eigenvectors ψ_0^j , $j = 0, 1, 2, \dots$; are known and that $-\infty < E_0^j < 0$ for each j . This can be achieved by choosing, for example, H_0 to be the hydrogenlike part of H . Let $\{E_0^j\}$ be ordered in a nondecreasing manner counting multiplicities and $\{\psi_0^j\}$ according to the order induced by $\{E_0^j\}$. If some of the eigenvalues are degenerate, some members of $\{E_0^j\}$ will be identical. In such a case the following analysis remains valid if the corresponding eigenvectors are taken to be any orthonormal set spanning the underlying subspace. Therefore we shall not distinguish this case from the one when all the eigenvalues have multiplicity equal to one.

Let the eigenvalues and the corresponding eigenvectors of H be denoted by $\{E^j\}$ and $\{\psi^j\}$, respectively. We shall obtain lower bounds to E^j for $j = 0, 1, \dots, J$; such that $E_0^0 \leq E^j < E_u < E_0^{J_0+1}$ with J_0 and E_u being arbitrary except for the stated inequality. Since E_u can be chosen arbitrarily close to $E_0^{J_0+1}$, the procedure enables one to obtain lower bounds to all $E^j < E_0^{J_0+1}$ and since J_0 is arbitrary, to any finite number of $E^j < 0$. Further, let p_0^j be the projection $\psi_0^j(\psi_0^j, \cdot)$. It is clear that $H_0' = H_0(1 - \sum_{j=0}^{J_0} p_0^j) \geq E_0^{J_0+1}$ and $A = H_0' + V - E_u \geq E_0^{J_0+1} - E_u > 0$.

Consider the $(J_0 + 1) \times (J_0 + 1)$ Hermitian matrix valued continuously differentiable function $\mathcal{A}(x)$ of $x \in (-\infty, E_u)$ given by $\mathcal{A}(x) = \tilde{\mathcal{A}}(x) - (\epsilon - E_u I)$, where I is the identity matrix, ϵ is the matrix with elements $\epsilon_{ij} = -E_0^j \delta_{ij}$, and $\tilde{\mathcal{A}}(x) = \epsilon^{1/2} \tilde{\mathcal{A}}(x) \epsilon^{1/2}$ with $\tilde{\mathcal{A}}(x)$ being the matrix with elements $\tilde{\mathcal{A}}(x)_{ij} = (A \psi_0^i, [(E_u - x)A + A^2]^{-1} A \psi_0^j)$. Let $\lambda^j(x)$,

^{a)}Supported in part by the NSERC grant No. A3604.

$j = 0, 1, \dots, J_0$; be the eigenvalues of $\mathcal{A}(x)$ which are all real.

Lemma 1: An $x_0 \in (-\infty, E_u)$ is an eigenvalue of H if and only if it is a fixed point of $\lambda^j(x)$ for some j .

Remark: By definition a fixed point x_0 of $\lambda^j(x)$ is a solution of $\lambda^j(x_0) = x_0$.

Proof: If x_0 is an eigenvalue of H , there is a vector ψ such that

$$(E_u - x_0 + A)\psi = - \sum_{j=0}^{J_0} E_j^0 \bar{\alpha}_j \psi_0^j, \quad (1)$$

where $\bar{\alpha}_j \psi_0^j = p_0^j \psi$. Since $x_0 < E_u$ and $A > 0$, we have that $(E_u - x_0 + A) > 0$. Therefore $(E_u - x_0 + A)^{-1}$ exists as a bounded operator and hence

$$\psi = - \sum_{j=0}^{J_0} E_j^0 \bar{\alpha}_j (E_u - x_0 + A)^{-1} \psi_0^j. \quad (2)$$

The identity $(E_u - x_0)(E_u - x_0 + A)^{-1} = 1 - A(E_u - x_0 + A)^{-1}$ and some algebraic manipulation yields

$$(\epsilon - \tilde{\mathcal{A}}(x_0)\epsilon - E_u I)\bar{\alpha} = -x_0 \bar{\alpha}, \quad (3)$$

where $\bar{\alpha}$ is the column vector with components $\bar{\alpha}_j$, $j = 0, 1, \dots, J_0$; and the other symbols are as defined above.

Setting $\alpha = \epsilon^{1/2} \bar{\alpha}$ we have that

$$\mathcal{A}(x_0)\alpha = x_0 \alpha, \quad (4)$$

i.e., $\mathcal{A}(x_0)$ has x_0 as its eigenvalue. Since $\mathcal{A}(x)$ is continuous, $\lambda^j(x)$ for each j , is continuous for $x \in (-\infty, E_u)$. Hence there must be $\lambda^j(x)$ for some j with x_0 as its fixed point.

Now, if there is a $\lambda^j(x)$ which has x_0 as its fixed point there must be an α such that (4) is satisfied and hence an $\bar{\alpha}$ satisfying (3). Define ψ by (2) with $\bar{\alpha}_j$ so obtained. Clearly ψ satisfies (1) and the proof will be complete if $\bar{\alpha}_j = (\psi_0^j, \psi)$ for then (1) is equivalent to the original eigenvalue equation.

This can be seen from the following.

Let $\bar{\beta}$ be the vector with elements (ψ_0^j, ψ) . It is easy to check that

$$(E_u - x_0)\bar{\beta} = (\epsilon - \tilde{\mathcal{A}}(x_0)\epsilon)\bar{\alpha}.$$

From (3) the right member is equal to $(E_u - x_0)\bar{\alpha}$ and since $E_u - x_0 \neq 0$, $\bar{\beta} = \bar{\alpha}$.

From Lemma 1 it is clear that the study of E^j , $j = 0, 1, \dots, J$, reduces to studying the fixed points of $\lambda^j(x)$, $j = 0, 1, \dots, J_0' \leq J_0$. In the following we establish a one to one correspondence between E^j and $\lambda^j(x)$ for each $j = 0, 1, \dots, J$. It is convenient, here, to introduce another continuously differentiable $(J_0 + 1) \times (J_0 + 1)$ matrix valued function $\mathcal{B}(x)$ of $x \in (-\infty, E_u)$: $\mathcal{B}(x) = \epsilon^{1/2} \tilde{\mathcal{B}}(x) \epsilon^{1/2}$, where $\tilde{\mathcal{B}}(x)_{ij} = (\psi_0^i, (E_u - x + A)^{-1} \psi_0^j)$. Since $\mathcal{B}(x)$ is Hermitian its eigenvalues $\beta^j(x)$, $j = 0, 1, \dots, J_0$; are real. Furthermore, it is straightforward to check that $\mathcal{A}(x)\alpha(x) = \lambda^j(x)\alpha(x)$ if and only if $\mathcal{B}(x)\alpha(x) = \beta^j(x)\alpha(x) = \{[E_u - \lambda^j(x)]/(E_u - x)\}\alpha(x)$. It is clear that the fixed points of $\lambda^j(x)$ in $(-\infty, E_u)$ i.e., for $j = 0, 1, \dots, J_0'$; are given by $\beta^j(x) = 1$, $j = 0, 1, \dots, J_0'$.

Lemma 2: For each j , $\lambda^j(x)$ has at most one fixed point in $(-\infty, E_u)$.

Proof: For each j , it follows from the Hellmann-Feynman theorem that

$$\frac{d\beta^j(x)}{dx} = \langle \alpha, \frac{d\mathcal{B}(x)}{dx} \alpha \rangle, \quad \langle \alpha, \alpha \rangle = 1,$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product in C^{J_0+1} , the space of $(J_0 + 1)$ -column vectors, and α is an eigenvector of $\mathcal{B}(x)$ corresponding to the eigenvalue $\beta^j(x)$. Substituting for $\mathcal{B}(x)$ one obtains that

$$\frac{d\beta^j(x)}{dx} = \langle u, [E_u - x + A]^{-2} u \rangle \geq 0,$$

where $u = \sum_{i=0}^{J_0} \alpha_i(x) \epsilon_i^{1/2} \psi_0^i$. But the equality implies that $(E_u - x + A)^{-1} u = 0$, i.e., $u = 0$, which in turn implies that $\alpha = 0$. Therefore $d\beta^j(x)/dx > 0$ and, hence, $\beta^j(x)$ is a strictly increasing function of $x \in (-\infty, E_u)$. Therefore $\beta^j(x) = 1$ can have at most one solution which implies the result.

It is obvious from Lemmas 1 and 2 that for $j = 0, 1, \dots, J$; E^j is the fixed point of one and only one $\lambda^j(x)$ and thus $J_0' = J$. For some values of j , $\lambda^j(x)$ may not have any fixed point in $(-\infty, E_u)$. This will happen if and only if J_0 is strictly greater than J , i.e., if and only if some of the perturbed eigenvalues have crossed the next unperturbed levels.

3. THE LOWER BOUNDS TO E^j

The first step in obtaining the lower bounds to E^j is to approximate $\lambda^j(x)$ from below. For this, consider the following set of linear equations

$$\sum_{k=1}^n \alpha_k^l(\phi_l, [(E_u - x)A + A^2] \phi_k) = (\phi_l, A \psi_0^j), \quad (5)$$

$$l = 1, 2, \dots, n,$$

where $\{\phi_k\} \subseteq \mathcal{D}(A)$ with $\mathcal{D}(\cdot)$ denoting the domain. Define a new scalar product $(\cdot, \cdot)_+$ on $\mathcal{D}(A)$ by $(u, v)_+ = (Au, Av)$, $u, v \in \mathcal{D}(A)$ and complete $\mathcal{D}(A)$ with respect to $(\cdot, \cdot)_+$ to obtain $\tilde{\mathcal{D}}(A) = \mathcal{H}_+ \subseteq \mathcal{H}$. Terms like $(u, A^2 v)$ are to be interpreted as (Au, Av) . The set of equations given by (5) takes the following form

$$\sum_{k=1}^n \alpha_k^l(x) (\phi_l, [1 + (E_u - x)B] \phi_k)_+ = (\phi_l, B \psi_0^j)_+, \quad (6)$$

$$l = 1, 2, \dots, n,$$

where B is the closure in \mathcal{H}_+ of B' defined by $(u, Av) = (u, B'v)_+$, $u \in \mathcal{H}$, $v \in \mathcal{D}(A)$. So defined, B is a self-adjoint non-negative bounded operator from \mathcal{H}_+ to \mathcal{H}_+ . Let $\{\phi_k\}$ be an orthonormal basis in \mathcal{H}_+ . Then (6) takes the following form:

$$[1 + (E_u - x)B_n] f_n^j(x) = P_n B \psi_0^j, \quad (7)$$

where $f_n^j(x) = \sum_{k=1}^n \alpha_k^j(x) \phi_k$, $B_n = P_n B P_n$ and P_n is the orthogonal projection on the subspace of \mathcal{H}_+ spanned by ϕ_k , $k = 1, 2, \dots, n$.⁴

Lemma 3: Let the symbols be as above. Then

(i) $f_n^j(x) \rightarrow (E_u - x + A)^{-1} \psi_0^j$ in \mathcal{H}_+ as well as in \mathcal{H} ,

(ii) let $f_n(x)$ be the solution of (7) with ψ_0^j replaced by an arbitrary $\phi \in \mathcal{H}_+$. Then $(A\phi, f_n(x)) \uparrow (A\phi, f(x))$, where $f(x) = (E_u - x + A)^{-1} \phi$.

Proof: The proofs of the stated results are standard. For example (i) is a special case of Theorem 1 of Ref. (4); the

convergence in (ii) follows from (i) with obvious replacement and the bound property follows from the easily derivable equality

$$(f - f_n, A\phi) = (f - f_n, B\phi)_+ \\ = (f - f_n, [1 + (E_u - x)B](f - f_n))_+ \geq 0.$$

Now, let $\mathcal{A}_n(x) = \tilde{\mathcal{A}}_n(x) - (\epsilon - E_u I)$, where $\tilde{\mathcal{A}}_n(x) = \epsilon^{1/2} \tilde{\mathcal{A}}_n(x) \epsilon^{1/2}$ with $\tilde{\mathcal{A}}_n(x)$ being the matrix with elements $(A\psi_0^i, f_n^j(x))$, $i, j = 0, 1, \dots, J_0$. Denote by $\lambda_n^j(x)$, $j = 0, 1, \dots, J_0$; the eigenvalues of $\mathcal{A}_n(x)$. In Theorem 1 we obtain lower approximations to $\lambda^j(x)$ and thus to E^j .

Theorem 1: With notation as above,

(i) $\lambda_n^j(x) \uparrow \lambda^j(x)$, $j = 0, 1, \dots, J_0$; $x \in (-\infty, E_u)$,

(ii) For each j such that $\lambda_n^j(E_u) < E_u$, $\lambda_n^j(x)$ has a unique fixed point $E_n^j \in (-\infty, E_u)$ and $E_n^j \uparrow E^j$ or $E_n^j \uparrow E_u$.

Proof: (i) For any $\alpha \in C^{J_0+1}$ with components α_j we have that

$$\langle \alpha, (\mathcal{A}(x) - \mathcal{A}_n(x))\alpha \rangle = \langle \epsilon^{1/2} \alpha, (\tilde{\mathcal{A}}(x) - \tilde{\mathcal{A}}_n(x))\epsilon^{1/2} \alpha \rangle \\ = (A\phi, f(x) - f_n(x)) \downarrow 0 \\ \text{[Lemma 3 (ii)]}$$

where $\phi = \sum_{i=0}^{J_0} \alpha_i \epsilon_i^{1/2} \psi_0^i$ and $f(x), f_n(x)$ are as defined in Lemma 3 (ii). The convergence of $\mathcal{A}_n(x)$ to $\mathcal{A}(x)$ from below in conjunction with the Hellmann-Feynman theorem implies that $\lambda_n^j(x) \uparrow \lambda^j(x)$, $j = 0, 1, \dots, J_0$; $x \in (-\infty, E_u)$.

(ii) Let $\beta_n^j(x) = [E_u - \lambda_n^j(x)] / (E_u - x)$. As in the case of $\beta^j(x)$, it follows from direct substitution that $\beta_n^j(x)$, for each j , is an eigenvalue of $\mathcal{B}_n(x) = \epsilon^{1/2} \tilde{\mathcal{B}}_n(x) \epsilon^{1/2}$, where $\tilde{\mathcal{B}}_n(x)$ is a matrix with elements $(B\psi_0^i, B_n^{1/2} [1 + (E_u - x)B_n]^{-1} B_n^{1/2} B\psi_0^j)_+$. Since $B \geq 0$, $B_n \geq 0$ and $B_n^{1/2}$ is well defined. Also the eigenvector α_n of $\mathcal{B}_n(x)$ corresponding to the eigenvalue $\beta_n^j(x)$ is the same as that of $\mathcal{A}_n(x)$ corresponding to the eigenvalue $\lambda_n^j(x)$. The facts that $\beta^j(x) \geq 0$ and $d\beta^j(x)/dx > 0$ (Lemma 2) imply that $\beta_n^j(x) > 0$. Since $\lambda_n^j(x) \leq \lambda^j(x)$ we have that $\beta_n^j(x) \geq \beta^j(x) > 0$.

Now

$$\frac{d\beta_n^j(x)}{dx} = \langle \alpha_n, \frac{d\mathcal{B}_n(x)}{dx} \alpha_n \rangle, \quad \langle \alpha_n, \alpha_n \rangle = 1 \\ = (Bu_n, B_n [1 + (E_u - x)B_n]^{-2} B_n Bu_n)_+ \geq 0,$$

where $u_n = \sum_{i=0}^{J_0} (\alpha_n(x))_i \epsilon_i^{1/2} \psi_0^i$. If $d\beta_n^j(x)/dx = 0$, then $[1 + (E_u - x)B_n]^{-1} B_n Bu_n = 0$ and hence

$$\beta_n^j(x) = (Bu_n, B_n^{1/2} [1 + (E_u - x)B_n]^{-1} B_n^{1/2} Bu_n) = 0.$$

This contradicts the result that $\beta_n^j(x) > 0$. Hence $d\beta_n^j(x)/dx > 0$.

It follows that $\beta_n^j(x)$ is an increasing function of $x \in (-\infty, E_u)$. Also, since $\mathcal{B}_n(-\infty) = 0$, $\beta_n^j(-\infty) = 0$ and since $\lambda_n^j(E_u) < E_u$, $\lim_{x \rightarrow E_u} \beta_n^j(x) = \infty$.

Hence $\beta_n^j(x) = 1$ has a unique solution $x = E_n^j < E_u$. It is clear that E_n^j is the fixed point of $\lambda_n^j(x)$.

Now, since $\lambda_n^j(x) \uparrow \lambda^j(x)$, $\beta_n^j(x) \downarrow \beta^j(x)$. This together with the continuity of $\beta^j(x)$ implies that E_n^j converges from below to the solution of $\beta^j(x) = 1$. If this solution is restricted to $(-\infty, E_u)$, then it is also the solution of $\lambda^j(x) = x$ i.e., it is E^j implying that if E_n^j does not converge to E_u then $E_n^j \uparrow E^j$.

Remark: (i) Since $\beta_n^j(x) \downarrow \beta^j(x)$ for each j the set $\{E_n^j\}$

cannot be smaller than the set $\{E^j\}$. If an upper bound to E^j is known then any E_n^j greater than the bound is irrelevant. However if $J = J_0$, this cannot happen.

(ii) $\lambda_n^j(E_u)$ cannot be greater than E_u for this will mean that $\beta_n^j(x) < 0$ for some $x < E_u$ which contradicts the non-negativity of $\mathcal{B}_n(x)$. The case $\lambda_n^j(E_u) = E_u$ is irrelevant for that will imply that $\lambda^j(x)$ has a fixed point equal to (or greater than) E_u .

4. LOWER BOUNDS TO E_n^j

From Theorem 1 the problem of obtaining the lower bounds reduces to evaluating the fixed points of $\{\lambda_n^j(x)\}$ in $(-\infty, E_u - \delta)$ with some $\delta > 0$. A value for δ can be determined by an upper bound E_u^j to E^j , e.g., $(E_u - \delta)$ may be taken to be E_u^j . If $J = J_0$, the fixed points of all of the $\lambda_n^j(x)$ are of interest. However, the exact evaluation of E_n^j is not likely to be possible in most cases of interest. In the following we show that the usual iterative method can be used to obtain a sequence $\{E_{nm}^j\}$ for each j and n such that $E_{nm}^j \uparrow E_n^j$. Thus $\{E_{nm}^j\}$ will provide converging lower bounds to E^j which can be computed numerically.

Lemma 4: For $x \leq E_n^j$ we have that

$$0 \leq d\lambda_n^j(x)/dx < \beta_n^j(x) \leq 1.$$

Proof: As in the case of $\beta_n^j(x)$, it is easy to check that $d\lambda_n^j(x)/dx$ is non-negative.

Now we have that

$d\beta_n^j(x)/dx = (d/dx)[E_u - \lambda_n^j(x)] / (E_u - x) > 0$ [Theorem 1 (ii)] which reduces to $(d/dx)\lambda_n^j(x) < \beta_n^j(x)$. Further, for $x \in (-\infty, E_n^j)$, $\beta_n^j(x) < 1$ otherwise $\beta_n^j(x)$ being an increasing continuous function will assume the value unity for some $x \in (-\infty, E_n^j)$. This implies the result.

After the fact that $d\lambda_n^j(x)/dx < 1$ has been established, a proof of the convergence of the iterative method is a standard procedure. However, we give a proof here in order to show, in addition, the bound property of the resulting sequence.

Theorem 2: Let $x_0 \leq E_n^j$, $x_{m+1} = \lambda_n^j(x_m)$, $m = 0, 1, 2, \dots$. Then

$$x_m = E_{nm}^j \uparrow E_n^j.$$

Proof: Using the fact that $\lambda_n^j(E_n^j) = E_n^j$ we obtain

$$E_n^j - x_{m+1} = \int_{x_m}^{E_n^j} \frac{d\lambda_n^j(x)}{dx} dx.$$

Since $0 \leq d\lambda_n^j(x)/dx < 1$, $x_m \leq E_n^j$ implies that $0 \leq E_n^j - x_{m+1} \leq E_n^j - x_m$, i.e., $x_m \leq x_{m+1} \leq E_n^j$. From this and the induction principle, it follows that $\{x_m\}$ is a nondecreasing sequence bounded by E_n^j and hence must converge from below to a limit $\bar{x} \leq E_n^j$. But $\bar{x} = \lim_{m \rightarrow \infty} x_{m+1} = \lim_{m \rightarrow \infty} \lambda_n^j(x_m) = \lambda_n^j(\bar{x})$, i.e., \bar{x} is a fixed point of $\lambda_n^j(x)$ and since λ_n^j has E_n^j as the unique fixed point, $\bar{x} = E_n^j$.

5. CONCLUDING REMARKS

In order to compute the lower bounds E_{nm}^j by using the present method, one needs a starting lower bound. This is not a serious limitation for crude lower bounds can easily be

estimated, e.g., E_0^j will serve as one for E_n^j or E_0^0 may be used as a starting bound for each E_n^j if so desired. If in Theorem 2 the starting value $x_0 \geq E_n^j$, the proof can be easily modified to show that the resulting $x_m \downarrow E_n^j$. This result, although not very interesting, may nevertheless be used to check the accuracy of E_{nm}^j . Further, it is necessary in the present method to solve the set of equations given by (5) for each x_m starting with x_0 . The value $x_{m+1} = \lambda_n^j(x_m)$ is then obtained by constructing the matrix $\mathcal{A}_n(x_m)$ from the solution and then diagonalizing it. This procedure has to be repeated for each j .

Since $x_m \xrightarrow{m \rightarrow \infty} \bar{x}$ and $f_n^j(\bar{x}) \xrightarrow{n \rightarrow \infty} (E_u - \bar{x} + A)^{-1} \psi_0^j$ (Lemma 3) the proof of Lemma 3(i) can easily be extended to conclude that $f_n^j(x_m) \xrightarrow{n, m \rightarrow \infty} (E_u - \bar{x} + A)^{-1} \psi_0^j$. Thus an approximation to ψ^j is given by $\psi_{nm}^j = \sum_{i=0}^{j_0} \alpha_{nmi}^j \epsilon_{ii}^{1/2} f_n^i(x_m)$, where α_{nmi}^j is the i th component of the eigenvector α_{nm}^j of $\mathcal{A}_n(x_m)$ corresponding to the eigenvalue $\lambda_n^j(x_m)$. The convergence of ψ_{nm}^j to ψ^j requires, further, the fact that

$\alpha_{nm}^j \xrightarrow{n, m \rightarrow \infty} \alpha^j$ where α^j is the eigenvector of $\mathcal{A}(\bar{x})$, corre-

sponding to the eigenvalue $\lambda^j(x)$. Thus the present method enables one to approximate also the eigenvector ψ^j when E^j is nondegenerate. In the case when E^j is degenerate it is easy to modify the computing procedure slightly to obtain an approximation to the corresponding eigenprojection.

If E_0^0 is nondegenerate and $E_0^0 < E^0 < E_0^1$ some simplifications occur. In that case if one chooses $E_u < E_0^1$, $\mathcal{A}(x)$ is a single function rather than a matrix with functions as entries. As a consequence the problem of obtaining the eigenvalues and eigenvectors of $\mathcal{A}_n(x)$ is nonexistent.

ACKNOWLEDGMENT

Thanks are due to Professor A. D. Stauffer for helpful comments and to Professor H. O. Pritchard for his hospitality.

¹J. G. Leopold, M. Cohen, and J. Katriel, J. Phys. B 8, 513 (1975) and references cited therein.

²M. Cohen, R. P. McEachran, S. Cameron, and J. A. Stauffer, J. Phys. B 4, 1109 (1971).

³P. O. Lowdin, Phys. Rev. A 139, 357 (1965).

⁴S. R. Singh, J. Math. Phys. 18, 1466 (1977).

⁵S. R. Singh and A. D. Stauffer, Nuovo Cimento B 25, 547 (1975).

Thermal blooming calculations with analytical diffraction approximated expressions

P. Hillion

Institut Henri Poincaré, 75231 Paris, France

S. Quinnez

Faculté des Sciences, Université de Dijon, 21000 Dijon, France

(Received 21 August 1979; accepted for publication 28 March 1980)

Using the analytical results obtained in a forthcoming paper, we discuss here the thermal blooming of collimated and focused laser beams in some simple steady state and transient cases where an approximation of hydrodynamic equations is available. We still obtain tractable expressions which make it possible to plot the iso-intensity curves.

PACS numbers: 42.65.Jx, 42.68.Rp

1. INTRODUCTION

In a forthcoming paper¹ we give the solutions to first order for collimated and focused beams of both equations:

$$\partial^j S(r) \partial_j S(r) = n^2(r) + \frac{1}{K_0^2} \sqrt{\frac{n(r)}{I(r)}} \partial^j \partial_j \sqrt{\frac{I(r)}{n(r)}} \quad (1)$$

and

$$\partial^j S(r) \partial_j \left[\frac{I(r)}{n(r)} \right] + \frac{I(r)}{n(r)} \partial^j \partial_j S(r) = 0, \quad (1')$$

where $S(r)$ is the eikonal, $I(r)$ the intensity of the light beam, $n(r)$ the refractive number, K_0 the wave number when $n(r)$ is

$$n(r) = 1 + \epsilon \mu(r) + o(\epsilon^2), \quad (2)$$

and when the second term on the right-hand side of (1) is approximated by $[1/\sqrt{I_0(r)}] \partial^j \partial_j \sqrt{I_0(r)}$, $I_0(r)$ being the solution of Eqs. (1) and (1') for $K_0 \rightarrow \infty$ and $\epsilon = 0$.

In Eqs. (1) and (1') we use the summation convention and the index j takes the values 1,2,3.

Leaving aside $S(r)$, we found for $I(r)$ in the case of a Gaussian beam propagating along the oz axis with z as parameter

$$\begin{aligned} \hat{I}_f[x(z), y(z), z] &= \frac{1}{\Delta_{2\epsilon}^2(z)} I \left[\frac{x^*(z)}{\Delta_{1\epsilon}(z)}, \frac{y^*(z)}{\Delta_{1\epsilon}(z)}, 0 \right] \\ &\times \exp \left\{ -\epsilon \int_0^z d\xi \int_0^\xi \partial^j \partial_j \mu[x(\rho), y(\rho), \rho] d\rho \right. \\ &\left. + o \left[\left(\epsilon + \frac{1}{K_0^2 a^2} \right)^2 \right] \right\}. \quad (3) \end{aligned}$$

The hat symbol means that diffraction is taken into account to order $(\epsilon + (1/K_0^2 a^2))$, a is a transverse characteristic of the beam. Notations are as follows:

$$\begin{aligned} x_\alpha^*(z) &= x_\alpha(z) - \epsilon v_\alpha(z), \\ v_\alpha(z) &= \int_0^z d\xi \int_0^\xi \partial_\alpha \mu(x(\rho), y(\rho), \rho) d\rho \quad \alpha = 1, 2, \quad (4) \end{aligned}$$

$$\begin{aligned} \Delta_{1\epsilon}^2(z) &= D_0^2(z) + \frac{2\epsilon}{f} (1 - z/f) v_3(z), \\ v_3(z) &= \int_0^z d\xi \int_0^\xi \partial_3 \mu(x(\rho), y(\rho), \rho) d\rho; \quad (5) \end{aligned}$$

$$\begin{aligned} \Delta_{2\epsilon}^2(z) &= D_0^2(z) + 2\epsilon \frac{z}{f} \left(1 - \frac{z}{f} \right) \\ &\times \left[\mu \left(\frac{x(z)}{1 - z/f}, \frac{y(z)}{1 - z/f}, 0 \right) + \frac{v_3(z)}{z} \right], \quad (5') \end{aligned}$$

with

$$D_0^2(z) = (1 - z/f)^2 + z^2/K_0^2 a^4. \quad (6)$$

In the following sections we use the expression (3) to discuss thermal blooming in the atmosphere, first in the easy case of a collimated beam, and then for a converging focused beam either in the near field or in the far field, combining Eq. (3) with linearized hydrodynamics.

For recent surveys on thermal blooming, one can consult Refs. 2 and 3. In particular, Smith³ and Gebhardt and Smith⁵ have previously obtained some of the following results.

2. THERMAL BLOOMING FOR A COLLIMATED BEAM

For a collimated beam, Eqs. (5) and (5') reduce to $\Delta_{1\epsilon}^2(z) = \Delta_{2\epsilon}^2(z) = 1 + (z^2/K_0^2 a^4)$, where $z^2/K_0^2 a^4$ is the correction term due to diffraction, whose effects begin to appear beyond the Rayleigh distance $K_0 a^2$, that is about 6 km for a CO₂ laser with transverse radius $a = 10$ cm. Here we assume $z \ll K_0 a^2$ so that one can neglect diffraction. Then it is shown in Ref. 1 that Eq. (3) becomes

$$\begin{aligned} I_\rho(x, y, z) &= I(x, y, 0) \exp \left[-\epsilon \int_0^z d\xi \int_0^\xi \left(\frac{\partial^\beta I(x, y, 0)}{I(x, y, 0)} \right. \right. \\ &\left. \left. \times \partial_\beta \mu(x, y, \rho) + \partial^j \partial_j \mu(x, y, \rho) \right) d\rho + o(\epsilon^2) \right], \quad (7) \end{aligned}$$

where the index β takes the values 1,2. When absorption is taken into account, one has instead of (1')

$$\partial^j S(r) \partial_j \left[\frac{I(r)}{n(r)} \right] + \frac{I(r)}{n(r)} \partial^j \partial_j S(r) = \alpha I(r), \quad (8)$$

where α is a constant linear absorption coefficient, and now one has

$$\begin{aligned} I_\rho(x, y, z) &= I(x, y, 0) e^{-\alpha z} \exp \left\{ -\epsilon \int_0^z d\xi \int_0^\xi \left[\frac{\partial^\beta I(x, y, 0)}{I(x, y, 0)} \right. \right. \\ &\left. \left. \times \partial_\beta \mu(x, y, \rho) + \partial^j \partial_j \mu(x, y, \rho) \right] d\rho + o(\epsilon^2) \right\}. \quad (9) \end{aligned}$$

Some authors (see, for instance Refs. 3 and 5) have ob-

tained in a heuristic manner the following expression

$$I'_p(x,y,z) = I(x,y,0)e^{-\alpha z} \exp \left\{ -\epsilon \int_0^z d\xi \int_0^\xi \left[\frac{\partial^\beta I(x,y,\rho)}{I(x,y,\rho)} \right. \right. \\ \left. \left. \times \partial_\beta \mu(x,y,\rho) + \partial^\beta \partial_\beta \mu(x,y,\rho) \right] d\rho + o(\epsilon^2) \right\}. \quad (9')$$

Although the two formulas (9), (9') do not agree, they give for thermal blooming the same practical results, due to the smallness of the absorption coefficient α (for another heuristic expression see Ref. 6).

A. Steady state case

Let us first consider propagation with forced convection. The thermal distortion of laser beams arises because the absorbed laser power in the medium changes the index of refraction and therefore changes the beam intensity itself. Indeed, one has as a function of temperature $T^{3,5}$

$$n(r) = 1 + \frac{dn}{dT} \Delta T(r) \quad \Delta T(r) = T(r) - T(0), \quad (10)$$

where dn/dT is the rate of change of the refractive index of the gas with respect to temperature at constant pressure.

For a continuous beam propagating along oz with a uniform wind of velocity v in the x direction, the hydrodynamic equation is

$$\rho c_p v \partial T(r) / \partial x = \alpha I(r), \quad \lim_{r \rightarrow \infty} T(r) = T_0, \quad (11)$$

where the quantities ρ, c_p are, respectively, the density and specific heat of the medium. Comparison of (11) and (10) gives

$$\epsilon \mu(r) = \frac{dn}{dT} \Delta T(r) + o(\epsilon^2),$$

and after integration of (11)

$$\epsilon \mu(r) = \alpha \frac{dn}{dT} \frac{1}{\rho c_p v} \int_{-\infty}^x I(x',y,z) dx' + o(\epsilon^2).$$

That is, taking (9) into account,

$$\epsilon \mu(r) = \alpha \frac{dn}{dT} \frac{1}{\rho c_p v} e^{-\alpha z} \int_{-\infty}^x I(x',y,0) dx' + o(\epsilon^2). \quad (12)$$

We now assume for the unperturbed intensity the collimated Gaussian profile

$$I_0(x,y,z) = I_0 \exp \left(-\frac{x^2 + y^2}{a^2} \right) e^{-\alpha z}, \quad (13)$$

where a is the e^{-1} beam radius. Then

$$\epsilon \mu(x,y,z) = \alpha \frac{dn}{dT} \frac{I_0}{\rho c_p v} e^{-(\alpha z + y^2/a^2)} \int_{-\infty}^x e^{-x'^2/a^2} dx + o(\epsilon^2) \\ = \left[\alpha \frac{dn}{dT} \frac{1}{\rho c_p v} a I_0 \right] \frac{\sqrt{\pi}}{2} e^{-(\alpha z + y^2/a^2)} \\ \times [1 + \operatorname{erf}(x/a)] + o(\epsilon^2),$$

which gives

$$\epsilon_c = \alpha \frac{dn}{dT} \frac{1}{\rho c_p v} a I_0, \quad (14)$$

$$\mu(x,y,z) = \frac{\sqrt{\pi}}{2} \exp(-\alpha z - y^2/a^2) \\ \times [1 + \operatorname{erf}(x/a)] + o(\epsilon),$$

so that one has $|\epsilon \mu(x,y,z)| \ll (\sqrt{\pi} \alpha |dn/dT| (1/\rho c_p v) a I_0)$. One can easily prove that ϵ is a very small parameter; for a seal-level propagation of 10.6 μm CO₂ laser radiation in atmosphere, typical conditions are⁵

$$\alpha = 1.6 \times 10^{-6} \text{ cm}^{-1}, \quad \frac{dn}{dT} = -10^{-6} \text{ }^\circ\text{C}^{-1}, \quad (14')$$

$$\rho = 1.2 \times 10^{-3} \text{ g cm}^{-3}, \quad C_p = 1.0 \text{ J g}^{-1} \text{ K}^{-1};$$

for a wind velocity $v = 40 \text{ cm s}^{-1}$ and a laser beam power $P = 80 \text{ kW}$ ($I_0 = P/a^2\pi$), one has $|\epsilon| = 1/4\pi \times 10^{-6}$, which justifies a first order theory.

In practical situations, a and α are respectively about 10 cm and 10^{-6} cm^{-1} so that one has

$$\alpha a \ll 1, \quad (15)$$

which will be used to simplify some expressions.

First, using (13) and (14) an easy calculation gives

$$h(x,y,z) = \partial^j \partial_j \mu(x,y,z) + \frac{\partial^\beta I(x,y,0)}{I(x,y,0)} \partial_\beta \mu(x,y,z) \\ = -\frac{2}{a^2} \left[\Phi_0(x,y) - \frac{a^2 \alpha^2}{2} \mu_1(x,y) \right] e^{-\alpha z} + o(\epsilon), \quad (16)$$

where one has

$$\mu_1(x,y) = \frac{1}{2} \sqrt{\pi} e^{-y^2/a^2} [1 + \operatorname{erf}(x/a)], \\ \Phi_0(x,y) = \frac{2x}{a} e^{-(x^2 + y^2)/a^2} + (1 - 4y^2/a^2) \mu_1(x,y). \quad (16')$$

But from (15) it follows that $\frac{1}{2} a^2 \alpha^2 |\mu_1(x,y)| \operatorname{ess} \ll |\Phi_0(x,y)|$, where $\operatorname{ess} \ll$ means that this inequality is valid except for x, y inside the two small ellipses $x^2 + \pi[(y \pm a)/2]^2 \leq \frac{1}{4} a^2 \frac{1}{16} \pi a^4$. Thus one may write

$$h(x,y,z) = -2a^{-2} \Phi_0(x,y) e^{-\alpha z} + o(\epsilon), \quad (17)$$

and one has

$$\int_0^z dz' \int_0^{z'} e^{-\alpha z''} dz'' = \frac{z}{a} \left(1 - \frac{1 - e^{-\alpha z}}{\alpha z} \right) = \frac{1}{2} z^2 g_1(\alpha z), \quad (18)$$

with

$$g_1(\alpha z) = \frac{2}{\alpha z} \left(1 - \frac{1 - e^{-\alpha z}}{\alpha z} \right) = 1 + o(\alpha z). \quad (18')$$

Substituting (13) and (17) into (9) and taking (18) into account one obtains

$$I_p(x,y,z) = I_0 e^{-\alpha z} \exp \left[-\frac{(x^2 + y^2)}{a^2} \right] \\ \times \exp \left[\epsilon_c z^2 a^{-2} \Phi_0(x,y) g_1(\alpha z) + o(\epsilon^2) \right] \\ = I_0 e^{-\alpha z} \exp \left[-\frac{(x^2 + y^2)}{a^2} \right] \\ \times \exp \left[-N_c \Phi_0(x,y) g_1(\alpha z) + o(\epsilon^2) \right], \quad (19)$$

where $N_c = -\epsilon z^2/a^2$ is a positive distortion parameter previously introduced in Ref. 5. One easily checks that in this case heuristic formula (9') leads to the same expression.

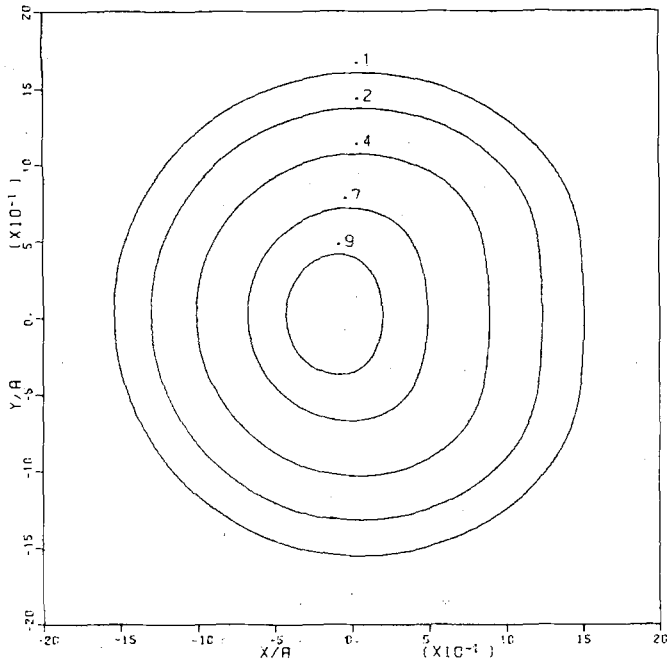


FIG. 1. Equal intensity contours collimated beam $N = 0.1$, $a = 10$ cm.

This result shows that the parameter to consider in the treatment of thermal blooming is N_c rather than ϵ_c , so that one has to determine the values of N_c , consistent with the relation (2) for $\mu(x,y,z)$ defined by (14); but when one substitutes (19) into (12) one obtains, instead of $\mu(x,y,z)$,

$$\begin{aligned} \mu_1(x,y,z;N_c) &= \frac{1}{a} e^{-(az + y^2/a^2)} \\ &\quad \times \int_{-\infty}^{x'} \exp\left[-\frac{x'^2}{2} - N_c \phi_0(x',y) g_1(\alpha z)\right] dx', \end{aligned}$$

and Eq. (2) becomes

$$\begin{aligned} n(x,y,z) &= 1 + \epsilon_c \mu_1(x,y,z;N_c) + o(\epsilon_c^2) \\ &= 1 + \epsilon_c \mu(x,y,z) \left(1 + \frac{\mu_1(x,y,z;N_c) - \mu(x,y,z)}{\mu(x,y,z)}\right) \\ &\quad + o(\epsilon_c^2). \end{aligned}$$

The consistency condition can be written (for a convenient norm)

$$\left\| \frac{\mu_1(x,y,z;N_c) - \mu(x,y,z)}{\mu(x,y,z)} \right\| < 1.$$

Assuming z small enough so that $g_1(\alpha z) \cong 1$, $(\mu_1 - \mu)/\mu$ does not depend on z ; besides, since the distortion is maximum for $y = 0$, we use as criterion

$$\sup_x \left| \frac{\mu_1(x,0,z;N_c) - \mu(x,0,z)}{\mu(x,0,z)} \right| < 1 \quad (19')$$

A numerical check of (19') shows that this condition is always fulfilled, but with a left-hand side very near unity as soon as $N_c > 1.5$, so it is advisable to keep $N_c < 1.5$, which corresponds, for a beam of transverse dimension $a = 10$ cm

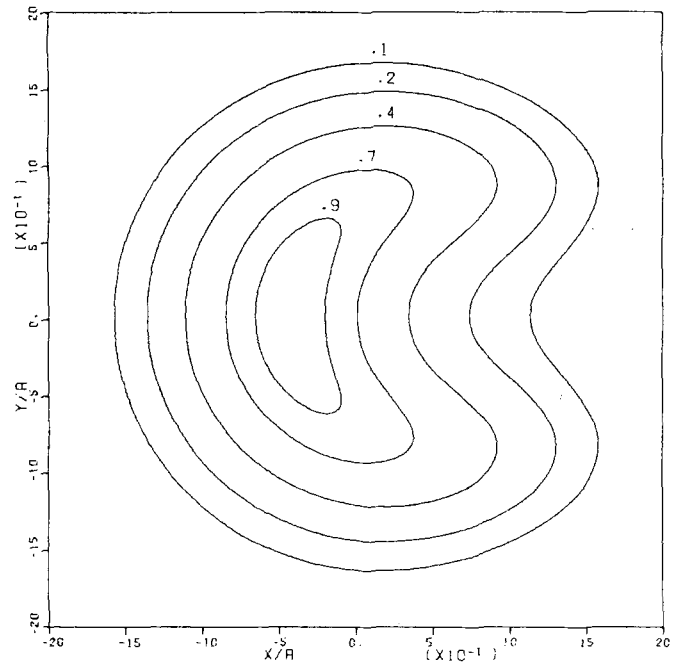


FIG. 2. Equal intensity contours collimated beam $N = 0.5$, $a = 10$ cm.

propagating in a medium with $\epsilon = (1/4\pi) \times 10^{-6}$, to a distance about 400 m.

We made a numerical application of Eq. (19), and Figs. 1-4 give the normalized isointensity curves $I_p(x,y,z)/\max_x I_p(x,0,z)$ for $N_c = 0.1$, $N_c = 0.5$, $N_c = 1$, and $N_c = 1.5$, while in Fig. 5 one has $I_p(x,0,z)/\max_x I_p(x,0,z)$ for the same

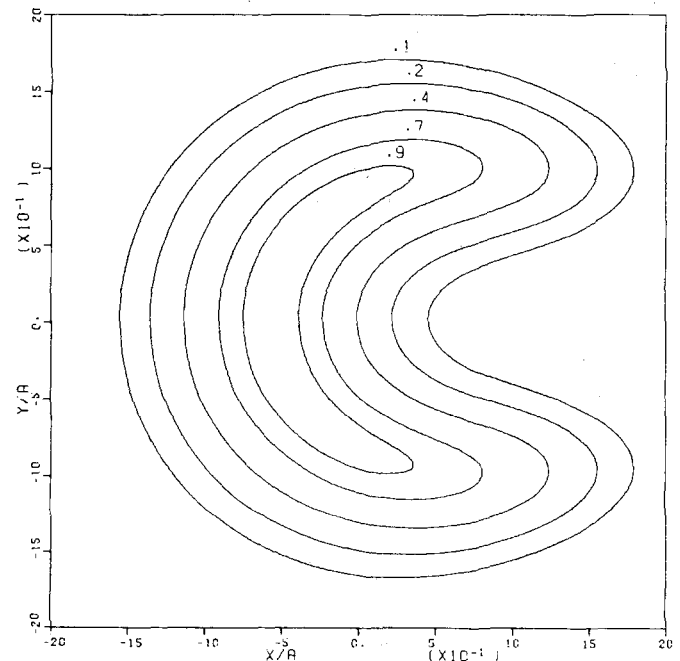


FIG. 3. Equal intensity contours collimated beam $N = 1$, $a = 10$ cm.

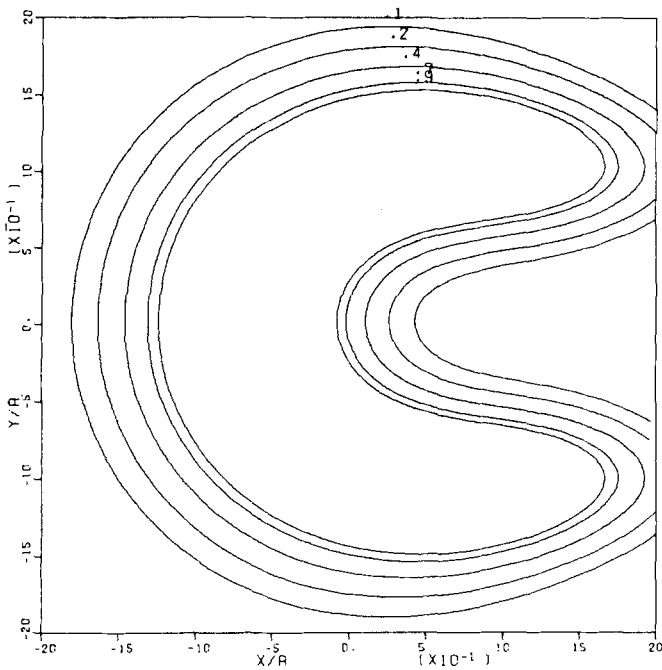


FIG. 4. Equal intensity contours collimated beam $N = 1.5$, $a = 10$ cm.

values of N_c . These curves were also obtained by other authors^{2,3,5}

For thermal-conduction-dominated propagation and for a Gaussian beam with circular symmetry (from now on r^2 means $x^2 + y^2$), the hydrodynamic energy equation is, in absence of convection,

$$\frac{k}{r} \frac{\partial}{\partial r} \left(r \frac{\partial T(r,z)}{\partial r} \right) = -\alpha I(r,z), \quad (20)$$

where k is the thermal conductivity, with $I(r,z) = I_0 e^{-(r^2/a^2)} e^{-\alpha z}$, and one has for the solution, without any singularity at $r = 0$,

$$\frac{\partial T(r,z)}{\partial r} = \frac{\alpha a^2 I_0}{2kr} (e^{-r^2/a^2} - 1) e^{-\alpha z}.$$

Now according to (10), one has $\partial n(r,z)/\partial r = (dn/dT)(\partial T(r,z)/\partial r)$ so that

$$\epsilon_D \frac{\partial \mu(r,z)}{\partial r} = \frac{dn}{dT} \frac{\alpha a^2 I_0}{2k} \frac{1}{r} (e^{-r^2/a^2} - 1) e^{-\alpha z} + o(\epsilon),$$

that is,

$$\epsilon_D = \frac{dn}{dT} \frac{a^2 \alpha}{2k} I_0, \quad (21)$$

$$\mu(r,z) = \int_0^r \frac{1}{r'} (e^{-r'^2/a^2} - 1) e^{-\alpha z} dr' + o(\epsilon).$$

As previously, let $h(r,z)$ be

$$\begin{aligned} h(r,z) &= \partial^i \partial_j \mu(r,z) + \frac{\partial^2 I(r,z)}{I(r,z)} \partial_\beta \mu(r,z) \\ &= \frac{\partial^2}{\partial r^2} \mu(r,z) + \frac{1}{r} \frac{\partial}{\partial r} \mu(r,z) + \frac{\partial^2}{\partial z^2} \mu(r,z) \\ &\quad - \frac{2r}{a^2} \frac{\partial}{\partial r} \mu(r,z). \end{aligned} \quad (22)$$

With (21) this results in

$$h(r,z) = -\frac{2}{a^2} e^{-\alpha z} (2e^{-r^2/a^2} - 1) + \alpha^2 \mu(r,z) + o(\epsilon).$$

According to (15), the last term on the right-hand side can be neglected

$$h(r,z) = -2a^{-2} e^{-\alpha z} (2e^{-r^2/a^2} - 1) + o(\epsilon) \quad (22')$$

Let $N_D = -\epsilon_D (z^2/a^2)$ be the distortion factor; then substituting (22') into (9) and using (18), one has

$$I_p(r,z) = I_0 e^{-\alpha z} e^{-r^2/a^2} \times \exp[-N_D g_1(\alpha z)(2e^{-r^2/a^2} - 1) + o(\epsilon^2)], \quad (23)$$

which is a particularly simple expression. Of course, to keep the theory consistent with (2), N_D (like N_c) must not be too large.

B. Transient case

In this section we consider a pulsed laser source and a Gaussian beam in two extreme situations.

1. Short time transient blooming

The laser pulse is assumed to be short compared to hydrodynamic time so that the density perturbation equation²

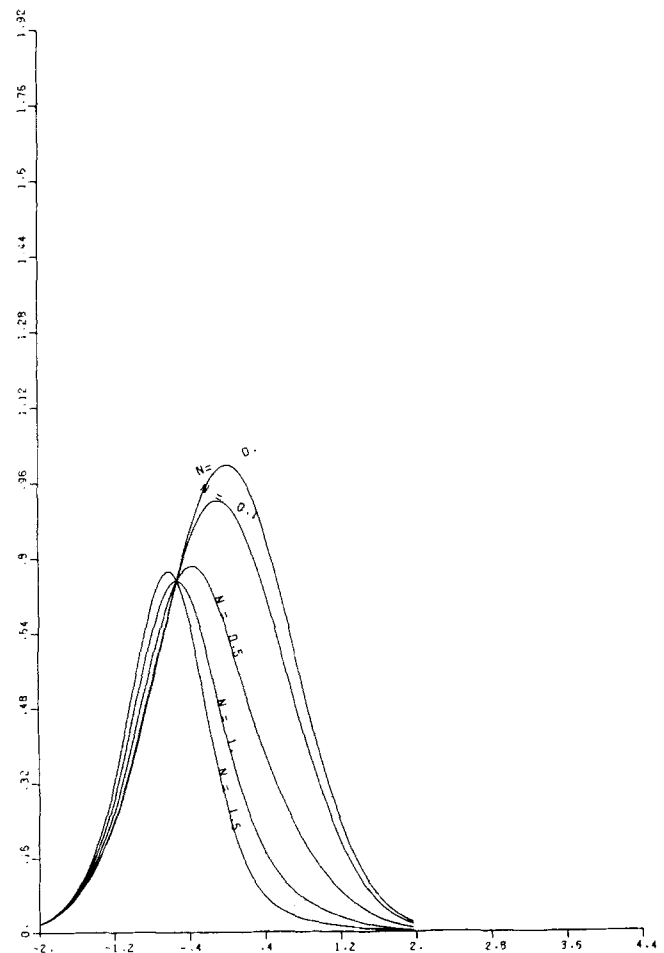


FIG. 5. Normalized intensity profiles for various values of N .

$$\left(\frac{\partial^2}{\partial t^2} - c_s^2 \partial_j^2\right) \frac{\partial \rho(r,z,t)}{\partial t} = (\gamma - 1) \alpha \partial_j^2 I(r,z,t), \quad (24)$$

where c_s is the acoustic velocity in the medium and γ the ratio of specific heat, reduces to

$$\partial^3 \rho(r,z,t) / \partial t^3 = (\gamma - 1) \alpha \partial_j^2 I(r,z,t),$$

which gives for a square pulse in time

$$\rho(r,z,t) - \rho_0 = \frac{1}{6} (\gamma - 1) \alpha \partial_j^2 I(r,z,t)^3. \quad (25)$$

To obtain $\epsilon \mu(r,z)$ one uses the Dale-Gladstone law

$$\epsilon \mu(r,z) = T \frac{dn}{dT} \frac{\Delta \rho(r,z)}{\rho}, \quad \Delta \rho(r,z) = \rho_0 - \rho(r,z), \quad (26)$$

together with the relation

$$(\gamma - 1) / C_p^2 = 1 / C_p T, \quad (26')$$

where C_p is the specific heat at constant pressure. With Eqs. (25), (26), and (26') one obtains for a Gaussian beam

$$\begin{aligned} \epsilon_{sc}(t) &= \frac{1}{6} \frac{dn}{dT} c_s^2 \alpha I_0 t^3 / \rho c_p a^2, \\ \mu(r,z) &= -a^2 \partial_j^2 \rho_j [e^{-\alpha z} e^{-r^2/a^2}] \\ &= 4(1 - r^2/a^2 - \frac{1}{2} \alpha^2 a^2) e^{-\alpha z} e^{-r^2/a^2} \end{aligned} \quad (27)$$

and still neglecting the $a^2 \alpha^2$ term,

$\mu(r,z) = 4(1 - r^2/a^2) e^{-\alpha z} e^{-r^2/a^2}$, which gives for the expression (22) of $h(r,z)$ (leaving aside the $a^2 \alpha^2$ term)

$$h(r,z) = -8a^{-2} (4 - 12r^2 a^{-2} + 4r^4 a^{-4}) e^{-r^2/a^2} e^{-\alpha z} + o(\epsilon). \quad (28)$$

substituting (28) into Eq. (9) and using (18), one obtains, with

$$\begin{aligned} N_{sc}(t) &= \epsilon_{sc}(t) (z^2/a^2), \\ I_p(r,z,t) &= I_0 e^{-\alpha z} e^{-r^2/a^2} \exp[-4N_{sc}(t)g_1(\alpha z) \\ &\quad \times (4 - 12r^2 a^{-2} + 4r^4 a^{-4}) e^{-r^2/a^2} + o(\epsilon^2)]. \end{aligned} \quad (29)$$

2. Long time transient blooming

We now assume that the laser pulse is long compared to the a/c_s time so that Eq. (24) becomes

$$\rho(t) - \rho_0 = -\frac{\gamma - 1}{C_s^2} \alpha I(r,z)t, \quad t \gg a/c_s, \quad (30)$$

which gives, with (26) and (26'),

$$\epsilon_{lc}(t) = \frac{dn}{dT} \frac{\alpha I_0 t}{\rho c_p}, \quad \mu(r,z) = e^{-\alpha z - (r^2/a^2)}, \quad (30')$$

and, still with the $a^2 \alpha^2$ term neglected, one has for $h(r,z)$

$$h(r,z) = -4a^{-2} (1 - 2r^2 a^{-2}) e^{-r^2/a^2} e^{-\alpha z} + o(\epsilon),$$

so that Eq. (9) becomes, with (18) and $N_{lc}(t)$

$$\begin{aligned} I_p(r,z) &= I_0 e^{-\alpha z} e^{-r^2/a^2} \exp[-2N_{lc}(t) \\ &\quad \times (1 - (2r^2/a^2)) e^{-r^2/a^2} g_1(\alpha z) + o(\epsilon^2)]. \end{aligned} \quad (31)$$

It is easy to show that in this case the condition (19') becomes $|e^{4e^{-3/2} N_{lc}} - 1| < 1$, which implies $N_{lc} < 0.8$. The analytical expressions of thermal blooming for collimated beams are very simple and easy to calculate. Eq. (31) can also be found in Refs. 3 and 5.

3. THERMAL BLOOMING FOR A FOCUSED BEAM IN THE NEAR FIELD

In this section, we assume $|z/f| < 1$, so that diffraction may be neglected. As proved in the Appendix of the forthcoming paper, but with absorption taken into account, Eq. (3) reduces to:

$$I_f(x(z), y(z), z) = \frac{I[x'(z), y'(z), 0]}{(1 - (z/f))^2 + \epsilon \beta(z)} e^{-\alpha z} \times \exp\{-\epsilon \psi(x(z), y(z), z)\} + o(\epsilon^2), \quad (32)$$

with $X'_\alpha(z) = (1 - (z/f))^{-1} X_\alpha(z)$, $\alpha = 1, 2$, and $\beta(z) = 2z/f(1 - (z/f))[\mu(x'(z), y'(z), 0) + (v_3(z)/z)]$

$$+ f^{-1} v_3(z) X_\alpha(z) \frac{\partial^\alpha I(x'(z), y'(z), 0)}{I(x'(z), y'(z), 0)}, \quad (33)$$

$$\begin{aligned} \psi(x(z), y(z), z) &= \frac{1}{1 - z/f} \frac{\partial^\alpha I(x'(z), y'(z), 0)}{I(x'(z), y'(z), 0)} \\ &\quad \times \int_0^z d\xi \int_0^\xi \partial_\alpha \mu(x(\rho), y(\rho), \rho) d\rho \\ &\quad + \int_0^z d\xi \int_0^\xi \partial_j^2 \mu(x(\rho), y(\rho), \rho) d\rho. \end{aligned} \quad (33')$$

Now, for a Gaussian beam, one has

$$\begin{aligned} I(x'(z), y'(z), 0) \\ = I_0 \exp[-x^\alpha(z) x_\alpha(z) / a^2 (1 - (z/f))^2], \end{aligned} \quad (34)$$

so that the unperturbed intensity is

$$I_{f0}(x(z), y(z), z) = \frac{I_0 e^{-\alpha z}}{(1 - (z/f))^2} \times \exp[-x^\alpha(z) x_\alpha(z) / a^2 (1 - (z/f))^2]. \quad (34')$$

We shall now use these equations to discuss thermal blooming of focused beams in the near field and this requires the following relation

$$\frac{\partial}{\partial z} X'_\alpha(z) = \frac{\partial}{\partial z} [X_\alpha(z) / \{1 - (z/f)\}] = 0 + o(\epsilon), \quad \alpha = 1, 2 \quad (35)$$

arising from the fact that $X'_\alpha(z)$ is the boundary condition $X_\alpha(0)$ for the unperturbed beam.

A. Steady state case

Substituting (34') into Eq. (12) gives

$$\begin{aligned} \epsilon \mu(x(z), y(z), z) &= \alpha \frac{dn}{dT} \frac{1}{\rho c_p v} \frac{I_0 e^{-\alpha z}}{(1 - (z/f))^2} \\ &\quad \times e^{-(y^2(z)/a^2)} \int_{-\infty}^x e^{-u^2/a^2 (1 - (z/f))^2} du \\ &= \left(\alpha \frac{dn}{dT} \frac{1}{\rho c_p v} a I_0\right) \frac{\sqrt{\pi}}{2} \frac{e^{-\alpha z}}{(1 - (z/f))} \\ &\quad \times e^{-y^2(z)/a^2} [1 + \operatorname{erf}(x'(z)/a)], \end{aligned} \quad (36)$$

which leads to:

$$\begin{aligned} \epsilon_c &= \alpha \frac{dn}{dT} \frac{1}{\rho c_p v} I_0 a, \\ \mu(x(z), y(z), z) &= \frac{1}{2} \sqrt{\pi} [e^{-\alpha z} / (1 - (z/f))] e^{-y^2(z)/a^2} \\ &\quad \times [1 + \operatorname{erf}(x'(z)/a)] + o(\epsilon_c). \end{aligned} \quad (36')$$

We now have to compute $\beta(z)$ and $\psi(x(z), y(z), z)$ with, form (34),

$$\begin{aligned} \partial^2 I(x'(z), y'(z), 0) / I(x'(z), y'(z), 0) \\ = -2x_\alpha(z) / a^2 (1 - z/f)^2. \end{aligned}$$

Let us write $\mu(x(z), y(z), z)$ in the form

$$\begin{aligned} \mu(x(z), y(z), z) &= \frac{e^{-az}}{(1 - z/f)} \mu_1(x'(z), y'(z)), \\ \mu_1(x'(z), y'(z)) &= \frac{\sqrt{\pi}}{2} e^{-(y'^2(z)/a^2)} [1 + \operatorname{erf}(x'(z)/a)] \\ &\quad + o(\epsilon_c), \end{aligned} \quad (37)$$

with, according to (35),

$$\frac{\partial}{\partial z} \mu_1(x'(z), y'(z)) = 0 + o(\epsilon). \quad (37')$$

From now on, to simplify, one writes X_α, X'_α for $X_\alpha(z), X'_\alpha(z)$, $\alpha = 1, 2$, and one introduces the following functions:

$$\begin{aligned} g_{0n}(\alpha z, f) &= \frac{1}{z} \int_0^z \frac{e^{-az'}}{(1 - (z'/f))^n} dz', \\ g_{1n}(\alpha z, f) &= \frac{2}{z^2} \int_0^z dz' \int_0^{z'} \frac{e^{-az''}}{(1 - (z''/f))^n} dz'' \end{aligned} \quad (38)$$

such that

$$\begin{aligned} \lim_{f \rightarrow \infty} g_{0n}(\alpha z, f) &= \frac{1}{\alpha z} (1 - e^{-\alpha z}) = 1 + o(\alpha z), \\ \lim_{f \rightarrow \infty} g_{1n}(\alpha z, f) &= \frac{2}{\alpha z} \left[1 - \frac{1 - e^{-\alpha z}}{\alpha z} \right] \\ &= g_1(\alpha z) = 1 + o(\alpha z). \end{aligned} \quad (38')$$

Then, according to (33) and (15) and taking into account (37) and (37'), one has for $\beta(z)$

$$\begin{aligned} \beta(z) &= \frac{2z}{f} (1 - (z/f)) \mu_1(x', y') \\ &\quad \times \left\{ 1 + \frac{1}{z} \left[1 - \frac{x'^\alpha x'_\alpha}{a^2 (1 - (z/f))} \right] \right. \\ &\quad \times \left. \int_0^z \left[\frac{e^{-az'}}{1 - (z'/f)} - 1 \right] dz' \right\} \\ &= \mu_1(x', y') \beta_1(z). \end{aligned} \quad (39)$$

$$\begin{aligned} \beta_1(z) &= \frac{2z}{f} (1 - (z/f)) \left\{ \frac{x'^\alpha x'_\alpha}{a^2 (1 - (z/f))} \right. \\ &\quad \left. + \left[1 - \frac{x'^\alpha x'_\alpha}{a^2 (1 - (z/f))} \right] g_{01}(\alpha z, f) \right\}. \end{aligned} \quad (39')$$

Let us now consider the first term on the right-hand side of (33'); using (34), (37), and (37'), one has

$$\begin{aligned} \frac{1}{1 - z/f} \frac{\partial^2 I(x', y', 0)}{I(x', y', 0)} \int_0^z d\xi \int_0^\xi \partial_\alpha \mu(x, y, \rho) d\rho \\ = - \frac{2}{a^2 (1 - (z/f))^2} \left[\frac{x'}{a} e^{-|(x'^2 + y'^2)/a^2|} - y'^2 a^{-2} \mu_1(x', y') \right] \\ \times \int_0^z dz' \int_0^{z'} \frac{e^{-az''}}{(1 - (z''/f))^2} dz'' \\ = - \frac{z^2 g_{12}(\alpha z, f)}{a^2 (1 - (z/f))^2} \left[\frac{x'}{a} e^{-|(x'^2 + y'^2)/a^2|} - y'^2 a^{-2} \mu_1(x', y') \right]. \end{aligned} \quad (40)$$

For the second term a simple computation gives

$$\begin{aligned} \partial^j \partial_j \mu(x, y, z) &= - \frac{2}{a^2} \left\{ \frac{e^{-az}}{(1 - (z/f))^3} \left[\frac{x'}{a} e^{-(x'^2 + y'^2)/a^2} \right. \right. \\ &\quad \left. \left. + (1 - 2y'^2 a^{-2}) \mu_1(x', y') \right] + \frac{\alpha^2 a^2 e^{-az}}{2(1 - (z/f))} \right. \\ &\quad \times \left[1 - \frac{2}{\alpha f (1 - (z/f))} + \frac{2}{\alpha^2 f^2 (1 - (z/f))^2} \right] \\ &\quad \left. \times \mu_1(x', y') \right\}. \end{aligned}$$

As in the collimated case, the $a^2 \alpha^2$ term of this last expression may be essentially neglected (that is except inside two very small ellipse with center at $(0, a/2)$, $(0, -a/2)$, so it becomes

$$\begin{aligned} \int_0^z d\xi \int_0^\xi \partial^j \partial_j \mu(x, y, \rho) d\rho \\ = - \frac{z^2}{a^2} g_{13}(\alpha z, f) \left[\frac{x'}{a^2} e^{-(x'^2 + y'^2)/a^2} \right. \\ \left. + (1 - 2y'^2 a^{-2}) \mu_1(x', y') \right]. \end{aligned} \quad (40')$$

Substituting (40) and (40') into (33') gives, for

$$\begin{aligned} F_c(x, y, z) &= \exp[-\epsilon_c \psi(x, y, z)], \text{ with } N_c = -\epsilon_c z^2 / a^2, \\ F_c(x, y, z) &= \exp \left\{ -N_c \left[g(\alpha z, f) \frac{x'}{a} e^{-(x'^2 + y'^2)/a^2} \right. \right. \\ &\quad \left. \left. + \left[g_{13}(\alpha z, f) - \frac{2y'^2}{a^2} g(\alpha z, f) \right] \mu_1(x', y') \right] \right\}, \end{aligned} \quad (41)$$

with

$$g(\alpha z, f) = \frac{g_{12}(\alpha z, f)}{(1 - (z/f))^2} + g_{13}(\alpha z, f). \quad (41')$$

Finally, the expression (32) of $I_f(x, y, z)$ becomes, with (39) and (41),

$$\begin{aligned} I_f(x, y, z) &= \frac{I_0 e^{-\alpha z} e^{-(x'^2 + y'^2)/a^2}}{(1 - (z/f))^2 + \epsilon \mu_1(x', y') \beta_1(z)} \\ &\quad \times \exp \left\{ -N_c g(\alpha z, f) \frac{x'}{a} e^{-(x'^2 + y'^2)/a^2} - N_c \right. \\ &\quad \times \left. [g_{13}(\alpha z, f) - 2y'^2 a^{-2} g(\alpha z, f)] \right. \\ &\quad \left. \times \mu_1(x', y') + o(\epsilon^2) \right\}. \end{aligned} \quad (42)$$

One can compare this result to (19) and it is easy to show that $\lim_{f \rightarrow \infty} I_f(x, y, z) = I_p(x, y, z)$. The most important perturbative factor in (42) is given by (41), and we made some numerical computations of the quantity $J(x, y, z) = e^{-|(x'^2 + y'^2)/a^2}$ $F_c(x, y, z)$ when αz is small enough to approximate $g_{12}(\alpha z, f)$ and $g_{13}(\alpha z, f)$ by the following expressions:

$$\begin{aligned} g_{12}(\alpha z, f) &= - \frac{2f^2}{z^2} [\log(1 - (z/f)) + (z/f)]; \\ g_{13}(\alpha z, f) &= (1 - (z/f))^{-1} \end{aligned}$$

But one first has to check the condition (19'), noticing that here the function $\mu(x(z), y(z), z)$ depends on both parameters, N_c and z/f . We made a numerical check of (19') in both cases:

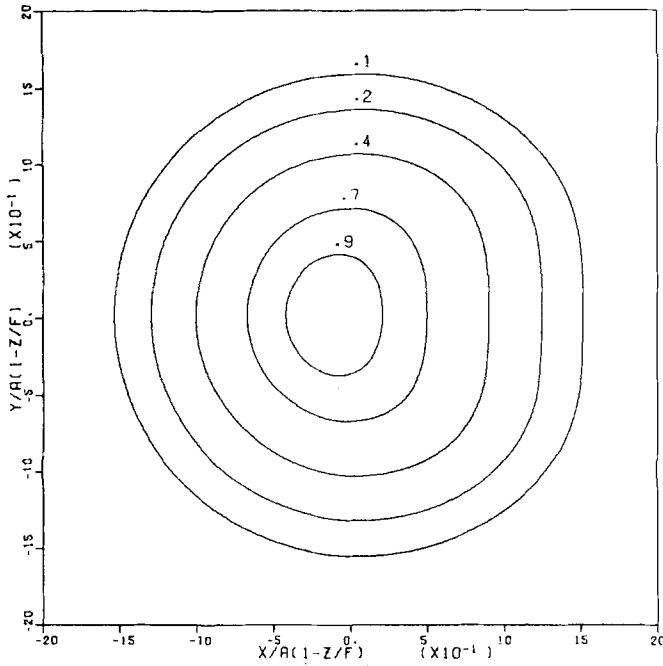


FIG. 6. Equal intensity contours focalized beam, $N = z^2/f^2 = 0.06$, $f = 1$ km, $a = 10$ cm.

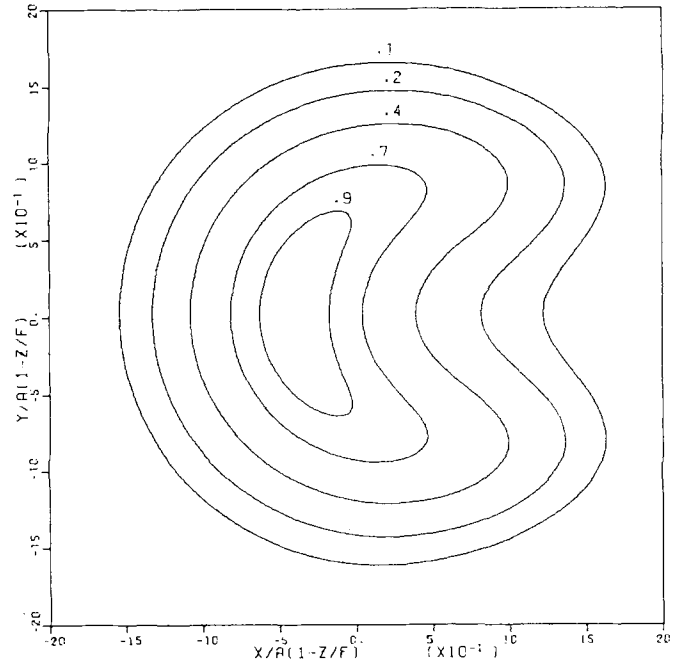


FIG. 8. Equal intensity contours focalized beam, $N = z^2/g^2 = 0.18$, $f = 1$ km, $a = 10$ cm.

$N_c = 10z^2/f^2$, which corresponds approximately to the conditions used for the collimated beam in the previous section, and $N_c = z^2/f^2$ (such a value can be obtained for a beam with a power ten times less than in the previous case). As limiting

values for N_c , we found, respectively, $N_c = 0.1$ ($z/f = 0.1$) and $N_c = 0.15$ ($z/f = 0.4$). These conditions are more severe than for a collimated beam, but for $f = 1$ km and $N_c = z^2/f^2$ the value $N_c = 0.15$ corresponds to 400 m, as in the collimated case. So, one gives in Figs. 6–9 the equal value contours for the normalized factor $J(x, y, z)/\max_x J(x, 0, z)$ for $N_c = z^2/f^2$ and $N_c = 0.06$, $N_c = 0.12$, $N_c = 0.18$, and $N_c = 0.21$. One must notice that in Figs. 6–9 the coordinates $x/a(1 - (z/f))$, $y/a(1 - (z/f))$ (with $a = 10$ cm, $f = 1$ km); then the comparison between these curves and those of Figs. 1–4 makes it possible to estimate the importance of thermal blooming for focused beams. Let us now consider the case of thermal conduction dominated propagation. Substituting (34') into Eq. (20) leads to

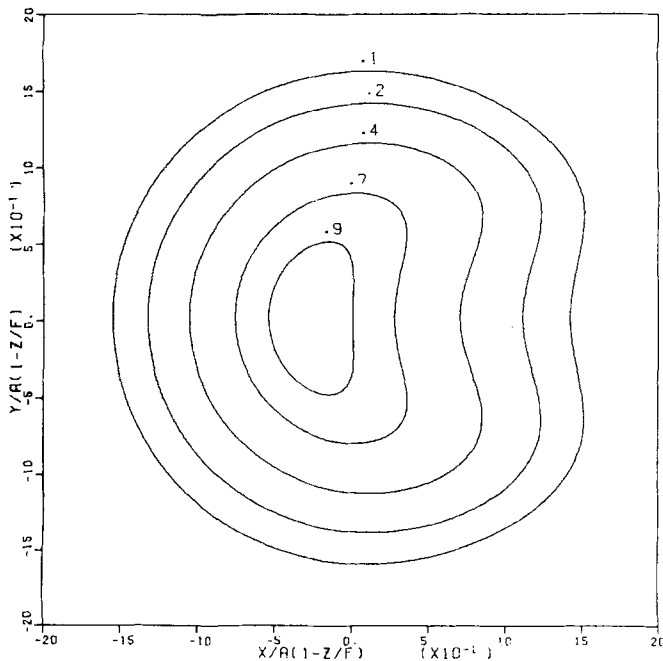


FIG. 7. Equal intensity contours focalized beam, $N = z^2/f^2 = 0.12$, $f = 1$ km, $a = 10$ cm.

$$\frac{k}{r} \frac{\partial}{\partial r} \left(r \frac{\partial I(r, z)}{\partial r} \right) = - \frac{\alpha I_0}{(1 - (z/f))^2} e^{-\alpha z} e^{-r^2/a^2(1 - (z/f))^2},$$

which gives after integration,

$$\frac{\partial I(r, z)}{\partial z} = \frac{\alpha a^2 I_0}{2kr} e^{-\alpha z} (e^{-r^2/a^2(1 - (z/f))^2} - 1),$$

and according to (10)

$$\epsilon_D \frac{\partial}{\partial r} \mu(r, z) = \frac{dn}{dT} \frac{\alpha a^2 I_0}{2kr} e^{-\alpha z} (e^{-r^2/a^2(1 - (z/f))^2} - 1) + o(\epsilon_D),$$

that is, with $r' = r/(1 - (z/f))$

$$\epsilon_D = \frac{dn}{dT} \frac{\alpha a^2 I_0}{2k};$$

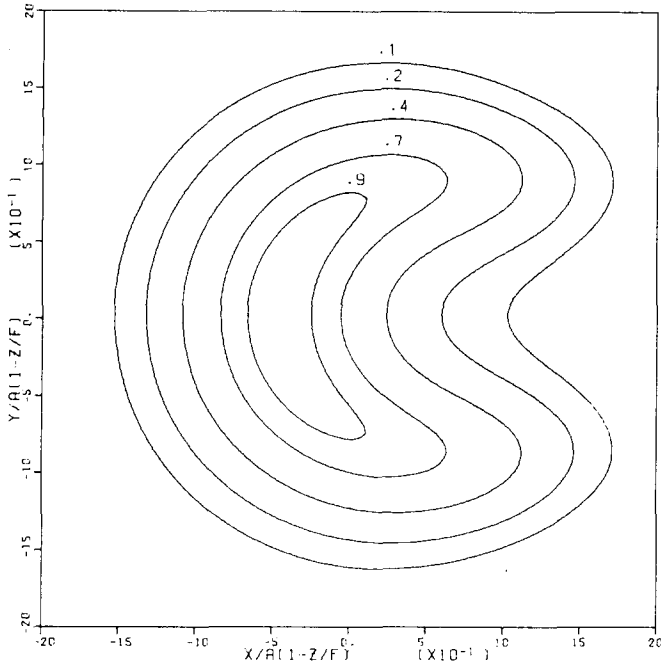


FIG. 9. Equal intensity contours focalized beam, $N = z^2/f^2 = 0.21$, $f = 1$ km, $a = 10$ cm.

$$\mu(r,z) = e^{-\alpha z} \int_0^r \frac{1}{r'} (e^{-r'^2/a^2} - 1) dr' + o(\epsilon_D). \quad (43)$$

Since the most important perturbative term is $F_D(x,y,z) = \exp[-\epsilon_D \psi(x,y,z)]$, we only compute this expression. One has

$$\frac{\partial \mu(r,z)}{\partial r} = \frac{e^{-\alpha z}}{(1-(z/f))^2} \frac{1}{r'} (e^{-r'^2/a^2} - 1),$$

so that one deduces from (33'), (34), and (43) (leaving aside the essentially negligible $\alpha^2 a^2$ term)

$$\begin{aligned} \psi(x,y,z) &= \frac{-2}{a^2(1-(z/f))^2} (-e^{r^2/a^2} - 1) \\ &\times \int_0^z dz' \int_0^{z'} \frac{e^{-\alpha z''}}{(1-(z''/f))^5} dz'' - \frac{2}{a^2} e^{-r^2/a^2} \\ &\times \int_0^z dz' \int_0^{z'} \frac{e^{-\alpha z''}}{(1-(z''/f))^2} dz'' \\ &= -\frac{z^2}{a^2} \left[\left[g_{12}(\alpha z, f) + \frac{g_{11}(\alpha z, f)}{(1-(z/f))^2} \right] \right. \\ &\left. \times e^{-r^2/a^2} - \frac{g_{11}(\alpha z, f)}{(1-(z/f))^2} \right]. \end{aligned}$$

Finally, using $N_D = -\epsilon_D(z^2/a^2)$, one has

$$\begin{aligned} F_D(x,y,z) &= \exp \left[-N_D \left[\left[g_{12}(\alpha z, f) + \frac{g_{11}(\alpha z, f)}{(1-(z/f))^2} \right] \right. \right. \\ &\left. \left. \times e^{-r^2/a^2} - \frac{g_{11}(\alpha z, f)}{(1-(z/f))^2} \right] \right], \quad (44) \end{aligned}$$

while in the collimated case, one has

$$\exp[-N_D g_1(\alpha z) [2e^{-r^2/a^2} - 1]].$$

Of course, Eq. (44) is only valid for the values of N_D such that the condition (19') is satisfied.

B. Transient case

For the short time transient blooming, using (25), and (26') together (34'), one obtains

$$\begin{aligned} \epsilon_{sc}(t) &= \frac{1}{6} \frac{dn}{dT} \frac{c_s^2 \alpha I_0 t^3}{\rho c_p a^2}, \\ \mu(r,z) &= -a^2 \partial^j \partial_j \left[\frac{e^{-\alpha z}}{(1-(z/f))^2} \right. \\ &\left. \times e^{-r^2/a^2(1-(z/f)^2)} \right] + o(\epsilon_{sc}). \quad (45) \end{aligned}$$

Taking (35) into account, Eq. (45) gives

$$\begin{aligned} \mu(r,z) &= -\frac{a^2 e^{-\alpha z}}{(1-(z/f))^2} \partial^\alpha \partial_\alpha [e^{-r^2/a^2(1-(z/f)^2)}] \\ &- a^2 e^{-r^2/a^2(1-(z/f)^2)} \\ &\times \frac{\partial^2}{\partial z^2} e^{-\alpha z} / (1-(z/f))^2 + o(\epsilon_{sc}) \\ &= \frac{4e^{-\alpha z}}{(1-(z/f))^4} \left[1 - \frac{r'^2}{a^2} \right] \\ &\times e^{-r^2/a^2} - \frac{a^2 \alpha^2 e^{-\alpha z} e^{-r^2/a^2}}{(1-(z/f))^2} \\ &\times \left[1 - \frac{4}{\alpha f(1-(z/f))} + \frac{6}{\alpha^2 f^2(1-(z/f))^2} \right] \\ &+ o(\epsilon_{sc}). \end{aligned}$$

As previously, the last term on the right-hand side is essentially negligible, so one has

$$\mu(r,z) = \frac{4e^{-\alpha z}}{(1-(z/f))^4} \left(1 - \frac{r'^2}{a^2} \right) e^{-r^2/a^2}. \quad (45')$$

Here also we only compute $F_{sc}(x,y,z,t) = \exp[-\epsilon_{sc}(t)\psi(x,y,z)]$, and leaving aside the $\alpha^2 a^2$ term, one deduces from Eqs. (33') and (45') (after some calculations)

$$\begin{aligned} \psi(x,y,z) &= -\frac{16}{a^2} \left[-\frac{r'^2}{a^2(1-(z/f))^2} \left[2 - \frac{r'^2}{a^2} \right] e^{-r^2/a^2} \right. \\ &\times \int_0^z dz' \int_0^{z'} \frac{e^{-\alpha z''}}{(1-(z''/f))^5} dz'' \\ &+ \left[2 - \frac{4r'^2}{a^2} + \frac{r'^4}{a^4} \right] e^{-r^2/a^2} \\ &\left. \times \int_0^z dz' \int_0^{z'} \frac{e^{-\alpha z''}}{(1-(z''/f))^6} dz'' \right] \\ &= -8 \frac{z^2}{a^2} e^{-r^2/a^2} \left[\left(2 - \frac{4r'^2}{a^2} + \frac{r'^4}{a^4} \right) g_{16}(\alpha z, f) \right. \\ &\left. - \frac{g_{15}(\alpha z, f)}{(1-(z/f))^2} \frac{r'^2}{a^2} \left(2 - \frac{r'^2}{a^2} \right) \right], \end{aligned}$$

which gives, with $N_{sc}(t) = -(z^2/a^2)\epsilon_{sc}(t)$,

$$\begin{aligned} F_{sc}(x,y,z,t) &= \exp \left[-8N_{sc}(t) e^{-r^2/a^2} g_{16}(\alpha z, f) \right. \\ &\times \left[2 - \frac{4r'^2}{a^2} + \frac{r'^4}{a^4} \right] + 8N_{sc}(t) e^{-r^2/a^2} \\ &\left. \times \frac{g_{15}(\alpha z, f)}{(1-(z/f))^2} \left[\frac{2r'^2}{a^2} - \frac{r'^4}{a^4} \right] \right]. \quad (46) \end{aligned}$$

For the long time transient blooming, one deduces from (30),

(26), and (26''), together with (34'),

$$\begin{aligned} \epsilon_{lc}(t) &= \frac{dn}{dT} \frac{\alpha I_0 t}{\rho c_p}, \\ \mu(r,z) &= \frac{e^{-\alpha z}}{(1 - (z/f))^2} e^{-r^2/a^2(1 - (z/f))^2}, \end{aligned} \quad (47)$$

which gives for $\psi(x,y,z)$ (still neglecting the term $(\partial^2/\partial z^2)\mu(r,z)$)

$$\begin{aligned} \psi(x,y,z) &= -\frac{4}{a^2} e^{-r^2/a^2} \left[(1 - r^2/a^2) \int_0^z dz' \right. \\ &\quad \times \int_0^{z'} \frac{e^{-\alpha z''} dz''}{(1 - (z''/f))^4} - \frac{r^2}{a^2(1 - (z/f))^2} \\ &\quad \times \left. \int_0^z dz' \int_0^{z'} \frac{e^{-\alpha z''} dz''}{(1 - (z''/f))^3} \right] \\ &= -\frac{2z^2}{a^2} e^{-r^2/a^2} \left[\left(1 - \frac{r^2}{a^2}\right) g_{14}(\alpha z, f) \right. \\ &\quad \left. - \frac{r^2}{a^2(1 - (z/f))^2} g_{13}(\alpha z, f) \right], \end{aligned}$$

so that the perturbative factor $F_{lc}(x,y,z,t)$ = $\exp\{-\epsilon_{lc}(t)\psi(x,y,z)\}$ becomes

$$\begin{aligned} F_{lc}(x,y,z,t) &= \exp \left[-2N_{lc}(t) e^{-r^2/a^2} \left(1 - \frac{r^2}{a^2}\right) g_{14}(\alpha z, f) \right. \\ &\quad \left. - \frac{r^2}{a^2(1 - (z/f))^2} g_{13}(\alpha z, f) \right], \\ N_{lc}(t) &= -\epsilon_{lc}(t)(z^2/a^2). \end{aligned} \quad (48)$$

This last result completes the discussion of thermal blooming for focused beams in the near field.

4. THERMAL BLOOMING IN THE FAR FIELD

To be brief, we only consider the steady state case with forced convection.

A. Collimated beam

We now assume that z is beyond the Rayleigh distance and, as proved in the forthcoming paper,¹ Eq. (13) becomes for a Gaussian beam (taking absorption into account)

$$\begin{aligned} \hat{I}_p(x,y,z) &= \frac{I_0}{d^2(z)} e^{-\alpha z} \exp \left(-\frac{x^\alpha x_\alpha}{a^2 d^2(z)} \right) \\ &\quad \times \exp \left[-\frac{\epsilon}{d(z)} \frac{\partial^\beta I(\hat{x}', \hat{y}', 0)}{I(\hat{x}', \hat{y}', 0)} \int_0^z d\xi \right. \\ &\quad \times \left. \int_0^\xi \partial_\beta \mu(x,y,\rho) d\rho - \epsilon \int_0^z d\xi \int_0^\xi \partial^j \partial_j \mu(x,y,\rho) d\rho \right] \\ &\quad + o \left[\left(\epsilon + (1/K_0^2 a^2) \right)^2 \right], \end{aligned} \quad (49)$$

with

$$d^2(z) = 1 + \frac{z^2}{K_0^2 a^4}, \quad \hat{x}'_\alpha = \frac{x_\alpha(z)}{d(z)}, \quad \alpha = 1, 2, \quad (49')$$

while similarly to (35), one has

$$\frac{\partial}{\partial z} \hat{x}'_\alpha(z) = \frac{\partial}{\partial z} \left(\frac{x_\alpha(z)}{d(z)} \right) = 0 + o \left(\frac{1}{K_0^2 a^2} \right), \quad \alpha = 1, 2. \quad (50)$$

According to (49) the unperturbed intensity is

$$\hat{I}_p(x,y,z) = \frac{I_0}{d^2(z)} e^{-\alpha z} \exp \left(-\frac{x^\alpha x_\alpha}{a^2 d^2(z)} \right) \quad (51)$$

and substituting (51) into Eq. (12) gives

$$\begin{aligned} \epsilon &= \alpha \frac{dn}{dT} \frac{I_0}{\rho c_p v} a, \\ \mu(x,y,z) &= \frac{\sqrt{\pi}}{2d(z)} \exp \left[-\alpha z - \frac{y^2}{a^2 d^2(z)} \right] \\ &\quad \times \left[1 + \operatorname{erf} \left(\frac{x}{ad(z)} \right) \right] + o(\epsilon). \end{aligned} \quad (52)$$

To simplify, we use the following notations

$$\begin{aligned} b_d(z) &= \frac{e^{-\alpha z}}{d(z)}, \\ \mu_{1d}(\hat{x}', \hat{y}') &= \frac{\sqrt{\pi}}{2} \exp(-\hat{y}'^2/a^2) [1 + \operatorname{erf}(\hat{x}'/a)], \end{aligned} \quad (53)$$

$$\mu_{2d}(\hat{x}', \hat{y}') = e^{-\hat{x}' \hat{x}'_\alpha / a^2},$$

and we introduce the functions

$$\hat{g}_{1d}(\alpha z) = \frac{2}{z^2} \int_0^z dz' \int_0^{z'} \frac{b(z'')}{d(z'')} dz'', \quad (53')$$

$$\hat{g}_{2d}(\alpha z) = \frac{2}{z^2} \int_0^z dz' \int_0^{z'} \frac{b(z'')}{d^2(z'')} dz''.$$

Let us first compute $\partial^j \partial_j \mu(x,y,z)$; using (50) one has

$$\begin{aligned} \partial^j \partial_j \mu(x,y,z) &= -\frac{2b_d(z)}{a^2 d^2(z)} \\ &\quad \times \left[\frac{\hat{x}'}{a} \mu_{2d}(\hat{x}', \hat{y}') + \left(1 - \frac{2\hat{y}'^2}{a^2}\right) \mu_{1d}(\hat{x}', \hat{y}') \right] \\ &\quad + \mu_{1d}(\hat{x}', \hat{y}') \frac{\partial^2 b_d(z)}{\partial z^2}. \end{aligned}$$

For the same reasons as in the previous sections, the last term on the right-hand side is essentially negligible, so that with (53') [(and still (50)]

$$\begin{aligned} &\int_0^z dz' \int_0^{z'} \partial^j \partial_j \mu(x,y,z'') dz'' \\ &= -\frac{z^2}{a^2} \left[\frac{\hat{x}'}{a} \mu_{2d}(\hat{x}', \hat{y}') + \left(1 - \frac{2\hat{y}'^2}{a^2}\right) \mu_{1d}(\hat{x}', \hat{y}') \right] \hat{g}_{2d}(\alpha z). \end{aligned} \quad (54)$$

In the same way, an easy computation gives

$$\begin{aligned} &\frac{1}{d(z)} \frac{\partial^\beta I(\hat{x}', \hat{y}', 0)}{I(\hat{x}', \hat{y}', 0)} \int_0^z d\xi \int_0^\xi \partial_\beta \mu(x,y,\rho) d\rho \\ &= -\frac{z^2}{a^2 d^2(z)} \left[\frac{\hat{x}'}{a} \mu_{2d}(\hat{x}', \hat{y}') - \frac{2\hat{y}'^2}{a^2} \mu_{1d}(\hat{x}', \hat{y}') \right] \hat{g}_{1d}(\alpha z). \end{aligned} \quad (54')$$

Substituting (54) and (54') into (49), we finally obtain for the diffracted intensity of a collimated beam

$$\begin{aligned} \hat{I}_p(x,y,z) &= \frac{I_0}{d^2(z)} e^{-\alpha z} e^{-\hat{x}' \hat{x}'_\alpha / a^2} \\ &\quad \times \exp \left[-N_c \left(\hat{g}_{2d}(\alpha z) + \frac{1}{d^2(z)} \hat{g}_{1d}(\alpha z) \right) \right. \\ &\quad \times \left(\frac{\hat{x}'}{a} \mu_{2d}(\hat{x}', \hat{y}') - \frac{2\hat{y}'^2}{a^2} \mu_{1d}(\hat{x}', \hat{y}') \right) \\ &\quad \left. - N_c \hat{g}_{2d}(\alpha z) \mu_{1d}(\hat{x}', \hat{y}') + o \left(\left(\epsilon + \frac{1}{K_0^2 a^2} \right)^2 \right) \right]. \end{aligned} \quad (55)$$

One easily sees that $\lim_{K_0 \rightarrow \infty} \hat{I}_p(x,y,z) = I_p(x,y,z)$.

B. Focused beam

In the Appendix of the forthcoming paper,¹ one proves that for a focused Gaussian beam, Eq. (3) can be written

$$\hat{I}_f[x(z), y(z), z] = \frac{I_0 e^{-\alpha z}}{D_0^2(z) + \epsilon \hat{\beta}(z)} \exp \left[-\frac{x^\alpha(z) x_\alpha(z)}{a^2 D_0^2(z)} \right] \times \exp \left[-\epsilon \hat{\psi}(x(z), y(z), z) + o \left(\left(\epsilon + \frac{1}{K_0^2 a^2} \right)^2 \right) \right],$$

with

$$D_0^2(z) = (1 - (z/f))^2 + (z^2/K_0^2 a^4) \quad (56')$$

and

$$\hat{\beta}(z) = \frac{2z}{f} \left(1 - \frac{z}{f} \right) \left[\mu \left[\frac{x(z)}{1 - (z/f)}, \frac{y(z)}{1 - (z/f)}, 0 \right] + \frac{\nu_3(z)}{z} \right] + \frac{\epsilon}{f} \nu_3(z) \frac{\partial^\alpha I(\hat{x}'(z), \hat{y}'(z), 0)}{I(\hat{x}'(z), \hat{y}'(z), 0)} X_\alpha(z), \quad (57)$$

$$\hat{\psi}(x(z), y(z), z) = \frac{1}{D_0(z)} \frac{\partial^\alpha I(\hat{x}'(z), \hat{y}'(z), 0)}{I(\hat{x}'(z), \hat{y}'(z), 0)} \times \int_0^z d\xi \int_0^\xi \partial_\alpha \mu(x(\rho), y(\rho), \rho) d\rho + \int_0^z d\xi \int_0^\xi \partial^j \partial_j \mu(x(\rho), y(\rho), \rho) d\rho, \quad (57')$$

where one has $\hat{x}'_\alpha(z) = x_\alpha(z)/D_0(z)$, satisfying

$$\frac{\partial}{\partial z} \hat{x}'_\alpha(z) = \frac{\partial}{\partial z} \left(\frac{x_\alpha(z)}{D_0(z)} \right) = 0 + o \left(\epsilon + \frac{1}{K_0^2 a^2} \right). \quad (58)$$

Since for the unperturbed beam the relation (56) gives

$$\hat{I}_f(x(z), y(z), z) = \frac{I_0 e^{-\alpha z}}{D_0^2(z)} \exp \left[-\frac{x^\alpha(z) x_\alpha(z)}{a^2 D_0^2(z)} \right], \quad (59)$$

the comparison of both sets of relations, on one hand (49), (49'), (50), and (51) and on the other hand (56), (56'), (57'), (58), and (59) shows that the computations of the collimated case are still valid provided that one uses $D_0^2(z)$ instead of $d^2(z)$. So, one has for the diffracted intensity of a focused beam

$$\hat{I}_f(x(z), y(z), z) = \frac{I_0 e^{-\alpha z}}{D_0^2(z) + \epsilon \hat{\beta}(z)} \exp \left[-\frac{\hat{x}'^\alpha(z) \hat{x}'_\alpha(z)}{a^2} \right] \times \hat{F}_c(x(z), y(z), z) + o \left(\left(\epsilon + \frac{1}{K_0^2 a^2} \right)^2 \right) \quad (60)$$

with

$$\hat{F}_c(x(z), y(z), z) = \exp \left(-N_c \left[\hat{g}_{2D_0}(\alpha z) + \frac{1}{D_0^2(z)} \hat{g}_{1D_0}(\alpha z) \right] \times \left[\frac{\hat{x}'(z)}{a} \mu_{2D_0}(\hat{x}'(z), \hat{y}'(z)) - \frac{2y'^2(z)}{a^2} \mu_{1D_0}(\hat{x}'(z), \hat{y}'(z)) \right] - N_c \hat{g}_{2D_0}(\alpha z) \mu_{1D_0}(\hat{x}'(z), \hat{y}'(z)) \right), \quad (60')$$

where μ_{1D_0} , μ_{2D_0} , \hat{g}_{1D_0} , and \hat{g}_{2D_0} are defined as in (53) and (53'). Of course this result is valid only for the values of N_c such that the condition (19') is fulfilled.

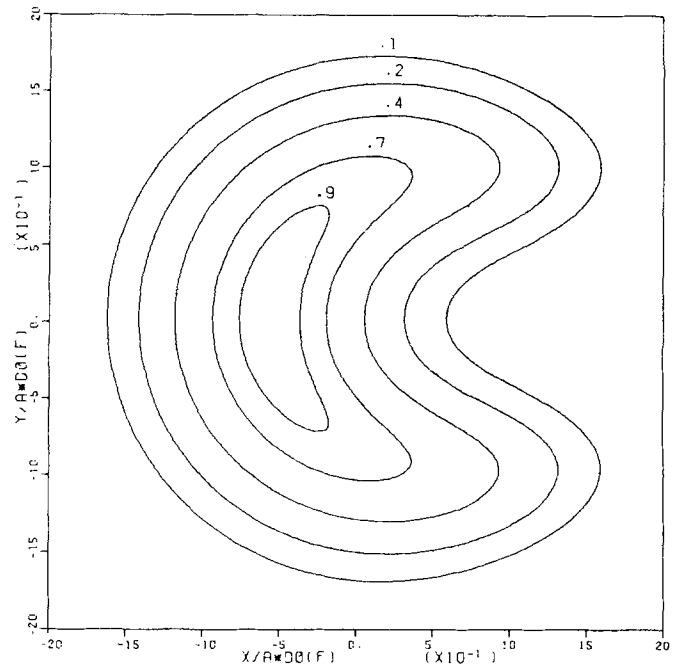


FIG. 10. Equal intensity contours in the local plane $N = 1$, $a = 10$ cm, $f = 10$ km.

One gives in Fig. 10 the equal value contours corresponding to the normalized quantity $\exp(-\hat{x}'^\alpha(z)\hat{x}'_\alpha(z)/a^2) \times \hat{F}_c(x(z), y(z), z)$, calculated in the geometrical focal plane $z = f$ and using the coordinates

$$\hat{x}'_\alpha(f) = x_\alpha(f)/aD_0(f) = x_\alpha(f)K_0 a/f, \quad \alpha = 1, 2,$$

but in order to have a clear picture of the diffractive effects in the focal plane one must restrict oneself to the case of weak thermal blooming and we considered a laser Gaussian beam with $a = 10$ cm, $f = 10$ km, and with a power small enough and propagating in such a medium that the distortion factor has the value $N = 1$ for $z = f$; then, since $K_0 = 6 \times 10^5 \text{ m}^{-1}$, one has $K_0 a = 6 \times 10^4$, and $f/a = 10^4$, hence $D_0(f) = 5/3$. We checked that for this value of N condition (19') is satisfied.

Remark: One easily sees that the other cases of thermal blooming can be deduced from the near-field results of the previous section only by changing $x'(z)$ and $(1 - (z/f))$ into $\hat{x}'(z)$ and $D_0(z)$.

5. CONCLUSIONS

These results show that in spite of its complexity the problem of the passage of a laser beam through an absorbing gas is amenable to an analytical solution, at least to first order approximation and when one may approximate hydrodynamic equations. But for continuous high energy beams, where thermal blooming is strong, (leaving aside the kinetic cooling phenomenon) the analytical solution breaks down when condition (19') is not fulfilled, so that one no

longer consider Eq. (2) as valid. And this generally happens for a distance less than the Rayleigh distance, where diffraction becomes important, except in some cases where ϵ is a decreasing function of z for—instance for a slewing beam with the transverse velocity $v = v_0 + \epsilon z$. It seems that this point was overlooked by those people comparing the geometrical solution to the computer results. On the contrary, for short pulsed high energy beams and for continuous moderate energy beams where Eq. (2) is valid, the analytical solution with approximation of diffraction gives excellent results and makes possible an easy and thorough discussion of the different parameters of a laser beam. Although we only considered thermal discussion for a Gaussian profile, the present method could also be applied to other smooth profiles, but the formulas are a bit more intricate.

APPENDIX

In this Appendix, one gives the equal intensity contours, at z fixed, of the normalized expression

$$J(x,y) = I_0(x,y)F(x,y,z)/\max_x[I_0(x,0)F(x,0,z)]$$

for a Gaussian beam of $1/e$ radius, $a = 10$ cm, either collimated or focused, and in the latter case, first in the near field and second in the geometrical focal plane. $I_0(x,y)$ is the intensity at $z = 0$, while $F(x,y,z)$ is the perturbative exponential factor due to thermal blooming. We only consider a continuous beam in the case of forced conduction, assuming αz small enough to approximate $e^{-\alpha z}$ by 1. The characteristics of the medium are those of a CO₂ laser beam propagating in normal air

$$\begin{aligned}\alpha &= 1.5 \times 10^{-6} \text{ cm}^{-1}, & \frac{dn}{dT} &= -10^{-6} \text{ }^\circ\text{C}^{-1}, \\ \rho &= 1.2 \times 10^{-3} \text{ g cm}^{-3}, & C_p &= 1.0 \text{ J g}^{-1} \text{ }^\circ\text{K}^{-1}, \\ v &= 40 \text{ cm s}^{-1},\end{aligned}$$

which gives $|\epsilon| = 10^{-7}$ for a laser beam power $P = 80$ kW. For a collimated beam, one has according to (19), since $e^{-\alpha z} \simeq 1$,

$$\begin{aligned}I_0(x,y)F(z,y,z) &= e^{-(x^2+y^2)/a^2} \exp\left[-N(z)\left[2xa^{-1}e^{-(x^2+y^2)/a^2}\right.\right. \\ &\quad \left.\left. + \frac{1}{2}\sqrt{\pi}e^{-y^2/a^2}(1-(4y^2/a^2))(1+\text{erf}(x/a))\right]\right], \quad (\text{A1})\end{aligned}$$

with $N(z) = \epsilon z^2/a^2 = 10^{-7} z^2/a^2$ ($a = 10$ cm). Figures 1 to 4 correspond respectively to the following values of N : 0.1, 0.5, 1.0, and 1.5, with the isovalue curves 1.0, 0.9, 0.7, 0.4, 0.2, and 0.1. The intensity profile $J(x,0)$ is given in Fig. 5 for the previous values of N .

For a focused beam ($f = 1$ km) in the near field, one has according to (42)

$$\begin{aligned}I_0(x,y)F(x,y,z) &= e^{-(x^2+y^2)/a^2} \exp\left(-N(z)\left\{g(\alpha z, f)x'a^{-1}e^{-(x^2+y^2)/a^2}\right.\right. \\ &\quad \left.\left.+ [g_{13}(\alpha z, f) - 2y'^2 a^{-2}g(\alpha z, f)]\right.\right. \\ &\quad \left.\left.\times \frac{1}{2}\sqrt{\pi}e^{-(y^2)/a^2}\left(1 + \text{erf}\frac{x'}{a}\right)\right\}\right), \quad (\text{A2})\end{aligned}$$

with

$$g(\alpha z, f) = \frac{g_{12}(\alpha z, f)}{(1-(z/f))^2} + g_{13}(\alpha z, f). \quad (\text{A2}')$$

Since $e^{-\alpha z} \simeq 1$ the expressions (38) reduce to:

$$\begin{aligned}g_{12}(\alpha z, f) &= -2f^2 z^{-2} [\ln(1-(z/f)) + (z/f)]; \\ g_{13}(\alpha z, f) &= 1/(1-(z/f)).\end{aligned} \quad (\text{A3})$$

For the collimated case one has $P = 8$ kW (so that $N = (z^2/f^2)$) and the equal value contours are given in Figs. 6–9 for $N = 0.06, 0.12, 0.18$, and 0.21 in the coordinate system $x'/a = x/a(1-(z/f))$, $y'/a = y/a(1-(z/f))$. At the geometrical focal plane $z = f$, the expressions (60) and (60') become

$$\begin{aligned}I(x,y)F(x,y) &= e^{-x'^2 a^2/a^2} \exp\left(-N(z)\left[\hat{g}_{D_0}(\alpha f)\frac{x'}{a}\right.\right. \\ &\quad \left.\left.\times e^{-(x'^2+y'^2)/a^2} + [\hat{g}_{2D_0}(\alpha f) - 2y'^2 a^{-2}\hat{g}_{D_0}(\alpha f)]\right.\right. \\ &\quad \left.\left.\times \frac{1}{2}\sqrt{\pi}e^{-(x'^2+y'^2)/a^2}\left(1 + \text{erf}\frac{x'}{a}\right)\right]\right),\end{aligned}$$

with

$$\hat{g}_{D_0}(\alpha f) = \frac{\hat{g}_{1D_0}(\alpha f)}{D_0^2(f)} + \hat{g}_{2D_0}(\alpha f), \quad D_0^2(f) = \frac{f^2}{K_0^2 a^4}.$$

Here we assume $P = 80$ kW, $f = 10$ km so that one has $N = 1$, $K_0 a = 6 \times 10^4$, $f/a = 10^5$, $D_0(f) = 5/3$, with $e^{-\alpha z} \simeq 1$, then the integrals (53') are

$$\begin{aligned}\hat{g}_{1D_0}(\alpha f) &= \frac{2}{f^2} \int_0^f dz' \int_0^{z'} \frac{dz''}{(1-(z''/f))^2 + D_0^2(f)(z''^2/f^2)} \\ &= \frac{\pi D_0(f) - 2 \ln D_0(f)}{1 + D_0^2(f)}\end{aligned}$$

$$\begin{aligned}\hat{g}_{2D_0}(\alpha f) &= \frac{2}{f^2} \int_0^f dz' \int_0^{z'} \frac{dz''}{[(1-(z''/f))^2 + D_0^2(f)(z''^2/f^2)]^{3/2}} \\ &= \frac{2}{D_0(f)}.\end{aligned}$$

The equal value curves at the focal plane for these numerical data are given in Fig. 10 in the coordinate system

$$\frac{x'}{a} = \frac{x}{aD_0(f)}, \quad \frac{y'}{a} = \frac{y}{aD_0(f)}.$$

¹P. Hillion and S. Quinnez, "Analytical solutions of geometric optics with an approximation of diffraction." To be submitted to J. Math. Phys.

²J. L. Walsh and P. B. Ulrich, in *Laser beam Propagation in the Atmosphere*, edited by J. W. Strohben (Springer, Berlin, 1978).

³D. C. Smith, Proc. IEEE 65, 1679 (1977).

⁴D. C. Smith, IEEE J. Quant. Elect. QE 5 600, (1969).

⁵F. G. Gebhardt and D. C. Smith, IEEE J. Quant. Elect. QE 7, 63 (1971).

⁶P. V. Avizonis, C. B. Hogge, R. R. Butts, and J. R. Kenemuth, Appl. Opt. 11, 554 (1972).

Analysis of a nonlinear integral equation arising in the study of the magnetic field in the critical state model of superconductivity

W. A. Beyer, G. M. Wing,^{a)} and A. Migliori
 Los Alamos Scientific Laboratory Los Alamos, New Mexico 87545

(Received 19 August 1980; accepted for publication 19 December 1980)

In the critical state theory of superconductors the magnetic field can be given as the solution of a nonlinear integral equation of Uryson type. It is shown that for sufficiently small values of physical and geometric parameters the equation always has a unique solution. A sufficient condition for the nonvanishing of the \mathbf{B} -field is derived and application is made to several models, two of which result from experiments with NbTi and Nb₃Sn. The investigation seems to cast doubt on the validity of some aspects of the critical state theory.

PACS numbers: 74.20. — z, 74.60.Jg

1. INTRODUCTION

The magnetic field associated with a type II superconductor carrying a transport current in zero applied field is given as the solution of a nonlinear integral equation of Uryson type if the critical state theory is used. When the superconductor is infinitely long and has a constant cross section, this equation is, in general, two-dimensional. We show in the case of a quite general cross section a unique solution always exists provided a certain parameter, dependent on both geometric and physical features, is sufficiently small. In addition, a criterion is found for the nonvanishing of the magnetic field within the core of the superconductor.

One theoretical model of a superconductor with a square cross section is discussed in detail, as are two actual experiments in which the cross section was rectangular. It is demonstrated that in all cases the existence and uniqueness theorem applies. In two of these cases it is shown, under very reasonable assumptions, that the magnetic field cannot vanish except in a relatively small part of the superconductor. (In the remaining model, no conclusion is reached.)

The existence and uniqueness of the B -field imply that the arbitrary imposition of the boundary condition of a shielded current free core on a superconductor carrying less than its critical current is, in general, invalid. The nonvanishing of the field found in certain of our investigations tends to cast some doubt on the validity of the critical state model.

In case the cross section of the superconductor is a circle or an annulus, the integral equation can be reduced to a nonlinear Volterra equation in one variable. It is shown that a unique solution then always exists and the field is zero inside the inner circle of the annulus, as is expected physically.

2. STATEMENT OF THE PROBLEM

We will present the problem in terms of classical vector analysis rather than using the more modern differential forms.¹ The level of rigor used in deriving the basic integral equation will be that of Stratton.² In particular, we assume

that all electromagnetic quantities involved have appropriate differentiability properties and that all integrals exists in some reasonable sense.

We consider a material which is homogeneous and isotropic. We also assume only the steady state case. The three equations of Maxwell

$$\begin{aligned} \text{curl} \mathbf{H} &= \mathbf{J}, \\ \text{div} \mathbf{B} &= 0, \\ \mathbf{B} &= \mu \mathbf{H}, \quad \mu \text{ const} \end{aligned}$$

lead to the law of Biot and Savart (Ref. 2, pp. 225–32):

$$\mathbf{B}(\mathbf{r}) = \frac{\mu}{4\pi} \int \frac{\mathbf{J}(\mathbf{r}') \times (\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} d^3\mathbf{r}'. \quad (2.1)$$

Here \mathbf{r} and \mathbf{r}' are vectors in three dimensional space, $d^3\mathbf{r}'$ denotes a volume element, and the integral is taken over all space. We shall use the Giorgi m.k.s. system of units throughout.

Now assume that the current \mathbf{J} is zero outside a cylindrical surface whose generators are parallel to the Z axis, that the cross section of the cylinder is a closed bounded region G , that the current \mathbf{J} is parallel everywhere to the Z axis, and that \mathbf{J} is independent of Z (see Fig. 1). Then

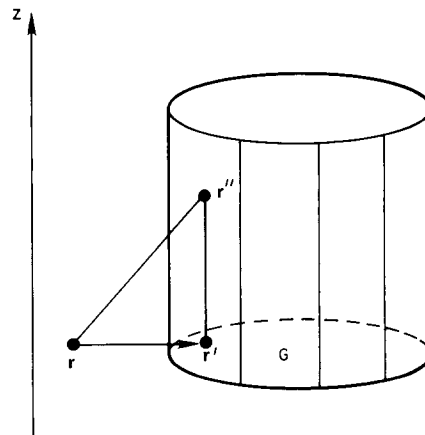


FIG. 1. The geometry for the cylindrical case of Eq. (2.1).

^{a)}Work done while on leave from Southern Methodist University, Dallas, Texas.

$$\begin{aligned} \mathbf{B}(\mathbf{r}) &= \frac{\mu}{4\pi} \int \frac{\mathbf{J}(\mathbf{r}'') \times [(\mathbf{r} - \mathbf{r}') + (\mathbf{r}' - \mathbf{r}'')] }{|\mathbf{r} - \mathbf{r}''|^3} d^3\mathbf{r}'' \\ &= \frac{\mu}{4\pi} \int \frac{\mathbf{J}(\mathbf{r}'') \times (\mathbf{r} - \mathbf{r}') }{|\mathbf{r} - \mathbf{r}''|^3} d^3\mathbf{r}'', \end{aligned}$$

since \mathbf{J} is parallel to $(\mathbf{r}' - \mathbf{r}'')$. Also $\mathbf{J}(\mathbf{r}'') = \mathbf{J}(\mathbf{r}')$ so that

$$\mathbf{B}(\mathbf{r}) = \frac{\mu}{4\pi} \int_G \mathbf{J}(\mathbf{r}') \times (\mathbf{r} - \mathbf{r}') d^2\mathbf{r}' \int_{-\infty}^{\infty} \frac{dz}{|\mathbf{r} - \mathbf{r}''|^3}.$$

Now

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{dz}{|\mathbf{r} - \mathbf{r}''|^3} &= \int_{-\infty}^{\infty} \frac{dz}{|(\mathbf{r} - \mathbf{r}') + (\mathbf{r}' - \mathbf{r}'')|^3} \\ &= \int_{-\infty}^{\infty} \frac{dz}{\{|\mathbf{r} - \mathbf{r}'|^2 + z^2\}^{3/2}} \\ &= \frac{2}{|\mathbf{r} - \mathbf{r}'|^2}. \end{aligned}$$

Thus

$$\mathbf{B}(\mathbf{r}) = \frac{\mu}{4\pi} \int_G \frac{\mathbf{J}(\mathbf{r}') \times (\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^2} d^2\mathbf{r}'. \quad (2.2)$$

In Ref. 3, Kim *et al.* assume that in a type II superconductor the magnitude of \mathbf{J} at a point is related to the magnitude of \mathbf{B} at that point by

$$|\mathbf{J}(\mathbf{r})| = F(|\mathbf{B}(\mathbf{r})|) = F(|\mathbf{B}|), \quad (2.3)$$

where F is a nonnegative, monotone decreasing function with a derivative for $|\mathbf{B}| > 0$ which is bounded in magnitude. This is called the critical state model. The possibility of F vanishing identically for $|\mathbf{B}|$ sufficiently large is allowed (see Fig. 2). Kim *et al.* suggest a possible form of F :

$$F(|\mathbf{B}|) \sim \alpha / (B_0 + |\mathbf{B}|), \quad \alpha > 0, B_0 > 0,$$

but we shall make no use of this explicit form.

Henceforth we shall be working entirely in the two-dimensional vector plane. Since this is equivalent to the complex plane it will be convenient to replace all vector quantities by complex numbers. Thus we shall replace $\mathbf{B}(\mathbf{r})$ by $b(z)$, etc. At all times z will denote a complex number and should not be confused with the coordinate Z . Equation (2.2) can now be rewritten as

$$b(z) = \frac{i\mu}{2\pi} \int_G \frac{F(|b(z')|)(z - z')}{|z - z'|^2} d^2z'. \quad (2.4)$$

Here d^2z' denotes two-dimensional measure in the plane.

Also $b(z)$ inherits the differentiability properties of $\mathbf{B}(\mathbf{r})$.

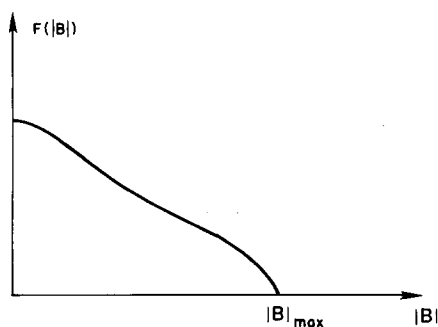


FIG. 2. The function $F(|\mathbf{B}(\mathbf{r})|) = |\mathbf{J}(\mathbf{r})|$.

3. THE CASE OF G AN ANNULUS OR CIRCLE

Suppose

$$G = \{z = re^{i\theta} | 0 < a < r < c\}.$$

From symmetry it is clear that $|b(z)|$ depends only upon the magnitude of z and we write $|b(z)| = \beta(r)$. Equation (2.4) then becomes

$$\beta(r) = \left| \frac{i\mu}{2\pi} \int_a^c \int_0^{2\pi} \frac{F(\beta(r'))(r - r'e^{i\theta'})}{|r - r'e^{i\theta'}|^2} r' dr' d\theta' \right|. \quad (3.1)$$

Now set

$$\begin{aligned} I &= \int_0^{2\pi} \frac{(r - r'e^{i\theta'})}{|r - r'e^{i\theta'}|^2} d\theta' = \int_0^{2\pi} \frac{d\theta'}{r - r'e^{-i\theta'}} \\ &= \frac{1}{ir} \int_0^{2\pi} \frac{ire^{i\theta'} d\theta'}{re^{i\theta'} - r'} \\ &= \frac{1}{ir} \int_C \frac{dz'}{z' - r'}, \end{aligned}$$

where C is the circle of radius r about zero in the z' plane. By Cauchy's theorem,

$$I = \begin{cases} 0, & r < r', \\ 2\pi/r, & r' < r. \end{cases} \quad (3.2)$$

Thus Eq. (3.1) yields

$$\beta(r) = \begin{cases} 0, & 0 < r < a, \\ (\mu/r) \int_a^r F(\beta(r')) r' dr', & a < r < c, \\ (\mu/r) \int_a^c F(\beta(r')) r', & c \leq r. \end{cases} \quad (3.3)$$

In view of the conditions imposed upon F we readily obtain

$$\begin{aligned} \frac{d}{dr} [r\beta(r)] &= \mu F(\beta(r)) r, \\ \beta(a) &= 0. \end{aligned} \quad (3.4)$$

Setting $y(r) = r\beta(r)$, we find

$$\frac{dy}{dr} = \mu F\left(\frac{y}{r}\right) r, \quad y(a) = 0. \quad (3.5)$$

For convenience extend F as an even function, $F(x) = F(-x)$, $x > 0$. Since F' is assumed bounded the function $\mu r F(y/r) = f(y, r)$ satisfies a uniform Lipschitz condition in any rectangle $0 < d < r < e < \infty$, $-\infty < p < y < q < +\infty$. The existence and uniqueness of the solution to (3.5) then follow from a classical result (see Ref. 4, Chap. 1, Theorem 2.3). This resolves the problem for the annulus.

For the case of G a circle, $a = 0$, we note that

$$\lim_{r \rightarrow 0} \mu r F(y/r) = \lim_{r \rightarrow 0} f(y, r) = 0.$$

By simply defining $f(y, 0) = 0$ the above reasoning may again be applied.

Theorem 1: Under the assumptions made on the function F , Eq. (2.4) possesses a unique solution $b(z)$ provided G is a circle or an annulus.

4. THE CASE OF A GENERAL G

The device employed in the previous section fails completely for a general bounded closed region G . We turn to the

Banach contraction mapping principle (see Ref. 5, pp. 141-42).

Let C be the Banach space of continuous complex-valued functions ϕ on the region G with the operations of addition and multiplication defined in the natural manner. Define the norm

$$\|\phi\| = \max_{z \in G} |\phi(z)|.$$

Consider the operator

$$T(\phi) = \lambda \int_G \frac{F(|\phi(z')|)|z - z'|}{|z - z'|^2} d^2z'.$$

Clearly T maps C into itself. Now

$$\begin{aligned} & \|T(\phi) - T(\psi)\| \\ &= \max_{z \in G} \left| \lambda \int_G \frac{F(|\phi(z')|) - F(|\psi(z')|)}{|z - z'|^2} (z - z') d^2z' \right| \\ &\leq \lambda \max_{z \in G} \int_G \frac{|F(|\phi(z')|) - F(|\psi(z')|)|}{|z - z'|} d^2z'. \end{aligned}$$

Since the derivative of F is bounded, F satisfies a Lipschitz condition:

$$\begin{aligned} |F(|\phi(z')|) - F(|\psi(z')|)| &\leq M \left| |\phi(z')| - |\psi(z')| \right| \\ &\leq M |\phi(z') - \psi(z')| \leq M \|\phi - \psi\|, \end{aligned}$$

where

$$M = \max_{\xi} |F'(|\xi|)|. \quad (4.1)$$

Thus,

$$\|T(\phi) - T(\psi)\| \leq \lambda \max_{z \in G} \int_G \frac{d^2z'}{|z - z'|} M \|\phi - \psi\|.$$

Let us define

$$\sigma_G = \max_{z \in G} \int_G \frac{d^2z'}{|z - z'|} \quad (4.2)$$

The mapping T is therefore a contraction provided

$$|\lambda| M \sigma_G < 1. \quad (4.3)$$

We can now assert the following:

Theorem 2: Under the assumptions made on F , the integral equation (2.4) has a unique continuous solution $b(z)$ provided $|\lambda| M \sigma_G < 1$, where $\lambda = \mu/2\pi$.

We pause at this point to note that all quantities in the contraction condition are physical parameters of the problem with the exception of σ_G . It is clear that σ_G contains information about the geometry, but its exact significance is not obvious. We seek a better understanding.

We introduce in the complex plane a polar coordinate system with its origin at the point z . Thus $|z - z'| = r$ and

$$\sigma_G = \max_z \iint_G \frac{r dr d\theta}{r} = \max_z \int_0^{2\pi} \int_0^{r(\theta)} dr d\theta$$

$$= \max_z \int_0^{2\pi} r(\theta) d\theta.$$

The inequality of Schwarz yields

$$\left(\int_0^{2\pi} r(\theta) d\theta \right)^2 \leq 2\pi \int_0^{2\pi} r^2(\theta) d\theta = 4\pi A_G,$$

where A_G is the area of G . Since the right-hand side of this inequality is clearly independent of z , we conclude

$$\sigma_G \leq (2\pi A_G)^{1/2}. \quad (4.4)$$

One can envision situations in which the area of G may be difficult to measure. If, however, the diameter d_G of G can be obtained, use may be made of the fact that of all plane sets of given diameter the circle has the largest area (see Ref. 6, p. 239). Thus

$$A_G \leq (\frac{1}{4}\pi) d_G^2,$$

and

$$\sigma_G \leq \pi d_G.$$

We can now reformulate Theorem 2:

Theorem 2': Under the assumptions made on F , the integral equation (2.4) has a unique continuous solution provided any of the following conditions holds:

$$(i) \quad \mu M \sigma_G / 2\pi < 1, \quad (4.5)$$

$$(ii) \quad (\mu M / \pi^2) A_G^{1/2} < 1, \quad (4.6)$$

$$(iii) \quad \frac{1}{2} \mu M d_G < 1. \quad (4.7)$$

5. CONDITION FOR NONZERO B

In this section we select an admissible function F and give a condition which must be satisfied at a point $z_0 \in G$ if the magnetic field is to vanish at z_0 .

Theorem 3: Suppose F has the form

$$F(|b|) = \begin{cases} \epsilon(b_0 - |b|), & |b| \leq b_0, \\ 0, & |b| > b_0, \end{cases} \quad (5.1)$$

where ϵ and b_0 are positive constants, and let the contraction condition (4.5) hold. Then for $z_0 \in G$, $b(z_0) = 0$ implies

$$\left| \int_G \frac{(z_0 - z')}{|z_0 - z'|^2} d^2z' \right| \leq 2\mu \epsilon A_G. \quad (5.2)$$

Proof: We note that F satisfies all the conditions imposed in Sec. 2. Indeed, $M = \max_{\xi} |F'(|\xi|)| = \epsilon$. Equation (2.4) becomes

$$b(z) = \frac{i\mu}{2\pi} \int_G \frac{\epsilon [b_0 - |b(z')|]}{|z - z'|^2} d^2z', \quad (5.3)$$

where $G^+ \subset G$ is that subregion on which $F > 0$. The contraction condition (4.5) becomes

$$\mu \sigma_G \epsilon / 2\pi < 1. \quad (5.4)$$

From (4.2), (5.3), and (5.4)

$$|b(z)| \leq \frac{\mu \epsilon b_0}{2\pi} \int_G \frac{d^2z'}{|z - z'|} \leq \frac{\mu \epsilon b_0}{2\pi} \sigma_G < b_0. \quad (5.5)$$

We can thus replace G^+ with G in (5.3) since

$$b_0 - |b(z')| > 0$$

for all z' in G .

Now suppose $b(z_0) = 0$. Then (5.3) yields

$$\frac{i\mu}{2\pi} \epsilon b_0 \int_G \frac{z_0 - z'}{|z - z'|^2} d^2 z' = \frac{i\mu \epsilon}{2\pi} \int_G |b(z')| \frac{z_0 - z'}{|z - z'|^2} dz'. \quad (5.6)$$

But, from (5.5),

$$\left| \int_G \frac{|b(z')|(z_0 - z')}{|z - z'|^2} d^2 z' \right| \leq \max_{z \in G} |b(z)| \sigma_G \leq \frac{\mu \epsilon b_0}{2\pi} \sigma_G^2. \quad (5.7)$$

Thus, (5.6) and (4.4) give

$$\left| \int_G \frac{z_0 - z'}{|z - z'|^2} d^2 z' \right| \leq \frac{\mu \epsilon}{2\pi} \sigma_G^2 \leq 2\mu \epsilon A_G,$$

which is (5.2).

It follows from (5.7) that if $b(z_0) = 0$,

$$\left| \operatorname{Re} \int_G \frac{z_0 - z'}{|z_0 - z'|^2} d^2 z' \right| \leq 2\mu \epsilon A_G. \quad (5.8)$$

An analogous inequality for the imaginary part of the integral also holds.

It is interesting to note that if G is an annulus, the left side of (5.2) vanishes for z_0 inside the inner circle. [This follows from (3.2).] Thus, no contradiction is encountered.

6. THE RECTANGULAR CROSS SECTION

In this section we discuss the case of a rectangular cross section. We again suppose F has the form (5.1). Define

$$\begin{aligned} R(x_0, y_0) &= \operatorname{Re} \int_G \frac{z_0 - z'}{|z_0 - z'|^2} d^2 z' \\ &= \int_G \frac{x_0 - x'}{(x_0 - x')^2 + (y_0 - y')^2} dx' dy' \end{aligned} \quad (6.1)$$

and

$$\begin{aligned} I(x_0, y_0) &= \operatorname{Im} \int_G \frac{z_0 - z'}{|z_0 - z'|^2} d^2 z' \\ &= \int_G \frac{y_0 - y'}{(x_0 - x')^2 + (y_0 - y')^2} dx' dy'. \end{aligned} \quad (6.2)$$

Suppose G is the rectangle \mathcal{R} with sides of length $2a$ and $2b$:

$$\mathcal{R} = \{(x, y) \mid -a < x < a, -b < y < b\}. \quad (6.3)$$

Thus

$$R(x_0, y_0) = \int_{-b}^b dy' \int_{-a}^a \frac{(x_0 - x')}{(x_0 - x')^2 + (y_0 - y')^2} dx', \quad (6.4a)$$

$$I(x_0, y_0) = \int_{-b}^b dy' \int_{-a}^a \frac{(y_0 - y')}{(x_0 - x')^2 + (y_0 - y')^2} dx'. \quad (6.4b)$$

Theorem 4: Let F be of the form (5.1) and assume that $0 \leq x_0 \leq a$, $0 \leq y_0 \leq b$. Then

$$R(x_0, y_0) < R(x_0, y), \quad y_0 > y, \quad (6.5)$$

$$R(x_0, y_0) < R(x, y_0), \quad x_0 < x, \quad (6.6)$$

$$I(x_0, y_0) > I(x_0, y), \quad y_0 > y, \quad (6.7)$$

$$I(x_0, y_0) > I(x, y_0), \quad x_0 < x. \quad (6.8)$$

Proof: In (6.4a), make the change of variables:

$$x_0 - x' = u, \quad y_0 - y' = v.$$

Then (6.4a) becomes

$$R(x_0, y_0) = \int_{y_0-b}^{y_0+b} dv \int_{x_0-a}^{x_0+a} \frac{udu}{u^2 + v^2}. \quad (6.9)$$

Note that $R(x_0, y_0) \geq 0$. Also

$$\frac{\partial R(x_0, y_0)}{\partial y_0} = \int_{x_0-a}^{x_0+a} u \left[\frac{1}{u^2 + (y+b)^2} - \frac{1}{u^2 + (y_0-b)^2} \right] du < 0,$$

since $0 \leq |y_0 - b| \leq y_0 + b$.

Also, recalling that $x_0 + a > 0$, $x_0 - a < 0$ we have

$$\frac{\partial R(x_0, y_0)}{\partial x_0} = \int_{y_0-b}^{y_0+b} \left[\frac{x_0+a}{(x_0+a)^2 + v^2} - \frac{x_0-a}{(x_0-a)^2 + v^2} \right] dv > 0,$$

which verifies (6.6).

If we augment our notation slightly and write

$$R(x_0, y_0) = R(x_0, y_0, a, b),$$

$$I(x_0, y_0) = I(x_0, y_0, a, b),$$

then it is easy to see from (6.4a) and (6.4b) that

$$R(y_0, x_0, b, a) = I(x_0, y_0, a, b),$$

and the remainder of the theorem follows.

We finally note that similar results may be obtained for (x_0, y_0) in other quadrants of the rectangle \mathcal{R} . However, it is clear from the overall symmetry that (5.8) and its analog need only be studied for (x_0, y_0) as in Theorem 4.

Theorem 4 is of importance since we wish to find values of x_0 (and y_0) for which (5.8) is violated. Inequalities (6.5) and (6.6) allow us to state that if such a violation occurs at (x_0, b) , then it also occurs for (x_1, y_1) provided $x_1 \geq x_0$, $|y_1| \leq b$. Inequalities (6.7) and (6.8) provide analogous results for the imaginary part of the integral in (5.2). Of course, even better information can be obtained by examining the entire left side of (5.2), but much more calculation is required.

The integration in (6.4a) yields

$$\begin{aligned} 2R(x_0, y_0) &= -(b - y_0) \log \frac{(a - x_0)^2 + (b - y_0)^2}{(a + x_0)^2 + (b - y_0)^2} \\ &\quad - (b + y_0) \log \frac{(a - x_0)^2 + (b + y_0)^2}{(a + x_0)^2 + (b + y_0)^2} \\ &\quad - 2(a - x_0) \left(\arctan \frac{b - y_0}{a - x_0} + \arctan \frac{b + y_0}{a - x_0} \right) \\ &\quad + 2(a + x_0) \left(\arctan \frac{b - y_0}{a + x_0} + \arctan \frac{b + y_0}{a + x_0} \right). \end{aligned} \quad (6.10)$$

[Note that $R(0, 0) = 0$.]

7. APPLICATION 1: A MODEL SUPERCONDUCTOR

In this section we discuss an application of the above theory to a square cross section by giving an example showing that the \mathbf{B} field is nonzero in over 85% of the cross section.

We make the following assumptions about the superconductor:

(a) The square-cross section has half side length:

$$a = 10^{-3} \text{ m},$$

(b) F has the form (5.1) with

$$\epsilon = 10^8 \text{ A/Wb}.$$

(This is a typical value for type II superconductors.) The constant $\mu = 4\pi \times 10^{-7} \text{ H/m}$ will be used throughout our applications.

The value of the left side of (4.6) is readily found to be 0.14. Therefore Theorem 2' applies.

The right-hand side of (5.2) is

$$2\mu\epsilon A_G = 1.0053 \times 10^{-3} \text{ m.} \quad (7.1)$$

From (6.10) with $y_0 = a$, $x_0 = 0.00046$ we compute

$$R(0.00046, 0.0001) = 1.0236 \times 10^{-3} \text{ m.} \quad (7.2)$$

Since (7.2) exceeds (7.1), it follows from Theorems 3 and 4 that when (x_0, y_0) is in the given square, but outside the square

$$|x| < 0.00046, \quad |y| < 0.00046,$$

the field must be nonzero. This square occupies about 21% of this total cross section. Hence the **B** field cannot vanish in about 79% of this cross section.

Detailed numerical evaluation of the left side of (5.2) indicates that actually the inequality (5.2) is violated in more than 85% of the given square, so that the field is nonzero in that fraction of the region.

8. APPLICATION 2: EXISTENCE AND UNIQUENESS RESULTS FOR TWO EXPERIMENTS

We consider two superconducting systems which have been investigated experimentally: an NbTi system⁷ and an Nb₃Sn system.⁸ NbTi generally loses superconductivity at 14 Wb/m² and Nb₃Sn at 24 Wb/m².⁹ These values are the abscissas B_{\max} in Fig. 2 at which F becomes zero. The peak current densities $F(0)$ measured in large samples at zero applied field are on the order of 3×10^8 A/m² for NbTi⁷ and 4×10^{10} A/m² for Nb₃Sn.⁸

One estimate for $M = \max\{|F'(|\mathbf{B}|)|\}$ is simply $F(0)/B_{\max}$. This is entirely consistent with the linear form of F used in Theorem 3, and we continue to assume that form. Thus

$$M_{\text{NbTi}} = 2.1 \times 10^7 \text{ A/Wb} \quad (8.1)$$

and

$$M_{\text{Nb}_3\text{Sn}} = 1.7 \times 10^7 \text{ A/Wb.} \quad (8.2)$$

The cross sections of the superconductors are rectangles. Using, as in Ref. 7, $2a = 0.0196$ m, $2b = 0.00091$ m we get $A_{\text{NbTi}} = 1.78 \times 10^{-5}$ m², and, from Ref. 8, $2a = 0.0127$ m and $2b = 1.2 \times 10^{-5}$ m so that $A_{\text{Nb}_3\text{Sn}} = 1.5 \times 10^{-7}$ m². The left side of (4.6) is 0.06 for NbTi and 0.47 for Nb₃Sn and both values are less than 1. Theorem 2 therefore implies the existence of a unique solution to (2.4) for those experiments.

9. APPLICATION 3: B-FIELDS INSIDE THE CORE FOR THE TWO EXPERIMENTS

For the rectangular cross sections for NbTi and Nb₃Sn discussed in Sec. 8, and for the material values used there, we have calculated the consequences of the criterion (5.2) for nonzero B given in Theorem 3.

The values of the right-hand side of (5.2) are respectively

$$2\mu\epsilon_{\text{NbTi}} A_{\text{NbTi}} = 9.395 \times 10^{-4} \text{ m} \quad (9.1)$$

and

$$2\mu\epsilon_{\text{Nb}_3\text{Sn}} A_{\text{Nb}_3\text{Sn}} = 6.409 \times 10^{-4} \text{ m.} \quad (9.2)$$

In the case NbTi, where

$$a = 0.0098, \quad b = 0.00045,$$

one finds

$$R(0.005, 0.000455) = 0.001 \text{ m} \quad (9.3)$$

and

$$I(0.0098, 0.00026) = 0.00096 \text{ m.} \quad (9.4)$$

Since (9.3) and (9.4) exceed (9.1), the area of the cross section for A_{NbTi} where the **B**-field must be nonzero is greater than 74% of the whole area. Numerical evaluation of the left side of (5.2) indicates that actually the inequality (5.2) is violated in more than 80% of the area of the rectangle.

Inequality (5.2) yielded no information in the case of Nb₃Sn. However, it must be emphasized that this does not imply the vanishing of the magnetic field. The criterion simply is not satisfied.

10. SUMMARY

In summary, we have shown that for certain real superconducting systems investigated experimentally, the critical state model described by Eq. (2.4) has a unique solution for the field **B**, and therefore it is inappropriate to apply boundary conditions on the interior which prescribe that the field is zero where the current is zero.

In Ref. 10 the same conclusion is reached. However, the discussion there must be considered somewhat heuristic, since the limiting processes used are not justified. This is not merely a mathematical nicety, for such processes frequently break down, especially in nonlinear problems. Our development leaves open completely any case in which the contraction condition (Theorem 2) does not apply, and the possibility of no solution or many solutions to (2.4) must be contemplated.

A physical argument is also given in Ref. 10 concerning the vanishing of the magnetic field as a solution to the integral equation inside the core, and it is correctly stated that such vanishing occurs in the case of the annulus. Whether this is the only geometry for which this is true, as conjectured in Ref. 10, remains to be seen. However, our studies certainly lend credence to this conjecture, and suggest that portions of the critical state theory must be questioned.

APPENDIX: POSSIBLE APPLICATION OF A THEOREM OF GOLOMB

It is clear that one wishes the contraction constant in Theorem 2 to be as small as possible. The analysis in Sec. 4 suggests that better estimates involving the geometry of the system may accomplish this. We mention one possible approach.

In Ref. 11 Golomb establishes the following result:

Theorem (Golomb): Consider the nonlinear integral equation

$$\psi(s) = \int_G K(s, t) f(t, \psi(t)) dt, \quad (A1)$$

where G is a measurable bounded set in n -dimensional Euclidean space, K is continuous for s and t in G , and f satisfies the Lipschitz condition

$$|f(t, u_1) - f(t, u_2)| \leq k |u_1 - u_2|.$$

Suppose $k^{\frac{1}{2}} < |A|$ where A is the eigenvalue of the smallest modulus of the iterated kernel

$$K^{(2)}(s, t) = \int_G K(s, r)K(t, r)dr.$$

Then (A1) has exactly one solution ψ which is continuous in G .

In our problem the identification

$$f = F(|b(z')|)(z - z')/|z - z'|, \quad K = 1/|z - z'|,$$

seems appropriate but by not means unique. However, all efforts thus far made to get an estimate on A to obtain stronger results than those appearing in Theorem 2' have been unsuccessful. The problem remains open.

¹H. Flanders, *Differential Forms with Applications to the Physical Sciences* (Academic, New York, 1963).

²J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill, New York, 1941).

³Y. B. Kim, C. F. Hempstead, and A. R. Strand, "Critical persistent currents in hard superconductors," *Phys. Rev. Lett.* **9**, 306-9 (1962).

⁴E. A. Coddington and H. Levinson, *Theory of Ordinary Differential Equations* (McGraw-Hill, New York, 1955).

⁵M. A. Krasnosel'skii, *Topological Methods in the Theory of Nonlinear Integral Equations* (Pergamon, New York, 1964).

⁶I. M. Yaglom and V. G. Boltyanskii, *Convex Figures* (Holt, Rinehart and Winston, New York, 1961).

⁷A. Migliori, R. D. Taylor, and R. J. Bartlett, "Magnetic field profiles of NbTi tapes," *IEEE Trans. Magn.* **MAG-13**, 198-200 (1977).

⁸A. Migliori, R. J. Bartlett, and R. D. Taylor, "Transport-current-induced magnetic field profiles of Nb₃Sn superconducting tape," *J. Appl. Phys.* **47**, 3266-3371 (1976).

⁹S. Foner, "High critical field superconductors," in *Superconductivity in d- and f-Band Metals*, edited by D. H. Douglass (Plenum, New York, 1976).

¹⁰A. Migliori, "The general critical state model in two dimensions and zero applied field: a uniqueness theorem and some consequences," *LA-UR 79-2099*, *J. Appl. Phys.* **51**, 3439 (1980).

¹¹M. Golomb, "Zur Theorie der Nichtlinearen Integralgleichungen, Integralgleichungssysteme und Allgemeinen Funktionalgleichungen," *Math. Z.* **39**, 45-75 (1934).

On the spontaneous magnetic order in the Hubbard model

J. Jędrzejewski^{a)} and A. Verbeure

Instituut voor Theoretische Fysica, Universiteit Leuven, B-3030 Leuven, Belgium

(Received 3 October 1980; accepted for publication 14 November 1980)

A rigorous proof is given of the absence of spontaneous magnetization in the Hubbard model in one and two dimensions.

PACS numbers: 75.10.Jm

In more than two dimensions rigorous proofs are given of the existence of magnetic transition for the Heisenberg models¹ as well as for other models.²

In one and two dimensions there is the well-known Mermin and Wagner³ result about the absence of spontaneous magnetization for the Heisenberg model. In this paper we complete the scheme by proving the same result for the Hubbard model.⁴

Although the result is expected on general physical grounds, no rigorous proof seems to be available in the literature. Attempts for a rigorous proof by exploiting a believed similarity of the model with the Heisenberg model have not been successful. We give a direct proof, exactly along the same lines as in Ref. 3.

Consider the ν -dimensional cubic lattice Z^ν and the local Hamiltonians H_A defined for all finite subsets $A \subset Z^\nu$ by

$$H_A = \sum_{x,y \in A} t(x-y) (C_{x,\sigma}^+ C_{y,\sigma} + C_{y,\sigma}^+ C_{x,\sigma}) + \frac{U}{2} \sum_{x \in A} n_{x,\sigma} n_{x,-\sigma} + h \sum_{x \in A} (n_{x,\sigma} - n_{x,-\sigma}), \quad (1)$$

where, for all $x \in A$, $t(x) \in \mathbb{R}$, $t(x) = t(-x)$, and $\tau = \sum_{x \in Z^\nu} x^2 |t(x)| < \infty$, $U \in \mathbb{R}$, $\sigma = \pm 1$ stand for the spin indices, C and C^+ are Fermion creation and annihilation operators, $n_{x,\sigma} = C_{x,\sigma}^+ C_{x,\sigma}$, and $h \in \mathbb{R}$ is the magnetic field.

Let ω_β^h be any lattice translation invariant β -KMS or equilibrium state for (1), which satisfies the Bogoliubov inequality

$$|\omega_\beta^h[A, B^*]|^2 \leq \lim_{\Lambda \rightarrow Z^\nu} \frac{\beta}{2} \omega_\beta^h(A^* A + A A^*) \omega_\beta^h[B^*, [H_\Lambda, B]] \quad (2)$$

for any pair A, B of local observables.

Now we prove that if $\nu \leq 2$

$$\lim_{h \rightarrow 0} \lim_{\Lambda \rightarrow Z^\nu} \frac{1}{N(\Lambda)} \sum_{x \in \Lambda} \omega_\beta^h(n_{x,\sigma} - n_{x,-\sigma}) = 0,$$

where $N(\Lambda)$ is the number of points in Λ .

The main thing for the proof is to choose the appropriate observables A and B in (2). Our choice is the following.

For $A \subset Z^\nu$ any centered cube, take

$$A_{\Lambda,\sigma}(k) = \sum_{x \in \Lambda} e^{ikx} C_{x,\sigma}^+ C_{x,-\sigma},$$

$$B_\Lambda(k) = \sum_{\sigma} A_{\Lambda,\sigma}(k),$$

where k is any element of the dual lattice $\tilde{\Lambda}$ of Λ . Then compute

$$\omega_\beta^h([B_\Lambda(k), A_{\Lambda,\sigma}(K)^*]) = N(\Lambda) m_\sigma(h) \quad (3)$$

where

$$m_\sigma(h) = \lim_{\Lambda \rightarrow Z^\nu} \frac{1}{N(\Lambda)} \sum_{x \in \Lambda} \omega_\beta^h(n_{x,\sigma} - n_{x,-\sigma})$$

is the magnetization. Further

$$\sum_{k \in \tilde{\Lambda}} \omega_\beta^h(A_{\Lambda,\sigma}(k)^* A_{\Lambda,\sigma}(k) + A_{\Lambda,\sigma}(k) A_{\Lambda,\sigma}(k)^*) \leq N(\Lambda)^2 \rho, \quad (4)$$

where

$$\rho = \sum_{\sigma = \pm 1} \lim_{\Lambda \rightarrow Z^\nu} \frac{1}{N(\Lambda)} \sum_{x \in \Lambda} \omega_\beta^h(n_{x,\sigma})$$

and

$$\begin{aligned} & \lim_{\Lambda \rightarrow Z^\nu} \omega_\beta^h([B_\Lambda(k)^*, [H_\Lambda, B_\Lambda(k)]]]) \\ &= 2 \sum_{\substack{x \in Z^\nu \\ \sigma' = \pm 1}} t(z-x) \exp[ik(z-x) - 1] \omega_\beta^h(C_{x,\sigma'}^+ C_{z,\sigma'} \\ &+ C_{z,\sigma'}^+ C_{x,\sigma'}) + 4hN(\Lambda) m_\sigma(h). \end{aligned} \quad (5)$$

It is important that this result is independent of the potential term.

From (2) it follows that (5) is positive. Hence we get an upper bound for (5) by adding the term with k replaced by $(-k)$. Using Schwartz inequality and $1 - \cos u \leq u^2/2$, one gets

$$\begin{aligned} & \lim_{\Lambda \rightarrow Z^\nu} \omega_\beta^h[B_\Lambda(k)^*, [H_\Lambda, B_\Lambda(k)]] \\ & \leq 4\rho k^2 N(\Lambda) \tau + 8hN(\Lambda) m_\sigma(h). \end{aligned} \quad (6)$$

Substitution of (3), (4), and (6) in (2) yields

$$\frac{1}{N(\Lambda)} \sum_{k \in \tilde{\Lambda}} \frac{m_\sigma(h)^2}{4\rho \tau k^2 + 8hm_\sigma(h)} \leq \frac{\beta}{2} \rho.$$

In the limit $\Lambda \rightarrow Z^\nu$ this becomes

$$\left(\frac{1}{2^\nu}\right) \int_{-\pi}^{\pi} d^\nu k \frac{m_\sigma(h)^2}{4\rho \tau k^2 + 8hm_\sigma(h)} \leq \frac{\beta}{2} \rho. \quad (7)$$

If $\nu \leq 2$, $\lim_{h \rightarrow 0} m_\sigma(h) \neq 0$ is in contradiction with (7). This

^{a)}On leave of absence from University of Wrocław, Poland.

proves the absence of spontaneous magnetization, and hence the absence of magnetic long range order,¹ for nonzero temperatures.

We remark that the proof does not work for more general Fermi lattice models with potentials of the type

$$\sum_{\substack{x,y \in A \\ \sigma = \pm 1}} v(x-y)n_{x,\sigma}n_{y,-\sigma}.$$

For these models the expression (5) does depend on the potential v and the argument breaks down. But it holds still if we replace our Hamiltonian (1) by the grand-canonical one.

Further, if we replace $A_{A,\sigma}(k)$ by $A_{A,\sigma}(r+K)$, and keep $B_A(k)$ unchange, where the vector K has the property that for each $x \in A$ $e^{iKx} = \pm 1$, the argument rules out the possibility of antiferromagnetic ordering in the one and two-dimensional Hubbard model. This conclusion is interesting in

connection with microscopic theories of metal-insulator transitions based on the Hubbard model. These theories consider the effect of correlations between electrons closely related to the existence of antiferromagnetic order in the system.⁵

¹F. J. Dyson, E. H. Lieb, and B. Simon, *J. Stat. Phys.* **18**, 335 (1978).

²J. Frölich, R. Israel, E. H. Lieb, and B. Simon, *Commun. Math. Phys.* **62**, 1 (1978); Part II to appear *J. Stat. Phys.*

³N. D. Mermin and H. Wagner, *Phys. Rev. Lett.*, **17**, 1133 (1966); see also P. C. Hohenberg, *Phys. Rev.* **158**, 383 (1967).

⁴J. Hubbard, *Proc. Soc. London, Ser. A* **276**, 238 (1963).

⁵N. F. Mott, *Metal-Insulator Transitions* (Taylor & Francis, London, 1974).

Erratum: A new method of matrix transformations. II. General theory of matrix diagonalizations via reduced characteristic equations and its application to angular momentum coupling [J. Math. Phys. 20, 2159 (1979)]

Shoon K. Kim

Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122

(Received 17 December 1980; accepted for publication 16 January 1981)

PACS numbers: 02.10.Sp, 99.10. + g

(1) The ninth line from the bottom on the right half of p. 2160 should read

$$P_v = \chi_v^{-1} \sum_{k=0}^{r-1} A^{r-1-k} \lambda_v^{(k)} = \chi_v^{-1} \sum_{k=0}^{r-1} A^{(k)} \lambda_v^{r-1-k} \quad (2.10')$$

$$\lambda_v^{(k)} = \sum_{i=0}^k c_i \lambda_v^{k-i}.$$

(2) In the first line of Theorem 2 on p. 2161, $A \in M(n \times n, p^{(r)}(x))$ should read

$$A \in M(n \times n, p^{(r)}(x)).$$

(3) In Eq. (3.6) on p. 2161, $\hat{T}_j = \chi_v [P_v]_j$.

(4) In the LHS of Eq. (4.1) on p. 2162, $[T\hat{T}]_{ij}$ should read $[\hat{T}T]_{ij}$.

(5) In Eq. (5.7) on p. 2163, $\{U^{-1}\}_{ij} = N_i S_{ij}$ should read $\{U^{-1}\}_{ij} = N_i \bar{S}_{ij}$.

(6) In the LHS of Eq. (5.14) on p. 2164, $[THT]_{ij}$ should read $[\hat{T}HT]_{ij}$.