On the relaxation to quantum-statistical equilibrium of the Wigner–Weisskopf atom in a one-dimensional radiation field. V. Exact solution for infinite systems

Russell Davidson

Chimie Physique II, Université libre de Bruxelles, Bruxelles 1050, Belgium

John J. Kozak

Department of Chemistry and Radiation Laboratory, University of Notre Dame, Notre Dame, Indiana 46556
(Received 5 November 1971)

In pursuit of the study begun in the preceding paper of the spontaneous emission of a Wigner–Weisskopf atom in a one-dimensional radiation field, the exact solution of this problem in the thermodynamic limit of a system of infinite extent is examined. The analysis reveals the existence in general of three contributions to the probability of the atom's being excited at any moment: a constant contribution and two sorts of time-dependent ones, exponential and nonexponential. For a specific choice of coupling function, this probability is written down in closed form. The link between the constant contribution and the appearance of ghost states and ergodicity is investigated, and criteria given to determine when this contribution does not arise. A numerical study is made of the expressions obtained, and the nonexponential part of the solution compared with approximations for it which have been previously obtained by various means. It is concluded that the predictions of the weak-coupling Prigogine–Résibois master equation are inadequate except when the model is ergodic in a certain sense, and some suggestions are made concerning the relevance of these studies to nonequilibrium statistical mechanics.

I. INTRODUCTION

This paper is a sequel to the preceding one (hereafter called IV) and to three previous papers of the authors.[1] A study is undertaken of the properties of an exact solution of the problem of the spontaneous emission of a Wigner–Weisskopf two-level atom in interaction with a one-dimensional radiation field, in the limit where the size of the system becomes infinite. This exact solution was obtained in III, where it was investigated for both finite and infinite systems in the limit of weak coupling between the atom and the field. Then, in IV, finite systems were again examined, without any approximation of weak coupling, and it was found that the exact solution and the approximate solution of III were indeed very similar for small couplings, and for all but the smallest systems. This was in contrast to another weak-coupling approach elaborated in II, and based on the Prigogine–Résibois master equation for the diagonal elements of the density matrix. This approach yielded an approximate solution of a different analytic form from those of III and IV, and which accorded with neither in a numerical comparison. Both weak-coupling treatments, however, led to the same result in the thermodynamic limit of an infinite system, namely, a simple exponential decay. But even in this limit, it was seen in I that the solution of the master equation contained nonexponential terms when one considered anything more than the lowest order in the coupling. It is therefore of interest to see if the exact solution for an infinite system manifests this property, and if so, to compare the nonexponential part of it with that found in I.

In order to treat the problem of spontaneous emission of the two-level atom, it has been seen to be sufficient to consider a Hilbert space spanned by the state vectors

$$|\mathfrak{A}\rangle \quad \text{and} \quad \{|\lambda\rangle\},$$

where $|\mathfrak{A}\rangle$ denotes the state with the atom excited and the radiation field completely de-excited, and $|\lambda\rangle$ that with the atom de-excited and one photon excited in the λth mode of the field. The index λ labels the modes of the field, the properties of which are to be determined by the geometry of the system, i.e., its size and dimensionality. The Hamiltonian can be written [see Eq. (IV-5)]

$$H = \hbar E|\mathfrak{A}\rangle\langle\mathfrak{A}| + \sum_\lambda \hbar\omega_\lambda|\lambda\rangle\langle\lambda| + \sum_\lambda [h_\lambda\sqrt{2}\,|\lambda\rangle\langle\mathfrak{A}| + h_\lambda^*\sqrt{2}\,|\mathfrak{A}\rangle\langle\lambda|], \quad (1)$$

where $\hbar E$ is the energy separating the two levels of the atom, $\hbar\omega_\lambda$ is the energy of a photon in the λth mode, and h_λ is a (possibly complex) coupling parameter. Now, if the system is in the state $|\mathfrak{A}\rangle$ at time $t = 0$, it has been shown that the probability amplitude for finding the state $|\mathfrak{A}\rangle$ at time t is given by (see Eq. [III-12])

$$\langle\mathfrak{A}|\Psi(\tau)\rangle = -\frac{1}{2\pi i}\int_C dz e^{-izt}\left(z - E - \sum_\lambda \frac{2|h_\lambda|^2}{\hbar(z-\omega_\lambda)}\right)^{-1}. \quad (2)$$

(C is a contour above the real axis and parallel to it.) This result, in the form of an inverse Laplace transform integral, was obtained directly by solving the Schrödinger equation for the problem. This form is particularly convenient for present purposes, since to proceed to the infinite system limit, it is sufficient to replace the quantity

$$\sum(z) \equiv \sum_\lambda [2|h_\lambda|^2/\hbar^2(z-\omega_\lambda)] \quad (3)$$

by its limit when the spectrum of energies $\hbar\omega_\lambda$ becomes continuous. For a one-dimensional system of length L, with periodic boundary conditions, the ω_λ are given by

$$\omega_\lambda = 2\pi|n|c/L, \quad (4)$$

where the nonzero integers n replace λ as the label. Finally, c is the velocity of light. As was discussed in the previous papers, it is best, with this choice of the ω_λ, to take the couplings as [see Eq. (IV-2)]

$$|h_\lambda|^2 = \hbar^2\alpha c E^{1+p}/L\omega_\lambda^p \quad (0 < p < 1), \quad (5)$$

where α is a dimensionless coupling constant, which corresponds, for a one-dimensional system to the fine-structure constant of quantum electrodynamics. This choice of $|h_\lambda|^2$ avoids difficulties associated with the infrared and ultraviolet divergences. With the specifications of Eqs. (4) and (5), Eq. (3) becomes

$$\sum(z) = 2\sum_{n=1}^{\infty} \frac{2\alpha c E^{1+p}}{L(2\pi nc/L)^p} \cdot \frac{1}{z - (2\pi nc/L)} \tag{6}$$

and as $L \to \infty$, this goes over to an integral:

$$\sum_{\infty}(z) = \frac{2}{\pi}\alpha E^{1+p}\int_0^{\infty} dk \cdot \frac{1}{k^p(z-k)}$$
$$= -2\alpha E^{1+p} z^{-p} e^{ip\pi}\csc(p\pi) \tag{7}$$

for $0 < \arg z < \pi$, as in Eq. (2). For the calculation of Eq. (2), it is convenient to employ dimensionless variables, and we shall define these as follows:

$$\tau = \alpha E t, \qquad \xi = z/\alpha E,$$

making use of the scaling by α which was found necessary to make a weak-coupling scheme meaningful. Then, with the further definition

$$\sigma(\xi) = (1/\alpha E)\sum(z) = (1/\alpha E)\sum(\alpha E\xi), \tag{8}$$

Eq. (2) becomes

$$\langle \mathfrak{A}|\Psi(\tau)\rangle = -\frac{1}{2\pi i}\int_c d\xi e^{-i\xi\tau}\left(\xi - \frac{1}{\alpha} - \sigma_\infty(\xi)\right)^{-1} \tag{9}$$

with

$$\sigma_\infty(\xi) = -z e^{ip\pi}\csc(p\pi)(\alpha\xi)^{-p}.$$

This result, Eq. (9), will be the principal object of study in the remainder of the paper. It was shown in III that in the limit $\alpha \to 0$ it leads to the result that the probability of finding the state $|\mathfrak{A}\rangle$ at time τ is

$$\rho(\tau) = |\langle \mathfrak{A}|\Psi(\tau)\rangle|^2 = e^{-4\tau}, \tag{10}$$

the simple exponential decay mentioned earlier.

As soon as one tries to go beyond this approximation, it is found that corrections to Eq. (10) are nonanalytic in α. This was the main conclusion of I, and the same phenomenon was pointed out in III for an infinite system. Nonanalyticity in α is also a feature even of the weak-coupling solution for a finite system, and this matter has been extensively discussed in II, III, and IV. Mazur and Siskens[2] have argued convincingly that nonanalyticity with respect to the coupling constant of the constants of motion is linked closely with the appearance of ergodic properties in the thermodynamic limit. Such nonanalyticity will also appear in the matrix elements of the transformation which diagonalizes the Hamiltonian [for our system these are displayed in Eq. (IV-12)], and a connection has been made by Cukier and Mazur[3] between the appearance of this phenomenon and of ergodicity of certain phase functions simultaneously in the thermodynamic limit. In their work, they considered a system which, like ours, is not a completely ergodic system even in this limit, but for which certain dynamical variables possess an ergodic property that may be studied by means of some bounds on time-averaged autocorrelation functions derived by Mazur.[4] In view of the formal similarities between our model and that of Cukier and Mazur, it is interesting to see whether the same kind of link can be made connecting the nonanalyticity of our model with ergodicity of the probability $\rho(\tau)$ for the state $|\mathfrak{A}\rangle$, since, if this quantity decays, by whatever means, to zero as $\tau \to \infty$ in the thermodynamic limit, then it will be ergodic.

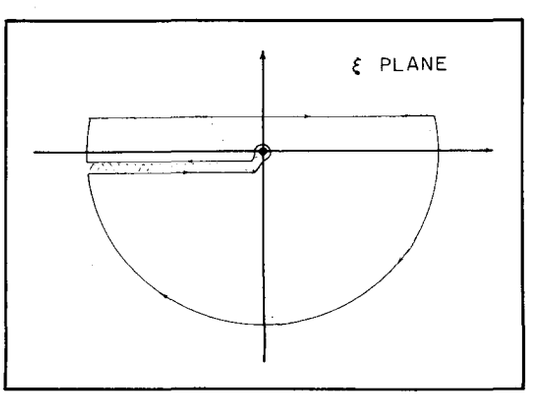


FIG. 1. The contour used in the study of Eq. (9) for $0 < p < 1$.

In the next section, the properties of Eq. (9) will be examined as a function of the time, τ, and then in Sec. III, for the particular choice $p = \frac{1}{2}$ of the exponent in Eq. (5), the contour integral will be explicitly calculated in closed form, and its properties studied. The matter of ergodicity will be considered in Sec. IV, and criteria will be established for its occurrence. A numerical evaluation of $\rho(\tau)$ is reported in Sec. V, along with a comparison of this exact result with the nonexponential contributions to $\rho(\tau)$ which arise from the approximate master equation treatment presented in I. Finally, there is some discussion of the results of this paper in Sec. VI and of their relevance to the statistical mechanical theory of nonequilibrium processes.

II. THE CONTOUR INTEGRAL

For the evaluation of the integral in Eq. (9), it is necessary to examine the singularities of the integrand:

$$e^{-i\xi\tau}[\xi - (1/\alpha) + 2e^{ip\pi}\csc(p\pi)(\alpha\xi)^{-p}]^{-1}. \tag{11}$$

The exponential factor here enables one to close the integration contour C by a large semicircle in the lower half-plane of ξ which will not contribute to the integral. Provision must be made by a cut for the branch point at $\xi = 0$ caused by the occurrence of $(\alpha\xi)^{-p}$ in the second factor. Since it is then clear that the only other singularities of the expression (11) within the closed contour are poles, consideration of the contour of Fig. 1 will allow the integral to be evaluated by Cauchy's theorem. The poles are located at the zeros of the expression

$$\varphi(\xi) \equiv \xi - 1/\alpha + 2e^{ip\pi}\csc(p\pi)(\alpha\xi)^{-p}.$$

This expression itself has no poles in the finite part of the plane, and so its zeros can be unambiguously located by use of the principle of the argument. It can readily be seen that the variation of the argument of $\varphi(\xi)$ around a contour, described in the positive sense, enclosing that part of the lower half-plane below the cut, is 2π. There is consequently a simple zero at $\xi = \xi_r$, say, in this region. Similarly one finds that there is no zero in the upper half-plane. On the positive real axis, the imaginary part of $\varphi(\xi)$ is $2(\alpha\xi)^{-p}$, which does not vanish. But on the negative real axis, the expression is purely real. Its value tends to $+\infty$ as $\xi \to 0^-$, and to $-\infty$ as $\xi \to -\infty$, and its derivative is always positive in this interval.

There is consequently one further simple zero on the negative real axis at $\xi = \xi_p$, say.

The location of the zero, ξ_r, can be determined as a power series in α. If one solves the equation $\varphi(\xi) = 0$ by Newton's method, using as the first approximant $\xi_0 = 1/\alpha$, then there results

$$\xi_r = (1/\alpha) - 2\omega - 4p\omega^2\alpha - 4\omega^3 p(3p + 1)\alpha^2 + \mathcal{O}(\alpha^3), \tag{12}$$

where we have written for convenience

$$\omega = e^{ip\pi} \csc p\pi = \cot p\pi + i.$$

It is seen at once that, for small α at least, the series, Eq. (12), converges to the zero ξ_r and not to any other zero of $\varphi(\xi)$. Further, for small α, ξ_r is in the right-hand quadrant of the lower half-plane. This implies that the cut in Fig. 1 may be displaced to lie along the imaginary axis. Since for negative ξ, $\varphi(\xi)$ is purely real, and since it has no zeros in the upper half-plane, the principle of reflection ensures that no zero is uncovered in the lower left-hand quadrant by this shift of the cut. As regards ξ_p, Newton's method can again be employed to give a series solution for it. Let us rearrange the equation $\varphi(\xi) = 0$ for negative ξ by putting

$$-\alpha\xi = r, \quad (2\alpha \csc p\pi)^{1/p} = \beta$$

to yield

$$r = \beta(1 + r)^{-1/p}.$$

With first approximant $r_0 = \beta$, the series solution is

$$r = \beta - (1/p)\beta^2 + [(p + 3)/2p^2]\beta^3 + \mathcal{O}(\beta^4),$$

that is,

$$\xi_p = -(1/\alpha)\{\beta - (1/p)\beta^2 + [(p + 3)/2p^2]\beta^3 + \mathcal{O}(\beta^4)\}.$$

Again, it is evident that for small enough α, this converges suitably. There, too, we may see the appearance, through β, of a nonanalyticity in α at $\alpha = 0$, except for the case $1/p$ = integer.

The residues of the integrand (11) at the two poles ξ_r and ξ_p are readily obtained in terms of the locations of these poles. Since each pole is simple, the residues take the form:

$$\mathrm{Res}(\xi = \xi_i) = e^{-i\xi_i\tau}\left[\frac{\partial}{\partial\xi}\{\varphi(\xi)\}\right]^{-1}_{\xi=\xi_i}$$

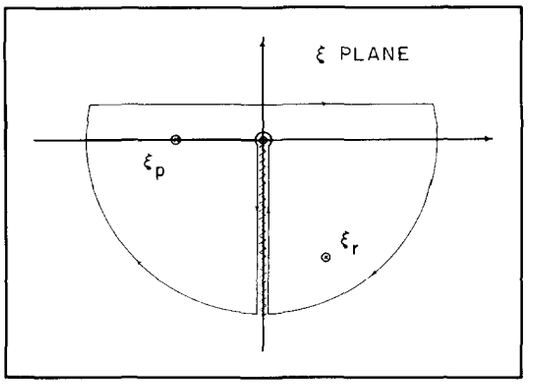


FIG. 2. The contour used in accounting for the contribution from the cut. See the discussion prefacing Eq. (14).

$$= e^{-i\xi_i\tau}[1 - 2p\omega\alpha(\alpha\xi_i)^{-p-1}]^{-1}. \tag{13}$$

If Eq. (13) is expanded out in powers of α for the residue at ξ_r, by using Eq. (12), one obtains

$$\exp(i\tau)[(1/\alpha) - 2\cot p\pi + \mathcal{O}(\alpha)]\exp(-2\tau) \times \{1 + 2p\omega\alpha + 4p(2p + 1)\omega^2\alpha^2 + \mathcal{O}(\alpha^3)\},$$

whence it is clear that the weak-coupling limit, Eq. (10), of $\rho(\tau)$, which involves no nonanalyticity in α, comes in its entirety from this residue.

When Cauchy's theorem is applied to calculate the integral of Eq. (9), in addition to the two residues given by Eq. (13), account must be taken of the contribution from the cut. This is most simply done using the contour of Fig. 2. The result is

$$-\frac{1}{2\pi i}\int_c d\xi \frac{e^{-i\xi\tau}}{\varphi(\xi)} = \mathrm{Res}(\xi = \xi_r) + \mathrm{Res}(\xi = \xi_p) - \frac{1}{2\pi i}\left(\int_0^{-i\infty} d\xi \frac{e^{-i\xi\tau}}{\varphi^{(r)}(\xi)} - \int_0^{-i\infty} d\xi \frac{e^{-i\xi\tau}}{\varphi^{(l)}(\xi)}\right), \tag{14}$$

where $\varphi^{(r)}(\xi)$ is the value assumed by $\varphi(\xi)$ for ξ on the right-hand side of the cut and $\varphi^{(l)}(\xi)$ for the left-hand side. A little rearrangement of the cut contribution yields for it the explicit expression:

$$\frac{1}{2\pi}\int_0^\infty dx e^{-x\tau/\alpha}\left(\frac{-1 + 2\alpha\cos(3p\pi/2)\csc(p\pi)x^{-p} + i[x - 2\alpha\sin(3p\pi/2)\csc(p\pi)x^{-p}]}{[-1 + 2\alpha\cos(3p\pi/2)\csc(p\pi)x^{-p}]^2 + [x - 2\alpha\sin(3p\pi/2)\csc(p\pi)x^{-p}]^2} - \frac{-1 + 2\alpha\cos(p\pi/2)\csc(p\pi)x^{-p} + i[x + 2\alpha\sin(p\pi/2)\csc(p\pi)x^{-p}]}{[-1 + 2\alpha\cos(p\pi/2)\csc(p\pi)x^{-p}]^2 + [x + 2\alpha\sin(p\pi/2)\csc(p\pi)x^{-p}]^2}\right). \tag{15}$$

For general p, this contribution can be studied best by numerical computation, but for the special case $p = \frac{1}{2}$, an explicit closed analytical form for Eq. (14) will be derived in the next section, which will manifest most of the interesting properties of the general case. It can also be seen by use of the Tauberian theorems for inverse Laplace transforms (see, for example, Ref. 5) that the asymptotic behavior of the expression (15) as $\tau \to \infty$ will be proportional to $\tau^{-(1+p)}$.

III. THE CASE $p = 1/2$

On setting $p = \frac{1}{2}$ in Eq. (9), we obtain for the solution of our problem

$$\langle \mathfrak{R} | \Psi(\tau) \rangle = -\frac{1}{2\pi i}\int_c d\xi e^{-i\xi\tau}\frac{1}{\xi - (1/\alpha) + 2i(\alpha\xi)^{-1/2}}.$$

For the evaluation of this integral, a somewhat different contour from that used in the last section will be considered. It is shown in Fig. 3, and its advantage is that its cut contribution can readily be expressed in terms of the error function. First of all, it is necessary to find closed expressions for the poles ξ_r and ξ_p.

The equation

$$\xi - (1/\alpha) + 2i(\alpha\xi)^{-1/2} = 0$$

can be reduced to a simple cubic by making the substitution

$$y = -i(\alpha\xi)^{1/2}. \tag{16}$$

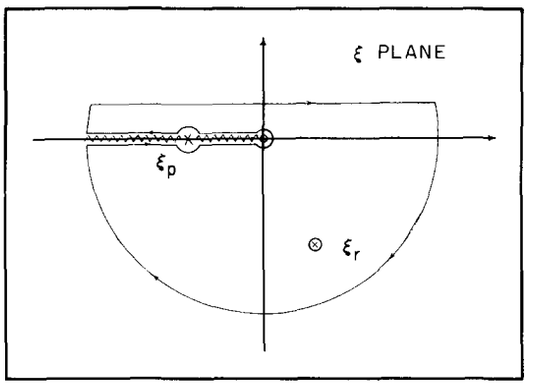


FIG. 3. The contour used in the study of Eq. (9) for the choice $p = \frac{1}{2}$.

This yields the equation:

$$y^3 + y - 2\alpha = 0. \tag{17}$$

To obtain the real root of this equation, one may set $y = u + v$, and it can be seen that y will be this root if u and v satisfy the equations

$$u^3 + v^3 = 2\alpha, \quad uv = -\tfrac{1}{3},$$

that is, if u^3 and v^3 are the roots of the quadratic

$$\Omega^2 - 2\alpha\Omega - \tfrac{1}{27} = 0.$$

Hence,

$$y = u + v = [\alpha + (\alpha^2 + \tfrac{1}{27})^{1/2}]^{1/3} + [\alpha - (\alpha^2 + \tfrac{1}{27})^{1/2}]^{1/3}.$$

We shall define

$$\theta_1 = u + v, \quad \theta_2 = u - v \tag{18}$$

and then the left-hand side of Eq. (17) can be factorized to yield the expression

$$(y - \theta_1)[y + \tfrac{1}{2}(\theta_1 + i\theta_2\sqrt{3})][y + \tfrac{1}{2}(\theta_1 - i\theta_2\sqrt{3})]. \tag{19}$$

From this result and Eq. (16) it is easily seen that the poles ξ_r and ξ_p are given by

$$\xi_p = -\theta_1^2/\alpha,$$
$$\xi_r = -(1/4\alpha)(\theta_1 + i\theta_2\sqrt{3})^2 = (1/2\alpha)(\theta_1^2 + 2 - i\theta_1\theta_2\sqrt{3}),$$

where, in the last equality, use has been made of the relation

$$\theta_2^2 = \theta_1^2 + \tfrac{4}{3},$$

which follows from the definitions (18). Further properties of θ_1 and θ_2 which will be needed are

$$\theta_2 > \theta_1 > 0.$$

It is clear from this that, with $p = \frac{1}{2}$, ξ_r is always in the lower right-hand quadrant, whatever α.

We may now write down the two pole contributions to $\langle \mathfrak{A} | \Psi(\tau) \rangle$ at once from Eq. (13). They are

$$\tfrac{1}{2}\ \mathrm{Res}(\xi = \xi_p) = \tfrac{1}{2} e^{i\theta_1^2\tau/\alpha}\theta_1^3/(\theta_1^3 + \alpha),$$
$$\mathrm{Res}(\xi = \xi_r) = \exp[(i\tau/4\alpha)(\theta_1 + i\theta_2\sqrt{3})^2]$$
$$\times\{1 + [\alpha/2(\theta_1^2 + 1)^3]\{(2\theta_1^2 + 3)\theta_1 - i\theta_2\sqrt{3}\}\}^{-1},$$

Only a half of the residue at ξ_p has been taken here of course, because of the use of the contour of Fig. 3. These two contributions can be put into a more convenient form by noticing that Eqs. (18) can be inverted to yield α in terms of θ_1:

$$\alpha = \tfrac{1}{2}\theta_1(\theta_1^2 + 1).$$

Some manipulation then gives, for the sum of the two expressions,

$$[1/(3\theta_1^2 + 1)]\{\theta_1^2 \exp(i\theta_1^2\tau/\alpha) + (2\theta_1^2 + 1 + i\theta_1/\theta_2\sqrt{3})$$
$$\times \exp[(i\tau/4\alpha)(\theta_1 + i\theta_2\sqrt{3})^2]\}. \tag{20}$$

There remains to be considered the cut contribution to $\langle \mathfrak{A} | \Psi(\tau) \rangle$. This is the integral of the discontinuity of

$$[\xi - (1/\alpha) + 2i(\alpha\xi)^{-1/2}]^{-1}$$

across the cut of Fig. 3. It is easily seen to be

$$-\frac{4}{2\pi i}\,\mathcal{P}\int_0^\infty d\xi\,\frac{(\alpha\xi)^{1/2}e^{i\xi\tau}}{4 - \alpha\xi(\xi + 1/\alpha)^2},$$

where the symbol $\mathcal{P}$ denotes the Cauchy principal part. The change of variable,

$$y = (\alpha\xi)^{1/2},$$

analogous to Eq. (16), puts the integral into a tractable form:

$$\frac{4\alpha i}{\pi}\,\mathcal{P}\int_0^\infty dy\,\frac{y^2 e^{iy^2\tau/\alpha}}{4\alpha^2 - y^2(y^2 + 1)^2}$$
$$= \frac{1}{2\pi i}\left(\mathcal{P}\int_{-\infty}^{+\infty} dy\,\frac{y^2 e^{iy^2\tau/\alpha}}{y^3 + y - 2\alpha} - \mathcal{P}\int_{-\infty}^{+\infty} dy\,\frac{y^2 e^{iy^2\tau/\alpha}}{y^3 + y + 2\alpha}\right). \tag{21}$$

If we denote by $I(\alpha_1\tau)$ the integral

$$\pm\,\mathcal{P}\int_{-\infty}^{+\infty} dy\,\frac{e^{iy^2\tau/\alpha}}{y^3 + y \pm 2\alpha},$$

then

$$-i\alpha\,\frac{\partial}{\partial\tau}I(\alpha_1\tau) = \pm\,\mathcal{P}\int_{-\infty}^{+\infty} dy\,\frac{y^2 e^{iy^2\tau/\alpha}}{y^3 + y \pm 2\alpha}$$

and so the expression Eq. (21) is just

$$\frac{\alpha}{\pi}\,\frac{\partial}{\partial\tau}I(\alpha, \tau). \tag{22}$$

The denominator $y^3 + y - 2\alpha$ can be factorized as in Eq. (19), and this factorization permits the evaluation of $I(\alpha_1\tau)$ to be reduced, by the decomposition of the integrand into partial fractions, to the consideration of interals of the form

$$\mathcal{P}\int_{-\infty}^{+\infty} dx\,\frac{e^{-x^2}}{x + a} \quad (a \text{ real}).$$

This integral, which occurs in the theory of the plasma dispersion function,[6] is related to the error function, and its value is

$$-\pi i e^{-a^2}\,\mathrm{erf}(ia), \tag{23}$$

where the erf function is defined as follows[7]:

$$\mathrm{erf}z = \int_0^z e^{-t^2}dt = \sum_{n=0}\frac{(-1)^n z^{2n+1}}{n!(2n + 1)}.$$

For complex a, this result can be extended:

$$\int_{-\infty}^{+\infty} dx\, \frac{e^{-x^2}}{x+a} = -\pi i e^{-a^2}[\mathrm{erf}(ia)+1] \quad (\mathrm{Im}a > 0) \tag{24a}$$

$$= -\pi i e^{-a^2}[\mathrm{erf}(ia)-1] \quad (\mathrm{Im}a < 0). \tag{24b}$$

Now the decomposition into partial fractions is

$$\frac{1}{y^3+y-2\alpha} = \frac{1}{3\theta_1^2+1} \times \left[\frac{1}{y-\theta_1} - \tfrac{1}{2}\left(1+\frac{i\theta_1\sqrt{3}}{\theta_2}\right)\frac{1}{y+\frac{1}{2}(\theta_1+i\theta_2\sqrt{3})} - \tfrac{1}{2}\left(1-\frac{i\theta_1\sqrt{3}}{\theta_2}\right)\frac{1}{y+\frac{1}{2}(\theta_1-i\theta_2\sqrt{3})}\right].$$

Next, by the use of Eqs. (23) and (24) one obtains the results

$$\mathcal{P}\int_{-\infty}^{+\infty} dy\, \frac{e^{iy^2\tau/\alpha}}{y-\theta_1} = \pi i e^{i\theta_1^2\tau/\alpha}\,\mathrm{erf}\left[\left(\frac{\tau}{\alpha}\right)^{1/2}\theta_1 e^{i\pi/4}\right] \times \int_{-\infty}^{+\infty} dy\, \frac{e^{iy^2\tau/\alpha}}{y+\frac{1}{2}(\theta_1+i\theta_2\sqrt{3})}$$

$$= -\pi i \exp\left(\frac{i\pi}{4\alpha}(\theta_1+i\theta_2\sqrt{3})^2\right) \times \left\{\mathrm{erf}\left[\left(\frac{\tau}{\alpha}\right)^{1/2}\cdot\tfrac{1}{2}(\theta_1+i\theta_2\sqrt{3})e^{i\pi/4}\right]+1\right\}$$

and

$$\int_{-\infty}^{+\infty} dy\, \frac{e^{iy^2\tau/\alpha}}{y+\frac{1}{2}(\theta_1-i\theta_2\sqrt{3})} = \pi i \exp\left(\frac{i\tau}{4\alpha}(\theta_1-i\theta_2\sqrt{3})^2\right) \times \left\{\mathrm{erf}\left[\left(\frac{\tau}{\alpha}\right)^{1/2}\tfrac{1}{2}(i\theta_2\sqrt{3}-\theta_1)e^{i\pi/4}\right]+1\right\}.$$

Thus it follows that

$$I(\alpha_1\tau) = \frac{\pi i}{3\theta_1^2+1}\left[-e^{i\theta_1^2\tau/\alpha}\,\mathrm{erf}\left[\left(\frac{\tau}{\alpha}\right)^{1/2}\theta_1 e^{i\pi/4}\right] - \tfrac{1}{2}\left(1+\frac{i\theta_1\sqrt{3}}{\theta_2}\right)\exp\left(\frac{i\tau}{4\alpha}(\theta_1+i\theta_2\sqrt{3})^2\right) \times\left\{\mathrm{erf}\left[\left(\frac{\tau}{\alpha}\right)^{1/2}\tfrac{1}{2}(\theta_1+i\theta_2\sqrt{3})e^{i\pi/4}\right]+1\right\} + \tfrac{1}{2}\left(1-\frac{i\theta_1\sqrt{3}}{\theta_2}\right)\exp\left(\frac{i\tau}{4\alpha}(\theta_1-i\theta_2\sqrt{3})^2\right) \times\left\{\mathrm{erf}\left[\left(\frac{\tau}{\alpha}\right)^{1/2}\tfrac{1}{2}(i\theta_2\sqrt{3}-\theta_1)e^{i\pi/4}\right]+1\right\}\right],$$

whence, since

$$\frac{d}{dx}\,\mathrm{erf}x = e^{-x^2},$$

there results from Eq. (22)

$$\text{cut contribution} = \frac{\alpha}{\pi}\frac{\partial}{\partial\tau}I(\alpha_1\tau)$$

$$= \frac{1}{2(3\theta_1^2+1)}\left[2\theta_1^2\exp\left(-\frac{i\theta_1^2\tau}{\alpha}\right)\mathrm{erf}\left[\left(\frac{\tau}{\alpha}\right)^{1/2}\theta_1 e^{i\pi/4}\right] - \left(2\theta_1^2+1+\frac{i\theta_1}{\theta_2\sqrt{3}}\right)\exp\left(\frac{i\tau}{4\alpha}(\theta_1+i\theta_2\sqrt{3})^2\right) \times\left\{\mathrm{erf}\left[\left(\frac{\tau}{\alpha}\right)^{1/2}\tfrac{1}{2}(\theta_1+i\theta_2\sqrt{3})e^{i\pi/4}\right]+1\right\} + \left(2\theta_1^2+1-\frac{i\theta_1}{\theta_2\sqrt{3}}\right)\exp\left(\frac{i\tau}{4\alpha}(\theta_1-i\theta_2\sqrt{3})^2\right) \times\left\{\mathrm{erf}\left[\left(\frac{\tau}{\alpha}\right)^{1/2}\tfrac{1}{2}(i\theta_2\sqrt{3}-\theta_1)e^{i\pi/4}\right]+1\right\}\right] \tag{25}$$

It is perhaps worthwhile to remark that when τ is set equal to zero in Eqs. (20) and (25), all the contributions to $\langle\mathfrak{A}|\Psi(0)\rangle$ add up to unity as they should. To see this, one has that Eq. (20) gives

$$\frac{1}{3\theta_1^2+1}\left(3\theta_1^2+1+\frac{i\theta_1}{\theta_2\sqrt{3}}\right) = 1+\frac{1}{3\theta_1^2+1}\left(\frac{i\theta_1}{\theta_2\sqrt{3}}\right),$$

whereas Eq. (25) gives

$$\frac{1}{2(3\theta_1^2+1)}\left[-\left(2\theta_1^2+1+\frac{i\theta_1}{\theta_2\sqrt{3}}\right)+\left(2\theta_1^2+1-\frac{i\theta_1}{\theta_2\sqrt{3}}\right)\right] = -\frac{1}{(3\theta_1^2+1)}\left(\frac{i\theta_1}{\theta_2\sqrt{3}}\right)$$

since $\mathrm{erf}(0) = 0$.

The asymptotic behavior and limit of $\langle\mathfrak{A}|\Psi(\tau)\rangle$ as $\tau\to\infty$ is of interest in determining the nature of the spontaneous emission of the excited atom. First of all, one can see that of the two terms in Eq. (20), the first is a purely oscillatory exponential which does not decay as $\tau\to\infty$. It will give rise to a constant contribution to the probability

$$\rho(\tau) = |\langle\mathfrak{A}|\Psi(\tau)\rangle|^2.$$

The second term has superimposed on its oscillatory part an exponential decay governed by the factor

$$\exp[-(\theta_1\theta_2\sqrt{3}/2\alpha)\tau],$$

to lowest order in α as can be seen by the definitions of θ_1 and θ_2, Eq. (18); this factor is

$$e^{-2\tau},$$

and is responsible, as remarked earlier, for the decay which is found in the weak-coupling solution

$$\rho(\tau) = e^{-4\tau}.$$

The cut contribution is more complicated, and can best be investigated by introducing the error function of complex argument,[7] $w(z)$, defined by the relation

$$w(z) = e^{-z^2}[1+\mathrm{erf}(iz)].$$

This function has the asymptotic development, valid as $|z|\to\infty, -\pi/4 < \arg z < 5\pi/4$:

$$w(z) \sim \frac{i}{z\sqrt{\pi}}\left[1+\sum_{m=1}^{\infty}\frac{1\cdot 3\cdot 5\ldots(2m-1)}{(2z^2)^m}\right]. \tag{26}$$

Now, Eq. (25) may be rewritten as

$$\frac{1}{2(3\theta_1^2+1)}\left[2\theta_1^2\exp\left(-\frac{i\theta_1^2\tau}{\alpha}\right)-2\theta_1^2 w\left[\left(\frac{\tau}{\alpha}\right)^{1/2}\theta_1 e^{3\pi i/4}\right]\right.$$

$$- \left(2\theta_1^2 + 1 - \frac{i\theta_1}{\theta_2\sqrt{3}}\right) w\left[\left(\frac{\tau}{\alpha}\right)^{1/2} \tfrac{1}{2}(\theta_1 + i\theta_2\sqrt{3})e^{-\pi i/4}\right]$$

$$+ \left(2\theta_1^2 + 1 - \frac{i\theta_1}{\theta_2\sqrt{3}}\right) w\left[\left(\frac{\tau}{\alpha}\right)^{1/2} \tfrac{1}{2}(\theta_1 - i\theta_2\sqrt{3})e^{3\pi i/4}\right] .$$

Again, the first term in square brackets in this expression is a nondecaying purely oscillatory exponential, equal to the term of this kind in Eq. (26). When the other terms are expanded using Eq. (22), it can be seen that the lowest order term, proportional to $\tau^{-1/2}$, has a vanishing coefficient, so that the asymptotic behavior comes from the next terms, which are proportional to $\tau^{-3/2}$, in accord with the remark at the end of Sec. II.

Lastly, in this section, we shall write out our solution for $\langle \mathfrak{A} | \Psi(\tau) \rangle$ in a compact form suitable for the numerical calculations described in Sec. V. When Eqs. (20) and (25) are grouped together, there results

$$\langle \mathfrak{A} | \Psi(\tau) \rangle = \frac{1}{2(3\theta_1^2 + 1)}$$

$$\times \left[2\theta_1^2 \exp\left(\frac{i\theta_1^2\tau}{\alpha}\right) \left\{\operatorname{erf}\left[\left(\frac{\tau}{\alpha}\right)^{1/2} \theta_1 e^{i\pi/4}\right] + 1\right\}\right.$$

$$+ \left(2\theta_1^2 + 1 + \frac{i\theta_1}{\theta_2\sqrt{3}}\right) \exp\left(\frac{i\tau}{4\alpha}(\theta_1 + i\theta_2\sqrt{3})^2\right)$$

$$\times \left\{\operatorname{erf}\left[-\left(\frac{\tau}{\alpha}\right)^{1/2} \tfrac{1}{2}(\theta_1 + i\theta_2\sqrt{3})e^{\pi i/4}\right] + 1\right\}$$

$$+ \left(2\theta_1^2 + 1 - \frac{i\theta_1}{\theta_2\sqrt{3}}\right) \exp\left(\frac{i\tau}{4\alpha}(\theta_1 - i\theta_2\sqrt{3})^2\right)$$

$$\left.\times \left\{\operatorname{erf}\left[\left(\frac{\tau}{\alpha}\right)^{1/2} \tfrac{1}{2}(-\theta_1 + i\theta_2\sqrt{3})e^{\pi i/4}\right] + 1\right\}\right] \tag{27}$$

IV. CRITERIA FOR ERGODICITY

A somewhat surprising feature in the analysis of the last two sections has been the appearance of the pole at $\xi = \xi_p$, on the negative real axis, in the integrand, Eq. (11). The residue from this pole gives rise, as we saw in the last section, to a constant contribution to $\rho(\tau)$, which persists after all the other terms, proportional to either a decaying exponential or to $\tau^{-(1+p)}$, have become arbitrarily small. The presence of this feature means that, contrary to the burden of the remarks in Sec. I, $\rho(\tau)$ is not, in this model, an ergodic function, since, inasmuch as the system we have considered is essentially at zero temperature in the thermodynamic limit, the time average of

$$\bar{\rho} = \lim_{T\to\infty} \frac{1}{T} \int_0^T d\tau \rho(\tau)$$

would be zero were the function ergodic. It is therefore of interest to seek the origin of the pole at ξ_p, and to determine which property of our model prevents the appearance of an ergodic $\rho(\tau)$. It should be remarked that it is sufficient to study the function $\rho(\tau)$ in this regard in order to make a connection with the work of Mazur,[4] where in a quantum-mechanical system, an observable, to be ergodic, needs to have the appropriate property for its associated *operator*. For it is clear from Eqs. (IV. 10, 11) that if the probability amplitude $\langle \mathfrak{A} | \Psi(\tau) \rangle$ tends to zero as $\tau \to \infty$, then so does $\langle \lambda | \Psi(\tau) \rangle$ for all λ, and in consequence also the full density matrix of the system. Thus consideration of $\rho(\tau)$ above is enough to enable one to decide if the Heisenberg operator $|\mathfrak{A}\rangle\langle\mathfrak{A}|$ is ergodic.

There are two clues to the problem. The first comes from quantum field theory, and has to do with the occurrence of ghost states. It was remarked in I that the model treated in these papers is closely related to the Lee model of field theory, and may be considered as a special case of it, with infinite-mass N and V particles. (For a discussion of the Lee model and of ghost states in it, see Refs. 8 and 9.) Now the ghost state of the Lee model is characterized as being a state of negative energy which is needed to complete the Hilbert space of the problem and which is associated with a negative norm, that is, the probability associated with it, in the expansion of a state of the system in terms of the complete set of eigenstates of the Hamiltonian, is negative. Negative probabilities cannot occur in our treatment of course, because all that has been done is to go to the limit of infinite size in a well-posed quantum-mechanical problem [that is, as prescribed by the Hamiltonian Eq. (1) and the boundary conditions] in which all probabilities are nonnegative by construction. But, although the ghost state, with negative norm, can only arise in an approach based on interacting fields, with the consequent complications of renormalization, both of mass and charge, it is fairly clear that the same property of the Hamiltonian is at work in producing such a state and also in giving rise to our pole at ξ_p. This remark will be made more explicit below. Here, it can be noted that in field theory, ghosts are related to particular choices of the form factor for the interaction term in the Hamiltonian. In our discussion, the "form factor" is essentially the quantity $|h_\lambda|^2$.

The second clue is found in the work of Cukier and Mazur,[3] who found that the dynamical variable that they examined could be either ergodic or not depending on the magnitude of the coupling constant in their problem.

When ergodicity failed, it was because, as here, a pole separated from an otherwise continuous spectrum gave rise to nondecaying contributions. In their work, however, which deals with an impurity atom in a harmonic chain, the separated pole corresponds to a frequency *greater* than that of the other modes of the system, rather than to a *negative* energy, and it is known from solid state physics and lattice dynamics[10] that such "isolated modes" do indeed occur.

That the pole at ξ_p in fact does correspond to a special eigenstate of the Hamiltonian, an isolated mode of the exact motion, can be seen by examining the eigenvalue spectrum of the Hamiltonian for a finite system. This has been studied in IV, where the nature of the roots of the dimensionless secular equation [compare Eqs. (8) and (9)]

$$\xi - (1/\alpha) - \sigma(\xi) = 0 \tag{28}$$

was examined. It was found that the lowest root, ξ_0, say, was always less than the smallest one-photon (dimensionless) energy of the unperturbed radiation field. Since the derivative $\sigma'(\xi)$ is always negative, and since $\sigma(\xi)$ has no poles for $\xi \leq 0$ and tends to 0^- as $\xi \to -\infty$, it follows that ξ_0 will be negative if (and only if)

$$\sigma(0) < -1/\alpha. \tag{29}$$

All other roots of Eq. (28) will of course always be positive. But if $\sigma(0)$ is looked at as a function of the length L of the system [see Eq. (6)], then it is seen at once that

$\sigma(0) \to -\infty$ as $L \to \infty$. Thus above a critical length, determined by

$$\sigma(0) = -1/\alpha,$$

the quantity ξ_0 will be negative, and it is clear that ξ_p is simply the limit of ξ_0 as $L \to \infty$, being in fact a zero of

$$\xi - (1/\alpha) - \sigma_\infty(\xi).$$

This analysis parallels that leading to the identification of the ghost state in the Lee model.

With the "form factor" of our Hamiltonian given, as it has been throughout this paper, by Eq. (5), it can now be seen that the isolated pole ξ_p is unavoidable, whatever α may be, since $\sigma_\infty(0) = -\infty$, and the condition, Eq. (29), is always satisfied. But it is possible, as with the Lee model, to find other choices of $|h_\lambda|^2$ which will remove the difficulty, at least for sufficiently small α. To achieve this, it is necessary and sufficient that $\sigma_\infty(0)$ be finite (it is always negative and cannot vanish), and then for

$$\alpha < -1/\sigma_\infty(0)$$

there will be no "ghost state." If one tries the choice of $|h_\lambda|^2$ used in I and II, namely

$$|h_\lambda|^2 = \hbar^2 \alpha c E/L, \tag{30}$$

then $\sigma(\xi)$ will diverge unless the summation which defines it, Eq. (6), is truncated at some upper bound for the one-photon energies. Such a procedure would yield [see Eq. (7)]:

$$\sigma_\infty(\xi) = \frac{2}{\pi} \int_0^\mu d\rho \, \frac{1}{(\xi - \rho)}$$

$$= \frac{2}{\pi} \log\left(\frac{\xi}{\xi - \mu}\right)$$

with μ as the (dimensionless) cutoff. But $\sigma_\infty(0)$ is still unbounded here, and, worse, the expression $\xi - (1/\alpha) - \sigma_\infty(\xi)$ has another isolated zero for $\xi > \mu$, of the type found by Cukier and Mazur. This is the case, too, for another choice of $|h_\lambda|^2$ mentioned in I [Eq. (I-11)]:

$$|h_\lambda|^2 = \hbar \alpha c \omega_\lambda / L, \tag{31}$$

where again a cutoff is needed for the convergence of Eq. (6). For this choice, however, the "ghost state" pole is removed, since $\sigma_\infty(0)$ is now finite, as it is for any $|h_\lambda|^2$ which behaves like a positive power of ω_λ near $\omega_\lambda = 0$. Thus no isolated pole will appear if $|h_\lambda|^2$ has this property for small ω_λ and yet remains nonzero (except perhaps at discrete points) for all finite ω_λ. An example of such a $|h_\lambda|^2$ may be cited, which has the same value, $\hbar^2 \alpha c E/L$, for $\omega_\lambda = E$ as in Eqs. (5), (30), and (31). It is

$$|h_\lambda|^2 = \begin{cases} \hbar^2 \alpha c \omega_\lambda / L & \text{for} \quad \omega_\lambda \leq E \qquad (32a) \\ \hbar^2 \alpha c E^{3/2} / L \omega_\lambda^{1/2} & \text{for} \quad \omega_\lambda \geq E. \qquad (32b) \end{cases}$$

This choice yields

$$\sigma_\infty(\xi) = \frac{2\alpha\xi}{\pi} \log\left(\frac{\alpha\xi}{\alpha\xi - 1}\right) - \frac{2}{\pi} + \frac{2}{\pi(\alpha\xi)^{1/2}} \log\left(\frac{1 - \alpha\xi}{1 + \alpha\xi + 2(\alpha\xi)^{1/2}}\right),$$

whence $\sigma_\infty(0) = -6/\pi$, and so for $\alpha < \pi/6$, the absence of isolated poles implies that the only contributions in Eq. (9) to $\langle \mathfrak{N} | \Psi(\tau) \rangle$ will be those from cuts and the pole which corresponds to the ξ_r of our analysis. In a system with form factor given by Eq. (32), then, $\rho(\tau)$ will indeed be an ergodic function.

V. NUMERICAL CALCULATIONS

In order to have an idea of the relative importance of the various contributions, discussed in the preceding sections, to the probability $\rho(\tau)$ of finding the two-level atom excited at a time τ, a numerical investigation has been performed based on Eq. (27). The real and imaginary parts of $\langle \mathfrak{N} | \Psi(\tau) \rangle$ were separately calculated from this equation, and from these results $\rho(\tau)$ itself was obtained. This function is plotted in Figs. 4 and 5 for the choices $\alpha = 0.1$ and $\alpha = 0.8$, respectively; included in these figures for comparison is the corresponding weak-coupling solution obtained in III [see Eq. (III. 27)], which,

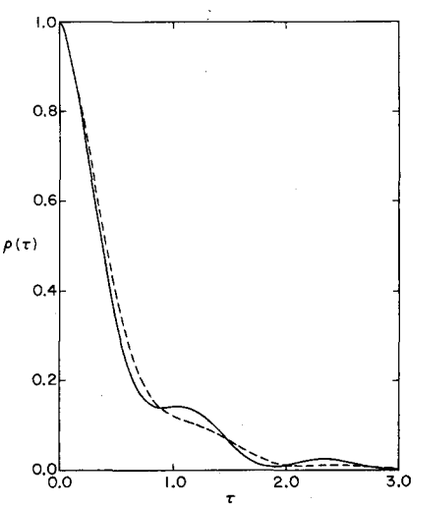


FIG. 4. A plot of $\rho(\tau)$ versus τ for $\alpha = 0.1, p = \frac{1}{2}$. The solid line represents the exact solution derived from Eq. (27), and the dashed line represents the Schrödinger weak-coupling solution given by Eq. (33). " For clarity the horizontal scale has been expanded by a factor of two."

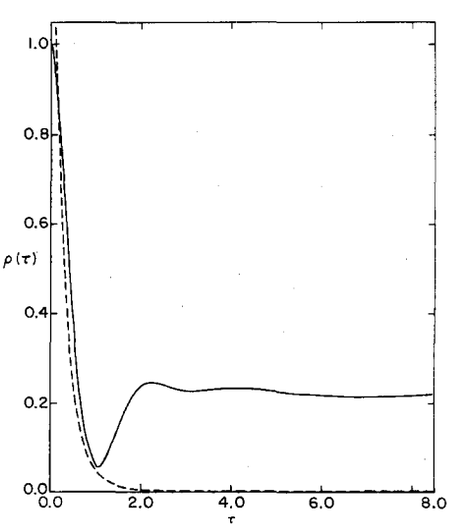


FIG. 5. A plot of $\rho(\tau)$ versus τ for $\alpha = 0.8, p = \frac{1}{2}$. The solid line represents the exact solution derived from Eq. (27), and the dashed line represents the Schrödinger weak-coupling solution given by Eq. (33).

for $p = \frac{1}{2}$, assumes the following form:

$$\rho(\tau) = \left[e^{-2\tau} + \frac{2}{\pi}\int_0^\alpha d\xi \sin\left(\frac{\tau}{\xi}\right)\left(\frac{\xi}{\alpha-\xi}\right)^{1/2}\right]^2 + \left[\alpha e^{-2\tau} - \frac{2}{\pi}\int_0^\alpha d\xi \cos\left(\frac{\tau}{\xi}\right)\left(\frac{\xi}{\alpha-\xi}\right)^{1/2}\right]^2. \quad (33)$$

For small values of α, it is clear that the constant term in $\rho(\tau)$, arising from the "ghost state" is very small, and the departure of $\rho(\tau)$ from ergodicity cannot be seen in the figure. On the other hand, for $\alpha = 0.8$, this contribution is significant and it can be seen that even over the time range shown, it rapidly becomes dominant. It is readily noted from Sec. III that the value of the contribution is

$$\left(\frac{2\theta_1^2}{3\theta_1^2+1}\right)^2, \quad (34)$$

that is, the square of the modulus of the full residue at $\xi = \xi_p$. For $\alpha = 0.8$, this quantity is 0.22.

The most noticeable feature of both these plots, however, is the appearance of nonexponential decay, coming from the cut integral in $\langle \mathfrak{N} | \Psi(\tau) \rangle$. This phenomenon is quite in agreement with the predictions of I, where, although nothing like the constant term, Eq. (34), can be found in the weak-coupling theory, nonexponential contributions which are also nonanalytic in α at $\alpha = 0$ are significant in $\rho(\tau)$. It may be remarked in passing that such nonanalyticity is evident in the exact solution of this paper, Eq. (27), through the arguments of the error functions, proportional to $\alpha^{-1/2}$. In view of these comments, the purely exponential part of $\rho(\tau)$, as calculated from Eq. (27), namely

$$\frac{4(\theta_1^2+1)^2}{3\theta_1^2+4}\exp[-(\theta_1\theta_2\sqrt{3}/\alpha)\tau],$$

has been substracted from $\rho(\tau)$, along with the constant term, Eq. (34), so as to make a quantitative comparison with the predictions of I.

A different choice of form factor was made in I from that used here. There, $|h_\lambda|^2$ was taken as given by Eq. (30), and, moreover, no cutoff was found to be necessary in the approximation scheme of that paper. So that a proper numerical comparison might be effected, the form factor used here, Eq. (5), was put into the formulas of I, and it follows easily that the nonexponential contribution to $\rho(\tau)$ as computed by those methods is

$$-\frac{1}{\pi}\int_0^\alpha d\xi \times (2\cos(\tau/\xi)[\xi^p(\xi+\alpha)^{-p} + 2\xi^p(\alpha-\xi)^{-p}\cos p\pi] + \sin(\tau/\xi)\{\xi^{-1} + 2[\xi^p(\xi+\alpha)^{-p}\cos p\pi - \xi^p(\alpha-\xi)^{-p}\cos 2p\pi]\csc p\pi\})/ \times (4\xi^2[\xi^p(\xi+\alpha)^{-p} + 2\xi^p(\alpha-\xi)^{-p}\cos p\pi]^2 + \{1 + 2\xi[\xi^p(\xi+\alpha)^{-p}\cos p\pi - \xi^p(\alpha-\xi)^{-p}\cos 2p\pi]\csc p\pi\}^2) - (2\cos(\tau/\xi)\xi^p(\xi+\alpha)^{-p} + \sin(\tau/\xi)\{\xi^{-1} + 2[\xi^p(\xi+\alpha)^{-p}\cos p\pi - \xi^p(\alpha-\xi)^{-p}]\csc p\pi\})/(4\xi^2[\xi^{2p}(\xi+\alpha)^{-2p}] + \{1 + 2\xi[\xi^p(\xi+\alpha)^{-p}\cos p\pi - \xi^p(\alpha-\xi)^{-p}]\csc p\pi\}^2). \quad (35)$$

For the choice $p = 0$, this expression reduces to Eq. (I. 62). For the choice $p = \frac{1}{2}$, in Figs. 6 and 7, the two nonexponential decays, form the exact solution and from Eq. (35), are plotted together, for $\alpha = 0.1$ and $\alpha = 0.8$, respectively. It is obvious at first glance that despite certain similarities in the behavior of the decays for small α, nothing like quantitative agreement is achieved. Considerable differences exist in the time scales over which they relax to zero, and for large α in the times at which they attain their maxima and minima. For $\alpha = 0.8$, since the theory of I is intended for weak coupling, this is not surprising, but for $\alpha = 0.1$, it is evident that there is a substantial and significant disagreement.

On the basis of the comparisons given in Figs. 4 and 5, it is possible to conclude that the results obtained using Eq. (33) are in good qualitative agreement with those obtained using the exact solution, Eq. (27), for the choice $p = \frac{1}{2}$; furthermore, given the comparisons presented in Figs. 6 and 7, it is possible to maintain that Eq. (33) is in better quantitative agreement with the exact solution than is Eq. (35). These conclusions strengthen the argument stated in IV (also given at the beginning of the next section) that the approach given in III is better all around than the approach adopted in I and II.

VI. DISCUSSION AND CONCLUSIONS

The effort of this paper has been directed towards understanding the properties of the quantum-mechanical solution of our problem of spontaneous emission in the thermodynamic limit, and the relations that exist between this solution and that of the master equation. It has been seen that the master equation, while reproducing certain qualitative features of the exact solution, fails to yield quantitative agreement for any of the nonexponential parts of it, in just the same way as it failed, as was seen in IV, to describe properly the behavior of a finite system. This tends, then, to reinforce the conclusion of IV that the master equation, having been constructed specifically to deal with large systems with weak coupling, cannot be pressed into effective service outside this regime.

In a sense, once one has seen the nature of the exact thermodynamic solution to the problem with $p = \frac{1}{2}$, Eq. (6), then it is evident that the master equation, although possessed, through its non-Markovian property [see Eq. (I. 24)], of a great deal more structure than say, a simple gain–loss equation of the Pauli type, cannot yield such a variety of behavior as is manifested in the exact solution. For the links between ergodicity of the function $\rho(\tau)$ and nonanalyticity, in the thermodynamic limit, of certain contributions to it, and of the important quantity $\sigma_\infty(\xi)$ [Eq. (9)], have been seen to be rather complex, depending on the properties of the form factor $|h_\lambda|^2$ at zero and infinity, as well as on the size of the coupling as measured by α. Nonetheless, the qualitative feature, predicted by the non-Markovian master equation, of nonexponential decay, a feature which does *not* appear in any simpler approach to spontaneous emission, is without any doubt a real property of the exact solution.

The problem of the ghost state in models of the kind we have discussed is a difficult one, but the point of greatest interest for present purposes is the rather direct connection that has appeared between the occurrence of such a state and the failure of ergodicity. If a form factor is chosen so that no ghost appears, or if coupling is sufficiently weak that its contribution to the time-dependence of $\rho(\tau)$ is small, then it is fair to say that the

master equation, while still somewhat wide of the mark when compared with the exact dynamics, gives a *better* description of the physics of the problem than when a ghost is present, that is, when $\rho(\tau)$ is not ergodic. Perhaps, it should be added to the list of conditions under which the master equation can be expected to be valid, that the system it describes should have the ergodic property as regards the observables of interest. Such an attitude rules out the possibility of using the master equation to *investigate* ergodic properties, but, as has already been remarked in IV, there are quite enough difficulties in understanding the rigorous status of the equation, and its links to the dynamics of finite systems, that this is no real loss. As a calculational tool for real physical and chemical problems, the master equation remains of the highest importance.

The conceptual problem of nonequilibrium statistical mechanics, namely of how it is that one may reasonably discuss the dynamics of infinite systems, is left largely untouched by the above arguments. But it seems to the authors that the possibility of examining, both analytically and numerically an exact solution to a problem which can have the ergodic property—by the elimination of the ghost—is very hopeful from the point of view of obtaining a better understanding of this matter. The notion of subdynamics[11] has recently been introduced into nonequilibrium theory, and it deals with the possibility of separating from the full dynamics of a large system a set of variables or modes, limited in number, whose time evolution depends only on these quantities themselves, and which determine the thermodynamics of the system. It has been pointed out by Levy[12] in a study of the definition of unstable states in the Lee model, and in a field-theoretic calculation very similar to that of this paper, that by "suitable" choice of initial conditions the nonexponential contribution to the evolution of his model can be made arbitrarily small. A thorough investigation of these matters is now possible within the framework of nonequilibrium theory by means of the model of this paper, and it may be hoped that in this way the application of subdynamics to irreversible systems, and the concepts of the thermodynamic limit and the master equation, may be better understood.

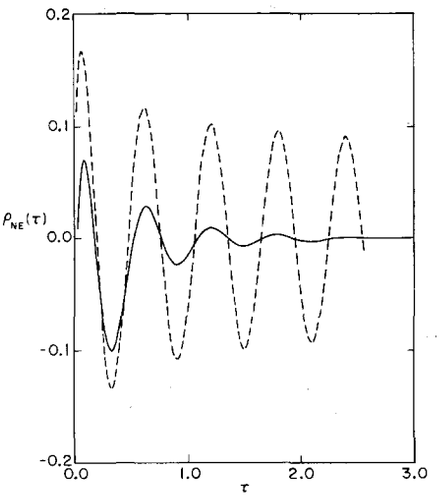


FIG. 6. A plot of the nonexponential contribution to the exact solution (solid line) as a function of τ for the choice $\alpha = 0.1, p = \frac{1}{2}$ (See the discussion following Eq. (34) in Sec. V). The dashed line represents the nonexponential contribution to the Liouville weak-coupling solution derived from Eq. (35) for the same choice of p and α.

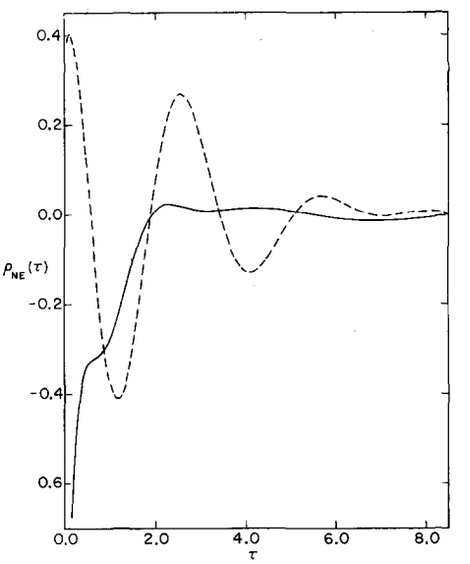


FIG. 7. A plot of the nonexponential contribution to the exact solution (solid line) as a function of τ for the choice $\alpha = 0.8, p = \frac{1}{2}$. The dashed line represents the nonexponential contribution to the Liouville weak-coupling solution derived from Eq. (35) for the same choice of p and α.

ACKNOWLEDGMENTS

One of us (R.D.) thanks the F.R.F.C. of Belgium for financial support while this work was being performed. The other (J.J.K.) has benefited from funds provided to the University of Notre Dame by the National Science Foundation. This work was supported in part by the United States Atomic Energy Commission through contract AT (11-1)-38.

[1] R. Davidson and J. Kozak, J. Math. Phys. **11**, 189 (1970); J. Math. Phys. **11**, 1420 (1970); J. Math. Phys. **12**, 903 (1971). In the text, these papers will be referred to as I, II, and III, respectively.
[2] P. Mazur and Th. J. Siskens, Physica (Utr.) **47**, 245 (1970).
[3] R. I. Cukier and P. Mazur, Physica (Utr.) **53**, 157 (1971).
[4] P. Mazur, Physica (Utr.) **43**, 533 (1969).
[5] B. van der Pol and H. Bremmer, *Operational Calculus* (Cambridge U.P., Cambridge, 1955) 2nd ed., Chap. VII.
[6] R. Balescu, *Statistical Mechanics of Charged Particles* (Interscience, New York, 1963).
[7] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).
[8] S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Harper and Row, New York, 1961).
[9] G. Barton, *Introduction to Advanced Field Theory* (New York, Interscience, 1963).
[10] A. A. Maradudin, E. W. Montroll, and G. H. Weiss, *Theory of Lattice Dynamics in the Harmonic Approximation*, Solid State Physics, supplement 3, edited by F. Seitz and D. Turnbull (Academic, New York, 1963).
[11] C. George, Bull. Acad. Roy. Belg. Classe Sci. **53**, 623 (1967); I. Prigogine, C. George and F. Henin, Physica (Utr.) **45**, 418 (1969); Proc. Natl. Acad. Sci. USA **65**, 789 (1970)
[12] M. M. Levy, Nuovo Cimento **13**, 115 (1959); Nuovo Cimento **14**, 612 (1959).

On the phase velocity and group velocity of guided waves

R. E. Eaves

Division of Engineering, Brown University, Providence, Rhode Island 02912
(Received 25 September 1972)

Relations between the phase velocity and group velocity of guided waves in homogeneous media and the restrictions under which they can be generalized to inhomogeneous media are discussed. Dirichlet, Neumann, and mixed boundary conditions are considered, including the case of frequency dependent media and boundary conditions.

1. INTRODUCTION

In treatments of homogeneous waveguiding structures subject to Dirichlet or Neumann boundary conditions, it is often demonstrated that phase velocity and group velocity are related as follows:

$$v_p v_g = c^2$$

and

$$v_p \geq c \geq v_g,$$

where c is the wave speed of the medium when unbounded.

Case[1] has recently given an interesting generalization for inhomogeneous media. Propagation of the form

$$\Phi = e^{i(kz-\omega t)}\Psi(r_p)$$

is considered, where Φ satisfies the wave equation

$$\nabla^2\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = 0$$

and c is independent of z but a general function of the transverse coordinate r_p. The variational expression

$$\omega^2 = \frac{\int_S |\nabla\Psi|^2 d^2r_p + k^2\int_S |\Psi|^2 d^2r_p}{\int_S (|\Psi|^2/c^2)\, d^2r_p} \tag{1}$$

is the starting point, and it is shown that

$$v_p v_g = \frac{1}{\overline{1/c^2}} \tag{2}$$

and

$$v_p \geq (1/c^2)^{-1/2} \geq v_g, \tag{3}$$

where

$$v_p = \frac{\omega}{k}, \quad v_g = \frac{d\omega}{dk}, \quad \overline{1/c^2} = \frac{\int_S |\Psi|^2 d^2r_p}{\int_S (|\Psi|^2/c^2)\, d^2r_p}.$$

2. DISCUSSION

In clarification of these results, it should be pointed out that (1) is valid for Dirichlet or Neumann boundary conditions,[2] and therefore (2) and (3) hold under that restriction. Implicit in the derivation of (2) and (3) is the assumption that c is not a function of ω.

Actually the cases of equality and inequality in (3) are easily separated. It is seen from (1) that equality is possible only for the solution $\Psi = \text{const}$, a solution which exists if and only if the boundary condition is Neumann and the media is homogeneous.

While (2) is a noteworthy result which may provide insight, one should be cautioned in its use. For inhomogeneous problems Ψ and consequently $\overline{1/c^2}$ are functions of ω. Therefore the product $v_p v_g$ is not constant but a function of ω. Since ω and v_p are mutually dependent, the product $v_p v_g$ is a function of v_p.

Extension of (2) and (3) to mixed boundary conditions is qualified and will be discussed in the following sections. Frequency dependent media and boundary conditions will also be considered.

3. MIXED BOUNDARY CONDITIONS

We denote the contour which is the boundary of the waveguide cross section by Γ and consider the boundary condition

$$\frac{\partial\Psi}{\partial n} + f\Psi = 0, \tag{4}$$

where f may be a function of transverse coordinates but not of frequency. The boundary may extend to infinity, but only modes of finite energy are considered, i.e., $\int_S |\nabla\Psi|^2 d^2r_p$ and $\int_S |\Psi|^2 d^2r_p$ are finite. Then it is simply shown that

$$\omega^2 = \frac{\int_S |\nabla\Psi|^2 d^2r_p + k^2\int_S |\Psi|^2 d^2r_p + \oint_\Gamma f|\Psi|^2 dr_p}{\int_S (|\Psi|^2/c^2)\, d^2r_p}, \tag{5}$$

which is stationary in Ψ. We consider ω to be varied by $\delta\omega$ and describe the corresponding variation in propagation by δk and $\delta\Psi$. Taking the limit $\delta\omega \to 0$ results in relation (2).

If the case $f \equiv 0$ (the previously treated Neumann problem) is excluded, we are assured that $\int_S |\nabla\Psi|^2 d^2r_p > 0$. Then

$$v_p^2 > \frac{1}{\overline{1/c^2}}\left(1 - \frac{\oint_\Gamma f|\Psi|^2 dr_p}{\int_S (|\Psi|^2/c^2)\, d^2r_p}\right)^{-1}$$

follows from (5). Therefore, $f \geq 0$ everywhere on Γ is a sufficient condition for

$$v_p > (\overline{1/c^2})^{-1/2} > v_g.$$

That this result does not hold for general f is demonstrated by the slow-wave ($v_p < c$) structure consisting of a homogeneous half-space bounded by an impedance surface on which $\partial\Psi/\partial n + f\Psi = 0$, with f a negative constant.

4. FREQUENCY DEPENDENT c AND f

If the problem is generalized to frequency dependent c and f, a variation $\delta\omega$ will be accompanied by change in the properties of the waveguiding structure, δc and δf, and by change in the propagation, described by δk and $\delta\Psi$. Effecting such variations in (5) and taking the limit $\delta\omega \to 0$, we get

$$v_p v_g = \frac{1}{\overline{1/c^2}}\left(1 - \omega\,\frac{\int_S (c_\omega/c)(|\Psi|^2/c^2)\,d^2r_p}{\int_S (|\Psi|^2/c^2)\,d^2r_p} - \frac{1}{2\omega}\,\frac{\oint_\Gamma f_\omega |\Psi|^2 dr_p}{\int_S (|\Psi|^2/c^2)\,d^2r_p}\right)^{-1}, \quad (6)$$

where the subscript ω indicates partial differentiation. This form does not invite interpretation as (2) does.

Now (5) is rearranged to give

$$v_p^2 = \frac{1}{\overline{1/c^2}} \times\left(1 - \omega^{-2}\frac{\int_S |\nabla\Psi|^2 d^2r_p}{\int_S (|\Psi|^2/c^2) d^2r_p} - \omega^{-2}\,\frac{\oint_\Gamma f |\Psi|^2 d^2r_p}{\int_S (|\Psi|^2/c^2) d^2r_p}\right)^{-1}. \quad (7)$$

Dividing (7) by (6), we have

$$\frac{v_p}{v_g} = \left(1 - \omega\,\frac{\int_S (c_\omega/c)(|\Psi|^2/c^2)\,d^2r_p}{\int_S (|\Psi|^2/c^2) d^2r_p} - \frac{1}{2\omega}\,\frac{\oint_\Gamma f_\omega |\Psi|^2 dr_p}{\int_S (|\Psi|^2/c^2) d^2r_p}\right) \times\left(1 - \omega^{-2}\frac{\int_S |\nabla\Psi|^2 d^2r_p}{\int_S (|\Psi|^2/c^2) d^2r_p} - \omega^{-2}\,\frac{\oint_\Gamma f |\Psi|^2 dr_p}{\int_S (|\Psi|^2/c^2) d^2r_p}\right)^{-1}. \quad (8)$$

Dirichlet and Neumann boundary conditions are the limit cases $f \to \infty$ and $f \to 0$, with $f_\omega = 0$. Comparison of (1) and (5) reveals that in both cases $\oint_\Gamma f |\Psi|^2 dr_p \to 0$. Therefore results for Dirichlet and Neumann boundary conditions can be obtained by setting $f_\omega = 0$ and $\oint_\Gamma f |\Psi|^2 dr_p = 0$ in (8):

$$\frac{v_p}{v_g} = \left(1 - \omega\,\frac{\int_S (c_\omega/c)(|\Psi|^2/c^2)\,d^2r_p}{\int_S (|\Psi|^2/c^2) d^2r_p}\right) \times\left(1 - \omega^{-2}\,\frac{\int_S |\nabla\Psi|^2 d^2r_p}{\int_S (|\Psi|^2/c^2) d^2r_p}\right)^{-1}$$

We first consider the solution $\Psi = \text{const}$, which exists if and only if the media is homogeneous and the boundary is Neumann. In that case the indirect approach taken here is unnecessary, since all propagation properties are simply determined from $k^2 = \omega^2/c^2(\omega)$.

In all other cases $\int_S |\nabla\Psi|^2 d^2r_p > 0$, so that sufficient conditions for $v_p > v_g$ are

$c_\omega \le 0$ everywhere in S for Dirichlet and Neumann boundary conditions,

$c_\omega \le 0$ everywhere in S, $f \ge 0$, $f_\omega \le 0$ everywhere on Γ for mixed boundary conditions.

5. CONCLUSIONS

For frequency independent waveguiding structures, the product $v_p v_g$ can be given in the suggestive form of (2) if boundary conditions are Dirichlet, Neumann, or mixed. The product is constant for homogeneous media. However, relation (3) applies without qualification only for Dirichlet and Neumann boundary conditions, with equality holding only for homogeneous media, Neumann boundary conditions, and $\Psi = \text{const}$. For the mixed boundary condition $\partial\Psi/\partial n + f\Psi = 0$, (3) holds with the restriction $f \ge 0$.

For frequency dependent waveguiding structures, restrictions on the media and boundary conditions have been given which are sufficient to assure $v_p > v_g$.

[1]K. M. Case, J. Math. Phys. **13**, 360 (1972).
[2]P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953), p. 1112–14.

Hamiltonian orderings and functional Integrals

I. W. Mayes and J. S. Dowker

Department of Theoretical Physics, University of Manchester, Manchester, England
(Received 23 August 1971; revised manuscript received 4 November 1971)

The factor ordering problem is discussed using functional integrals and the nonuniqueness of the quantum Hamiltonian is more closely investigated. Systems whose Hamiltonians are quadratic in the momenta are considered in detail and the equivalence of the canonical and Lagrangian approaches is explicitly confirmed by an averaging procedure. It is important that the integrand be in an averaged form if formal manipulations such as partial integrations are to be performed. This is probably significant for perturbation calculations.

INTRODUCTION

The functional integral formulations of quantum mechanics and quantum field theory often provide conveniently compact expressions suitable for formal manipulations and for the generation of perturbation expansions. It is not necessary here to justify the use of the method and we refer to the books of Katz,[1] Rosen,[2] and Feynman and Hibbs[3] for extensive background material.

Recently, functional integrals have been used in discussions of the ordering problem. Given a classical quantity, say the Hamiltonian, what is the corresponding quantum operator?

Kerner and Sutcliffe[4] and Cohen[5] employ the canonical approach in which the functional integral is one over paths in phase space, i.e., paths $\{p(t), q(t)\}$. According to Cohen[5] the ambiguity in the ordering is due to the different choices one can make for the action A for small time intervals, $t'' - t'$,

$$A[p,q] = \int_{t'}^{t''} (pdq - H_c dt). \tag{1}$$

Typically, A occurs in a functional integral of the form

$$K(q'', t''|q't') = \mathfrak{N} \iint e^{iA[p,q]} d(p)d(q), \tag{2}$$

where, to obtain the usual propagator, we integrate over all paths $\{p(t), q(t)\}$ subject to $q(t') = q'$ and $q(t'') = q''$. There is no restriction on p.

To give a precise meaning to such an integral, it is first of all necessary to turn the functional integrals into lattice ones. Next, the total action A is split up into pieces corresponding to the lattice intervals, and it is here that the ambiguities occur.[5]

In the present work we wish to investigate this question of ambiguities further and also to relate the canonical approach to the strictly Lagrangian one discussed by Rosen[2] and, more generally, by DeWitt.[6] The motivation for this is provided, partly, by the use of functional methods, canonical and Lagrangian, in the theory of chiral dynamics.[7] It seems to us that in such a nonlinear theory one must be more than usually careful when using functional integrals.

QUANTUM HAMILTONIAN

To obtain the quantum Hamiltonian for which (2) is the propagator, we proceed formally and differentiate (2) with respect to t'' in order to obtain a Schrödinger equation.

We have

$$i\frac{\partial}{\partial''} K(q'', t''|q', t') = \mathfrak{N} \iint -\frac{\partial A}{\partial t''} e^{iA[p,q]} d(p)d(q)$$

$$= \mathfrak{N} \iint H_c(t'') e^{iA} d(p)d(q). \tag{3}$$

Using the bracket notation, we can write (3) as

$$i\frac{\partial}{\partial t''} \langle q''t''|q't'\rangle = \langle q''t''|T^*(\underline{H}(t''))|q't'\rangle, \qquad H = H_c(\underline{p},\underline{q}), \tag{4}$$

where T^* is the time ordering operator of Nishijima. Expression (4) is not well defined when the operators involved in $\underline{H}$ are at equal times, and recourse to the lattice definition is required to resolve the ambiguities. The method is that of Katz.[1]

To obtain the derivative of $\int_{t'}^{t''} H_c dt$, we continue it by a time Δt which is then divided into a network of $2m + 1$ lattice points with the p and q arranged alternately and equally spaced. The expression so obtained is next divided by Δt and the limits $m \to \infty$ and $\Delta t \to 0$ taken.

We find

$$\lim_{\substack{m\to\infty \\ \Delta t\to 0}} \sum_{k=0}^{m} \left\{\frac{\epsilon}{\Delta t} H[p(t_{k+1/2}), q(t_{k+1}), q(t_k)]\right\} \tag{5}$$

where $\epsilon = \Delta t/(2m + 1)$ and where $H(p_{k+1/2}, q_{k+1}, q_k)$ is determined only by the requirement that it should give the Hamiltonian $H_c(p_k, q_k)$ in the limit of ϵ tending to zero.[5]

It is expression (5) that should be substituted into (4) for $\underline{H}$, with p, q replaced by operators $\underline{p}, \underline{q}$.

If we choose

$$H(p, q'', q') = \tfrac{1}{2}[H_c(p, q'') + H_c(p, q')],$$

then for (4) we find

$$i\frac{\partial}{\partial t''} \langle q''t''|q't'\rangle = \lim_{\epsilon\to 0} \langle q''t''|T^*\{H_c[p(t''), q(t'' + \epsilon)] + H_c[p(t''), q(t'' - \epsilon)]\}|q't'\rangle$$

$$= \langle q''t''|\underline{H}_s(\underline{p}(t''), \underline{q}(t''))|q't'\rangle,$$

where now the time ordering is well defined and clearly yields the symmetrically ordered Hamiltonian[1,5] $\underline{H}_s$, i.e., the average of the expressions with all the $\underline{p}$ on the left and all the $\underline{p}$ on the right.

In order to obtain the Weyl ordering rule, it is necessary to consider Hamiltonians of the form

$$H_c = q^m p^n.$$

We choose a lowest order term $H(p,q'',q')$ as

$$H(p,q'',q') = H_c(p, \tfrac{1}{2}(q'' + q'))$$

and find that (4) becomes

$$i\frac{\partial}{\partial t''}\langle q''t''|q't'\rangle = \lim_{\epsilon\to 0}\left\langle q''t''\right| T^* \times \left\{\frac{1}{2^m}[\underline{q}(t''+\epsilon) + \underline{q}(t''-\epsilon)]^m \underline{p}^n(t'')\right\}\left|q't'\right\rangle. \quad (6)$$

The time ordering is performed after expanding the power by the binomial theorem,

$$[\underline{q}(t+\epsilon) + \underline{q}(t-\epsilon)]^m = \sum_{r=0}^{m} C_r^m \underline{q}^r(t+\epsilon)\underline{q}^{m-r}(t-\epsilon). \quad (7)$$

[The fact that $q(t+\epsilon)$ and $q(t-\epsilon)$ do not commute does not worry us here as the commutator will be of order ϵ and will vanish in the limit.] Substituting (7) into (6), we find the quantum Hamiltonian to be

$$2^{-m}\sum_{0}^{m} C_r^m \underline{q}^r \underline{p}^n \underline{q}^{m-r},$$

which is equivalent to the Weyl ordering.[8]

Instead of determining the quantum Hamiltonian through a Schrödinger equation, one can proceed directly to the Heisenberg equations of motion. These follow from the functional integration by parts lemma.[2]

Thus we have

$$\iint \frac{\delta}{\delta p} e^{iA} d(p)\, d(q) = 0$$

and (8)

$$\iint \frac{\delta}{\delta q} e^{iA} d(p)\, d(q) = 0$$

whence, formally, we find

$$\dot{\underline{q}} = T^*\left\{\frac{\partial H_c(\underline{p},\underline{q})}{\partial \underline{p}}\right\}$$

and (9)

$$\dot{\underline{p}} = T^*\left\{\frac{\partial H_c(\underline{p},\underline{q})}{\partial \underline{q}}\right\}.$$

As before we must return to the lattice form of the integral to define the time ordering, and it is clear from our previous results that we shall find the Heisenberg equations of motion,

$$\dot{\underline{q}} = \frac{\partial \underline{H}}{\partial \underline{p}} = i[\underline{H},\underline{q}]$$

and

$$\dot{\underline{p}} = \frac{\partial \underline{H}}{\partial \underline{q}} = i[\underline{H},\underline{p}],$$

to be in accord with the Hamiltonian in the Schrödinger picture.

To summarize the main points, we can say that the formal expression (2) will give different propagators depending on what lowest order term one chooses for the action,

$$\lim_{\epsilon\to 0}\int_{t-\epsilon}^{t+\epsilon}(p\,dq - H_c\,dt) = \Big(p(t)\frac{1}{2\epsilon}[q(t+\epsilon) - q(t-\epsilon)] - H(p(t), q(t+\epsilon), q(t-\epsilon))\Big)2\epsilon. \quad (10)$$

Expression (10) is taken to be valid inside the functional integral, as part of the implicit formalism, and this is important because it enables us perform the functional integration by parts and pass from (8) to (9). For this to be valid, we require that

$$\frac{dA}{dt} = p\dot{q} - H_c$$

under the functional integral sign.

CANONICAL VERSUS LAGRANGIAN

If we wish to compare and ultimately show the equivalence of the canonical and Lagrangian approaches, it will be wise to discuss systems for which they are not obviously different at the start.[2,8] Thus we shall restrict ourselves to systems the Hamiltonians for which are quadratic in the momenta. This class of systems is sufficiently large to make the discussion relevant to real life. The quantization of such systems has been discussed at length by DeWitt,[6] and we shall use the results of his work.

We write down the classical Lagrangian and Hamiltonian

$$L(q,\dot{q}) = \tfrac{1}{2} g_{\alpha\beta}(q)\dot{q}^\alpha\dot{q}^\beta, \quad \dot{q}^\alpha = \frac{dq^\alpha}{dt},$$
$$H(p,q) = \tfrac{1}{2} g^{\alpha\beta}(q)p_\alpha p_\beta, \quad p_\alpha = g_{\alpha\beta}\dot{q}^\beta. \quad (11)$$

Now, instead of arguing from the classical to the quantum Hamiltonian, it is easier to reverse this and take the quantum equations (Schrödinger or Heisenberg) as given and then ask for the corresponding functional integral. Again we concentrate on the propagator $K(q'',t''|q',t')$, which we will assume satisfies the covariant equation

$$\left(i\frac{\partial}{\partial t''} + \tfrac{1}{2}\Delta_2''\right)K(q'',t''|q',t') = i\delta(t''-t')\delta(q',q''),$$

Δ_2 being the Laplace–Beltrami operator for the metric $g_{\alpha\beta}$. The quantum Hamiltonian is thus fixed and reads[6,9]

$$\underline{H} = \tfrac{1}{2}\underline{g}^{-1/4}\underline{p}_\alpha\underline{g}^{\alpha\beta}\underline{g}^{1/2}\underline{p}_\beta\underline{g}^{-1/4} \quad (12)$$

in terms of the momentum operator p_α. Passing to the Heisenberg picture the operator equations of motion follow directly if the canonical commutation rules

$$[\underline{q}^\alpha,\underline{q}^\beta] = 0 = [\underline{p}_\alpha,\underline{p}_\beta], \quad [\underline{p}_\alpha,\underline{q}^\beta] = -i\delta_\alpha^\beta \quad (13)$$

are employed. However, since the resulting, rather long, expressions have already been given by DeWitt[10] they will not be repeated at this point.

We now ask whether it is possible to write K in the form (2) with A given by (1) and H_c by (11). Under the assumption (10) with $H(p,q'',q')$ chosen to give the symmetrical ordering, say, it is easily checked that the quantum Hamiltonian is not that of (12), and this is true whatever ordering we choose.

To make the quantum Hamiltonian coming from (2) the same as the $\underline{H}$ of (12), one must take H_c in (1) not as the classical Hamiltonian H of (11) but as $H + B(q)$, where the B term is to be determined and will depend on the particular correspondence rule opted for. For the symmetrization rule we find

$$H_c(p,q) = H(p,q) + \tfrac{1}{4}\partial_\alpha\partial_\beta g^{\alpha\beta} - \tfrac{1}{2}g^{1/4}\Delta_2 g^{-1/4}. \quad (14)$$

This is the expression given in our earlier paper[11] where the ambiguities inherent in the calculation were not appreciated.

In general, then, we have the canonical expression

$$K(q''t''|q't') = \mathfrak{N}\iint \exp\left(i\int_{t'}^{t''}(p_\alpha\dot{q}^\alpha - H - B)dt\right)d(p)d(q) \quad (15)$$

with the understanding that in the lattice meaning of the integrals the action is approximated to first order in the lattice spacing ϵ according to (10), and the B term will depend on the particular $H(p,q'',q')$ chosen.[12]

Thus we have different-looking, formal integrals for the same propagator.

The origin of this situation can be discovered by actually constructing the functional integral for K as DeWitt[6] does. He, however, works directly with the Lagrangian method, and so let us formally integrate over p to obtain

$$K(q''t'|q't') = \mathfrak{N}'\int \exp\left(i\int_{t'}^{t''}(L-B)dt\right)\mathfrak{D}(q), \quad (16)$$

where

$$\mathfrak{D}(q) = \exp[\tfrac{1}{2}\delta(0)\int \ln g\,dt]\,d(q).$$

By folding together the short time propagators DeWitt finds for K the formal expression

$$K(q''t''|q't') = \mathfrak{N}'\int\exp\left(iS(q''t''|q't') + i\tfrac{1}{6}\int_{t'}^{t''}R(q)\,dt\right)\mathfrak{D}(q), \quad (17)$$

where

$$S(q''t''|q't') = \int_{t'}^{t''}L(q)\,dt$$

and where R is the scalar curvature for the metric $g_{\alpha\beta}$.

The fact that (17) differs from (16) is not surprising because, as with (2), one has to say what one means by the functional integral, i.e., one has to say how one treats the integrand in a lattice approximation. For (16) this is already defined, through (10), i.e., we choose the lowest order term for the action $S = \int_{t'}^{t''}L dt$. This is not true for (17), where one has to expand S in powers of $(q''-q')$ as far as necessary (usually to the fourth order). The paths that contribute to the functional integral are really Brownian paths for which $\delta q \sim \epsilon^{1/2}$ and so $(\delta q)^4/\epsilon \sim \epsilon$.[13]

It will be demonstrated that expressions (16) and (17) are, in fact, equivalent, the quantity B arising from averaging terms of higher order in $(q''-q')$ in the expansion of $S(q''t''|q't')$ in (17), as suggested before.[11]

The first step is to write (15) in lattice form, integrate over p and compare with the lattice form of (17). For definiteness we shall choose $H(p,q'',q')$ to give the symmetrical ordering, viz.,

$$H(p,q'',q') = \tfrac{1}{2}[H_c(p,q'') + H_c(p,q')].$$

The action then becomes

$$\sum_{k=0}^{n}\Big\{p_{\alpha,k+1/2}(q^\alpha_{k+1}-q^\alpha_k) - \epsilon\mathcal{G}^{\alpha\beta}(q_{k+1},q_k) \times p_{\alpha,k+1/2}p_{\alpha,k+1/2} - \tfrac{1}{2}\epsilon[B(q_k)+B(q_{k+1})]\Big\} = \epsilon A_\epsilon, \quad (18)$$

where

$$\mathcal{G}^{\alpha\beta}(q_{k+1},q_k) = \tfrac{1}{4}[g^{\alpha\beta}(q_{k+1}) + g^{\alpha\beta}(q_k)]$$

and where the index k refers to the lattice point, i.e., $q_{k+1} = q(t_{k+1})$, etc. For the propagator we have

$$K(q''t''|q't') = \mathfrak{N}\,g'^{-1/4}g''^{-1/4}\iint e^{i\epsilon A_\epsilon}d(p)d(q) \quad (19)$$

where

$$\mathfrak{N} = (2\pi)^{-(n+1)r},$$

and

$$d(p) = \prod_{k=0}^{n}\prod_{\alpha=1}^{r}dp_{\alpha,k+1/2}, \qquad d(q) = \prod_{k=1}^{n}\prod_{\alpha=1}^{r}dq^\alpha_k.$$

r is the dimension of the space involved, $\alpha = 1,2,\ldots,r$.

The integration over p can now be performed by the standard method of completing the square and translating the integration variables. We find

$$g'^{1/4}g''^{1/4}K(q''t''|q't') = \mathfrak{N}(\pi i\epsilon)^{-(n+1)r/2}\int_{q'}^{q''}d(q)\prod_k[\det\mathcal{G}^{\alpha\beta}(q_{k+1},q_k)]^{-1/2} \times\exp\Big(i\sum_k\{(1/4\epsilon)(q^\delta_{k+1}-q^\delta_k)(q^\gamma_{k+1}-q^\gamma_k) \times\mathcal{G}_{\delta\gamma}(q_{k+1},q_k) - \tfrac{1}{2}\epsilon[B(q_k)+B(q_{k+1})]\}\Big) \quad (20)$$

where $\mathcal{G}_{\alpha\beta}$ is the matrix inverse of $\mathcal{G}^{\alpha\beta}$.

We compare (20) with the exact expression derived by DeWitt.[6] This is

$$\mathfrak{N}(2\pi\epsilon i)^{-(n+1)r/2}\int_{q'}^{q''}d(q)\prod g^{1/4}(q_{k+1})g^{1/4}(q_k) \times\exp\Big(i\sum_k[S(q_{k+1},t_{k+1}|q_k,t_k) + \tfrac{1}{6}\epsilon R(q_k)]\Big) \quad (21)$$

with the expansion of the action[6]

$$\begin{aligned}\epsilon S(q_{k+1},t_{k+1}|q_k,t_k) &= \tfrac{1}{2}g_{\alpha\beta}(q_k)(q^\alpha_{k+1}-q^\alpha_k)(q^\beta_{k+1}-q^\beta_k)\\ &+ \tfrac{1}{12}P_3 g_{\alpha\beta,\gamma}(q_k)(q^\alpha_{k+1}-q^\alpha_k)(q^\beta_{k+1}-q^\beta_k)(q^\gamma_{k+1}-q^\gamma_k)\\ &+ \tfrac{1}{72}[P_6 g_{\alpha\beta,\gamma\delta}(q_k) - P_3 g^{\epsilon\lambda}(q_k)[\alpha\beta,\epsilon]_k[\gamma\delta,\lambda]_k]\\ &\times(q^\alpha_{k+1}-q^\alpha_k)(q^\beta_{k+1}-q^\beta_k)(q^\gamma_{k+1}-q^\gamma_k)(q^\delta_{k+1}-q^\delta_k).\end{aligned} \quad (22)$$

Here P_3 and P_6 stand for summations over the indicated number of distinct terms obtained by permuting the free indices on the object following.

If we use the symbol $\doteq$ to denote equivalence under the multiple integral sign, then we wish to show that the following relationship is true

$$\begin{aligned}2^{-r/2}[\det\mathcal{G}^{\alpha\beta}(q_{k+1},q_k)]^{-1/2}\exp\{i[(1/4\epsilon)\mathcal{G}_{\delta\gamma}(q_{k+1},q_k) &\times(q^\delta_{k+1}-q^\delta_k)(q^\gamma_{k+1}-q^\gamma_k) - \epsilon B(q_k)]\}\\ \doteq g^{1/4}(q_{k+1})g^{1/4}(q_k)\exp\{i[S(q_{k+1},t_{k+1}|q_k,t_k)&\\ + \tfrac{1}{6}\epsilon R(q_k)]\},&\end{aligned} \quad (23)$$

where we have replaced $\frac{1}{2}[B(q_{k+1}) + B(q_k)]$ by $B(q_k)$, the difference being of higher order in ϵ.

This equivalence is shown by an averaging procedure. We are going to extract a factor

$$\exp(i/2\epsilon)g_{\alpha\beta}(q_k)(q^\alpha_{k+1} - q^\alpha_k)(q^\beta_{k+1} - q^\beta_k) \tag{24}$$

and then average the remaining terms against this using the easily derived theorems,

$$(2\pi i\epsilon)^{-r/2} \int dq'' f(q'')(q''^\alpha - q'^\alpha)(q''^\beta - q'^\beta) \times \exp(i/\epsilon)g'_{\gamma\delta}(q''^\gamma - q'^\gamma)(q''^\delta - q'^\delta) = i\epsilon g'^{\alpha\beta} f(q') + O(\epsilon^{3/2}) \tag{25}$$

and

$$(2\pi i\epsilon)^{-r/2} \int dq''(q''^\alpha - q'^\alpha)(q''^\beta - q'^\beta)(q''^\gamma - q'^\gamma) \times(q''^\delta - q'^\delta) f(q'') \exp(i/\epsilon)g'_{\epsilon\lambda}(q''^\epsilon - q'^\epsilon)(q''^\lambda - q'^\lambda) = -\epsilon^2 f(q')[g'^{\alpha\beta}g'^{\gamma\delta} + g'^{\alpha\gamma}g'^{\beta\delta} + g'^{\alpha\delta}g'^{\beta\gamma}] + O(\epsilon^{5/2}), \tag{26}$$

the first of which was given by DeWitt.[6] The averages of higher powers of $(q'' - q')$ can also be calculated but will not be needed since they contribute at least to order ϵ^3.

Before applying these theorems an expansion of $[\det\mathcal{G}^{\alpha\beta}(q_{k+1}, q_k)]^{-1/2}$ in the form $g^{1/4}(q_{k+1})g^{1/4}(q_k)[1 + O(q_{k+1} - q_k)^2]$ is needed. Higher order terms are unnecessary because they contribute to higher order in ϵ and will tend to zero as ϵ vanishes. We also require an expansion of $\mathcal{G}_{\alpha\beta}$.

A little algebra reveals that, correct to second order,

$$\tfrac{1}{4}\mathcal{G}_{\alpha\beta}(q_{k+1}, q_k) = \tfrac{1}{2}g_{\alpha\beta}(q_k) + \tfrac{1}{4}g_{\alpha\beta,\gamma}(q_k)(q^\gamma_{k+1} - q^\gamma_k) - \tfrac{1}{8}[g_{\alpha\lambda}(q_k)g_{\gamma\beta}(q_k)g^{\lambda\gamma}{}_{,\delta\epsilon}(q_k) + g_{\alpha\lambda}(q_k)g_{\gamma\beta,\delta}(q_k)g^{\lambda\gamma}{}_{,\epsilon}(q_k)](q^\delta_{k+1} - q^\delta_k)(q^\epsilon_{k+1} - q^\epsilon_k). \tag{27}$$

Use of the determinant theorem

$$\det(A + B) = \det A \det(1 + A^{-1}B) = \det A\{1 + \operatorname{tr}(A^{-1}B) + \tfrac{1}{2}[\operatorname{tr}(A^{-1}B)]^2 - \tfrac{1}{2}\operatorname{tr}(A^{-1}B)^2 + \cdots\}$$

then enables one to show that $[\det\mathcal{G}^{\alpha\beta}(q_{k+1}, q_k)]^{-1/2}$ is equal to the following, to second order:

$$2^{r/2}g^{1/4}(q_{k+1})g^{1/4}(q_k)[1 + \tfrac{1}{16}g_{\alpha\beta,\gamma}(q_k)g^{\alpha\beta}{}_{,\delta}(q_k) \times (q^\gamma_{k+1} - q^\gamma_k)(q^\delta_{k+1} - q^\delta_k)]. \tag{28}$$

The calculation now proceeds by substituting (27) and (28) into the basic requirement (23), extracting the factor (24), and using the averaging equations (25) and (26) after the remaining exponentials have been expanded in power series up to fourth order in $(q_{k+1} - q_k)$.

The algebra is rather tedious, and we therefore simply state the result that

$$B(q) = Q(q) + \tfrac{1}{4}g^{\alpha\beta}{}_{,\alpha\beta}(q),$$

where

$$Q(q) = -\tfrac{1}{2}g^{1/4}\Delta_2 g^{-1/4}.$$

Q is the Q term of DeWitt.[10]

A glance at equation (14) shows that indeed we have shown the equivalence of the canonical and Lagrangian expressions (15), or (16), and (17), respectively.

It will be obvious from the preceding calculation that the form of B will depend on the choice of the lowest order "Hamiltonian," $H(p, q'', q')$. This is as it should be if we are to have unique equations of motion since the choice of $H(p, q'', q')$ decided the factor ordering in the quantum Hamiltonian as previously demonstrated. Differences in the Hamiltonian due to different factor orderings are compensated by different B terms in such a way that the total Hamiltonian remains unchanged.

For completeness we now give the equations of motion that come from an application of the functional integration by parts lemma [cf. Eqs. (8) and (9)]. They are

$$\dot{\underline{p}}_\alpha + \tfrac{1}{2}\underline{p}_\beta \underline{g}^{\beta\gamma}{}_{,\alpha}\underline{p}_\gamma + \underline{Q}_{,\alpha} = 0$$

and

$$\dot{\underline{q}}^\alpha = \tfrac{1}{2}(\underline{g}^{\alpha\beta}\underline{p}_\beta + \underline{p}_\beta \underline{g}^{\alpha\beta})$$

in agreement with the equations of DeWitt.[10] The canonical commutation relations (13) can also be derived by the same lemma. The method is a generalization of that given by Rosen[2] and will not be detailed here.

We emphasize that these manipulations are only possible because $dA = (p\dot{q} - H_c)dt$ is valid inside the functional integrals.

LAGRANGIAN METHOD

In most field theoretic discussions it is from the Lagrangian formulation that perturbation theory is developed, and it is therefore natural to enquire whether one needs the canonical approach.

We can begin by writing down the "naive" functional integral

$$K(q''t'' \mid q't') = \mathcal{N} \int \mathcal{D}(q) \exp\left(i \int_{t'}^{t''} L\,dt\right), \tag{29}$$

L being given by (11), and asking what it means. To answer this question, we must split the functional integral into a lattice one over variables $q^\alpha(t_k) = q^\alpha_k$, $k = 0, 1, \ldots, n$, with $q^\alpha_{k+1} - q^\alpha_k = \epsilon$ and let ϵ tend to zero, and n to infinity, as usual.

Exactly as in the canonical case the action

$$S(q''t'' \mid q't') = \int_{t'}^{t''} L(q, \dot{q})\,dt$$

is broken into pieces corresponding to the lattice intervals and, again, it is here that apparent ambiguities arise.

We assume that, for small ϵ, the action is approximated by

$$\lim_{\epsilon\to 0} S = \sum_{k=1}^{n} L_\epsilon(q_k, q_{k-1})\epsilon,$$

where

$$\lim_{\epsilon\to 0} L_\epsilon(q_k, q_{k+1}) = L(q, \dot{q})$$

is the only restriction on L_ϵ. Clearly, different approximations will effectively describe different quantum systems, in general. Thus, on this basis, the propagator for a given quantum system will have various functional integral representations depending on the choice of L_ϵ. As before, let us consider the system defined by the quantum Hamiltonian $\underline{H}$ of (12) and compare DeWitt's expression (17) with (29).

The equations of motion are covariant under point

transformations of the q^α and this shows up in the covariance of expression (17). Now, expression (29) looks covariant at first sight but is not so in fact. The reason is that (29) is a stochastic integral with terms of order $(q_k - q_{k-1})^4/\epsilon$ contributing to order ϵ. Thus the lowest order truncation of $S(q_k, t|q_{k-1}, t-\epsilon)$ to $L_\epsilon(q_k, q_{k-1})\epsilon$, which is implicit in (29), is *not* a covariant procedure. In order that (29) should be equivalent to (17), it is necessary therefore to add a term $-B(q)$ to the Lagrangian $L(q,\dot{q})$ to allow for this truncation and to restore covariance.

We thus arrive at the form (16) previously derived from the canonical method.

The B quantity can be calculated by equating (16) and (17) and averaging over the higher order terms in the expansion of S, as was done in the canonical method. Its value can then be checked by showing that the (Lagrangian) equations of motion correspond to those written down at the end of the last section. Of course, we gain nothing over the canonical method, but that is not our object.

The use of (16) with the truncated expression for S and the inclusion of the compensatory term B is very convenient since one has $(d/dt)\int L\,dt = L$ inside the functional integral, whereas in (17) such a statement is untrue due to the stochastic nature of the integration.[13] However, this is just the condition that must be satisfied if we are to apply the functional integration by parts lemma straightforwardly.

Obviously the form of B will depend on the choice of L_ϵ which will determine the operator ordering of q and $\dot{q}$ in $\underline{L}(\underline{q},\underline{\dot{q}})$, i.e., in the equations of motion. The particular L_ϵ that gives the simplest (symmetrical) ordering of $\underline{q}$ and $\underline{\dot{q}}$, i.e.,

$$T^*(\underline{\dot{q}}^2 f(\underline{q})) = \tfrac{1}{2}[\underline{\dot{q}}^2 f(\underline{q}) + f(\underline{q})\underline{\dot{q}}^2] + \{\text{terms arising from ordering } \dot{q}^2\},$$

is

$$L_\epsilon[q(t+\epsilon), q(t)] = (1/4\epsilon^2)[q^\alpha(t+\epsilon) - q^\alpha(t)] \times [q^\beta(t+\epsilon) - q^\beta(\epsilon)]\{g_{\alpha\beta}[q(t+\epsilon)] + g_{\alpha\beta}[q(t)]\},$$

and we shall determine B and the equations of motion appropriate to this choice. We should point out that the ordering here is not the same as the symmetrical ordering of p and q. However, for practical purposes, we might need the simplest ordering of fields and their derivatives, in quantum field theory, for example.

The evaluation of B and the equations of motion is rather complicated and we can only sketch the method here. The details are in the Manchester thesis of one of us (I.W.M.).

In contrast to the canonical method more care is needed regarding the normalization. One has to ensure that the unitarity requirement is satisfied.[14] If one does this, the basic condition for equivalence of (16) and (17) is

$$D^{1/2}(q_{k+1}, q_k)\exp\{i\{(1/4\epsilon)(q^\alpha_{k+1} - q^\alpha_k)(q^\beta_{k+1} - q^\beta_k) \times [g_{\alpha\beta}(q_{k+1}) + g_{\alpha\beta}(q_k)] - \epsilon B(q_k)\}\) \doteq \epsilon^{n/2} g^{1/4}(q_{k+1}) g^{1/4}(q_k) \times \exp\{i[S(q_{k+1}, t+\epsilon|q_k, t) + \tfrac{1}{6}\epsilon R(q_k)]\}, \quad (30)$$

where D is, basically, the van Vleck determinant, defined by

$$D(q_{k+1}, q_k) = \det D_{\alpha\beta},$$

$$D_{\alpha\beta} = -\frac{\partial^2}{\partial q^\alpha_{k+1}\partial q^\beta_k}[L_\epsilon(q_{k+1}, q_k) - B(q_k)]\epsilon.$$

Exactly as before we need an expansion of $D^{1/2}$ to second order. This is given in DeWitt.[6] We now substitute this series into (30), expand S up to fourth order in $(q_{k+1} - q_k)$ and average terms of order higher than $(q_{k+1} - q_k)^2$ using Eqs. (25) and (26). A little heavy algebra then yields the result

$$B = Q + \tfrac{1}{4} g^{\alpha\beta}{}_{,\alpha\beta} - \tfrac{1}{8}(g_{\alpha\beta} g^{\beta\gamma}{}_{,\gamma} g^{\alpha\epsilon}{}_{,\epsilon} + g_{\alpha\beta} g^{\alpha\gamma}{}_{,\epsilon} g^{\epsilon\beta}{}_{,\gamma} + g_{\gamma\alpha,\epsilon} g_{\delta\beta,\lambda} g^{\gamma\lambda} g^{\delta\epsilon} - g_{\gamma\alpha,\epsilon} g_{\beta\lambda,\delta} g^{\gamma\lambda} g^{\delta\epsilon}) + \tfrac{1}{4} g^{\gamma\delta} g_{\alpha\gamma,\delta\beta}$$

with Q as before.

Thus we now have the propagator in the form

$$K(q''t''|q't') = \lim \int \prod_{k=1}^{n-1} dq_k \prod_{k=0}^{n} D^{1/2}(q_{k+1}, q_k) \times \exp\left(i \sum_{k=0}^{n} [\epsilon L_\epsilon(q_{k+1}, q_k) - \epsilon B]\right), \quad (31)$$

where everything is known.

The equations of motion are determined by means of an integration by parts, which can be written formally and then put onto a lattice using (31) in order to evaluate the time ordered products that arise. The calculation is again a little involved and one has to be very careful to keep all those terms which might contribute in the limit of ϵ tending to zero. Not only this but it seems that one needs the commutator function $\underline{G}^{\alpha\beta}(\underline{q}_k, \underline{q}_{k-1})$ to second order in ϵ. $\underline{G}^{\alpha\beta}$ is defined by

$$[\underline{q}^\alpha_{k+1}, \underline{q}^\beta_k] = i\underline{G}^{\alpha\beta}(\underline{q}_{k+1}, \underline{q}_k).$$

Thus one is lead to a somewhat circular situation. Of course, it is always possible to use the evolution equation

$$\underline{q}^\alpha_{k+1} = e^{i\underline{H}\epsilon}\underline{q}^\alpha_k e^{-i\underline{H}\epsilon},$$

which can be obtained from the canonical method, and if this is done, one obtains, after a further heavy calculation, the equations of motion in the form

$$\tfrac{1}{2}\frac{d}{dt}(\underline{\dot{q}}^\beta \underline{g}_{\alpha\beta} + \underline{g}_{\alpha\beta}\underline{\dot{q}}^\beta) = \tfrac{1}{2}\underline{\dot{q}}^\beta \underline{g}_{\gamma\beta,\alpha}\underline{\dot{q}}^\gamma - \underline{Q}_{,\alpha} + \underline{W}_\alpha$$

with

$$\underline{W}_\alpha = \tfrac{1}{4}[\underline{g}^{\gamma\beta}(\underline{g}^{\delta\epsilon}{}_{,\epsilon}\underline{g}_{\delta\beta,\alpha})_{,\gamma} + \tfrac{1}{2}\underline{g}^{\delta\epsilon}{}_{,\epsilon}\underline{g}^{\gamma\beta}\underline{g}_{\delta\gamma,\alpha}]$$

which is just what DeWitt[9] has given. These equations are identical to the ones at the end of the last section.

DISCUSSION AND CONCLUSION

In reality the result of the previous section is no more than a check that we have done our algebra correctly; however, it does show, firstly, that the Lagrangian method is more messy than the canonical one. This is partly because limits of the form

$$\lim_{\substack{q''\to q' \\ \epsilon\to 0}} \frac{f(q''-q')}{\epsilon}$$

do not occur in the canonical calculation and partly because, whereas in the canonical method $T^*(\underline{p}_\alpha(t)\underline{p}_\beta(t))$

$= \underline{p}_\alpha(t)\underline{p}_\beta(t)$, it is not true that the equivalent expression arising in the Lagrangian formalism, $\lim_{\epsilon\to 0} \epsilon^{-2} T^*[(\underline{q}^\alpha(t+\epsilon) - \underline{q}^\alpha(t))(\underline{q}^\beta(t+\epsilon) - \underline{q}^\beta(t))]$, is equal to $\underline{\dot{q}}^\alpha\underline{\dot{q}}^\beta$.

In other words, there are problems ordering the $\dot{q}$.

Further, it appears necessary to know the commutator $G^{\alpha\beta}$, to second order in ϵ, in order to obtain the equations of motion. This itself is a partial solution to the problem and it appears that we must know this solution in order to find the complete one. An appeal to the canonical method is therefore called for at this point, which is rather annoying.

The nonuniqueness of the operator orderings has been treated using the ideas of Cohen[5] who showed that the uniqueness result of Kerner and Sutcliffe[4] was incorrect. We have seen that it leads to various expressions for the effective potential B, and it is possible that suitable choices of ordering and of coordinate system could simplify considerably any calculations. For example, if the Riemannian space is the manifold of a semisimple group, as in chiral theory, Q is a constant in canonical coordinates.

Instead of arguing from the classical to the quantum Hamiltonian we have done the reverse. We think that this is both more practical and more fundamental as the world is really a quantum mechanical one.

Finally we should like to emphasise that for calculational purposes it is necessary, in many cases, to write the functional integrals in the averaged forms (15) and (16) and it seems to us that this could be significant in any renormalization of chiral loops, for example.

[1] A. Katz, *Classical Mechanics, Quantum Mechanics and Field Theory* (Academic, New York, 1969).

[2] G. Rosen, *Formulations of Classical and Quantum Dynamical Theory* (Academic, New York, 1969).

[3] R. P. Feynman and A. R. Hibbs, *Quantum mechanics and path integrals* (McGraw-Hill, New York, 1965).

[4] E. Kerner and W. Sutcliffe, J. Math. Phys. **11,** 391 (1970).

[5] L. Cohen, **1,** 3296 (1970); F. J. Testa, J. Math. Phys. **12,** 1471 (1971).

[6] B. S. DeWitt, Rev. Mod. Phys. **29,** 377 (1957).

[7] J. Honerkamp and K. Meetz, Phys. Rev. D 3, 1996, (1970); A. Salam and J. Strathdee, Phys. Rev. D 2, 2869 (1970). D. Boulware, Ann. Phys. (N.Y.) 56, 140 (1968), J. M. Charap, Phys. Rev. D3, 1998 (1971).

[8] H. J. Groenewold, K. Dan. Vidensk. Selsk. Mat.-Fys. Medd. **30** (19), (1956).

[9] B. Podolsky, Phys. Rev. **32,** 812 (1928).

[10] B. S. DeWitt, Phys. Rev. **85,** 653 (1952).

[11] J. S. Dowker and I. W. Mayes, Nucl. Phys. B **29,** 259 (1971).

[12] The effect of this nonuniqueness can be discussed for more general orderings than the ones considered here. Then B in (15) may contain a term proportional to p, $B = ap_a\partial_\beta g^{\alpha\beta} + b\partial_a\partial_\beta g^{\alpha\beta} + Q$. Equation (16) must then be altered by replacing B by $ag_{\alpha\beta}q^\alpha\partial_\gamma g^{\gamma\beta} + b\partial_a\partial_\beta g^{\alpha\beta} - (1/2)a^2 g_{\alpha\beta}\partial_\gamma g^{\gamma\alpha}\partial_\delta g^{\delta\beta} + Q$. The numbers a and b are related to the derivatives of the function defining the general ordering. Our paper "Canonical Functional Integrals in General Coordinates," submitted to Proc. Roy. Soc. Lond., can be consulted for the details.

[13] L. S. Schulman, Phys. Rev. **176,** 1558 (1968); S. F. Edwards and Y. V. Gulyaev, Proc. R. Soc. A **279,** 229 (1964); R. P. Feynman, Rev. Mod. Phys. **20,** 267 (1948).

[14] Cecille Morette, Phys. Rev. **81,** 848 (1951).

On the Wigner coefficients of $Sp(4)$ and $SO(5)$

Wayne J. Holman III*
Department of Physics and Astrophysics, University of Colorado, Boulder, Colorado 80302
(Received 21 August 1972)

The matrix elements of totally symmetric tensor operators in $Sp(4)$, the spinor covering group of $SO(5)$, are obtained by direct factorization of the corresponding $Sp(4) * Sp(4)$ matrix element. The general $Sp(4) * Sp(4)$ matrix element is factorized, and coefficients are obtained which possess a degree of symmetry higher than that of the canonical Wigner coefficients of $Sp(4)$, but which do not form an orthonormal set in the multiplicity space. The canonical Wigner coefficients can, in principle, be derived from those obtained in different ways by a Schmidt orthogonalization process.

1. INTRODUCTION

In recent years the Wigner coefficients of $Sp(4)$ have come under investigation as a result of the applicability of the irreducible representations of this group to the classification of nuclear states. Hecht[1,2] has calculated special cases of these coefficients, and Wong[3] has also calculated matrix elements of the tensor operator $\langle 1\ 0\rangle$. Ališauskas and Jucys[4,5] have studied the general cases of multiplicity-free tensor operators; but in general expressions for the matrix elements of tensor operators which are not multiplicity-free have so far not been available.

A method exists, however, for the extraction of such matrix elements by factorization of the $Sp(4)*Sp(4)$ matrix elements. This procedure has been suggested by Biedenharn, Giovannini, and Louck[6,7] for the case of the group $U(n)$, and all $U(n)*U(n)$ boson states, as well as those of $Sp(4)*Sp(4)$ have been constructed.[8] General factorization of matrix elements in the orbital group, however, is a very complicated process. No nontrivial case of such a factorization has yet been accomplished for any of the unitary groups, though the trivial (multiplicity-free) case of the $U(2)*U(2)$ factorization is well known. The present work achieves the factorization of the $Sp(4)*Sp(4)$ matrix element of a general boson state into Wigner coefficients of $Sp(4)$. The construction is described in detail in the hope that it may have some paradigmatic value for more complicated factorization processes. The factorization, however, realizes canonical matrix elements of tensor operators in $Sp(4)$ only in degenerate cases, i.e., only in cases that are multiplicity-free or that one of the three representations being coupled is totally symmetric. In the most general case, we obtain Wigner coefficients which have redundant multiplicity labels. The Wigner coefficients obtained by our procedure, then, must be subjected to a Schmidt orthogonalization process in order to yield an orthonormal set in the multiplicity space.

The procedure followed here is based on identities for the general and singly stretched $(9-j)$ symbols found in a previous paper.[9] It is to be hoped that the derivation of these identities will indicate how the "$(9-j)$ symbols" in $U(n-1)$ which occur in the general $U(n)*U(n)$ matrix element are to be treated in order to allow the factorization of this matrix element into Wigner coefficients of $U(n)$.

In addition to the applications of the $Sp(4)$ Wigner coefficient in nuclear physics, it is also useful in the theory of generalized harmonic analysis of scattering amplitudes. This coefficient, when analytically continued in the three parameters Φ of the representations being coupled (in Hecht's notation $\Phi = J_m + \Lambda_m + 1$), contains all Wigner coefficients of the Poincaré group as asymptotic forms. It thus provides an "analytic medium" within which the different Wigner coefficients of the Poincaré group may be reached from one another by analytic continuation and asymptotic expansion.

In Sec. 2 we consider the matrix elements of totally symmetric tensor operators and the general case in Sec. 3. We use the notation of Refs. 1 and 8 for the state labels throughout.

2. MATRIX ELEMENTS OF TOTALLY SYMMETRIC TENSOR OPERATORS IN $Sp(4)$

Our procedure will be to evaluate the matrix element

$$\left\langle \begin{pmatrix} (\max) \\ J''_m\ \Lambda''_m \\ J''\ \Lambda'' \\ M''_J\ M''_\Lambda \end{pmatrix} \middle| B \begin{pmatrix} m'_j\ m'_\lambda \\ j'\ \lambda' \\ J'_m\ 0 \\ J'\ \Lambda' \\ M'_J\ M'_\Lambda \end{pmatrix} \middle| \begin{pmatrix} (\max) \\ J_m\ \Lambda_m \\ J\ \Lambda \\ M_J\ M_\Lambda \end{pmatrix} \right\rangle$$

$$= \left[\frac{\mathfrak{M}''}{\mathfrak{M}}\right]^{1/2} \sum_q \left\langle \begin{matrix} J''_m\ \Lambda''_m \\ (\max) \end{matrix} \middle| \left\langle \begin{matrix} J'^{\,q}_m\ 0 \\ j'\ \lambda' \\ m'_j\ m'_\lambda \end{matrix} \right\rangle \middle| \begin{matrix} J_m\ \Lambda_m \\ (\max) \end{matrix} \right\rangle$$

$$\times \left\langle \begin{matrix} J''_m\ \Lambda''_m \\ J''\ \Lambda'' \\ M''_J\ M''_\Lambda \end{matrix} \middle| \left\langle \begin{matrix} J'^{\,q}_m\ 0 \\ J'\ \Lambda' \\ M'_J\ M'_\Lambda \end{matrix} \right\rangle \middle| \begin{matrix} J_m\ \Lambda_m \\ J\ \Lambda \\ M_J\ M_\Lambda \end{matrix} \right\rangle, \tag{2.1}$$

where we have denoted the single multiplicity parameter as q. The relation (2.1) is obtained from the Biedenharn factorization lemma for the unitary groups, which may be written

$$B \begin{pmatrix} (\mu)_{n-1} \\ [m]_n \\ (m)_{n-1} \end{pmatrix} = \sum_\Gamma \mathfrak{M}^{1/2} \left\langle \begin{matrix} \Gamma \\ [m]_n \\ (\mu)_{n-1} \end{matrix} \right\rangle \left\langle \begin{matrix} \Gamma \\ [m]_n \\ (\mu)_{n-1} \end{matrix} \right\rangle \mathfrak{M}^{-1/2}, \tag{2.2}$$

in which the tensor operators act in the upper and lower spaces, respectively, of the orbital group $U(n)*U(n)$. The relation (2.2) expresses the factorization of a boson state of the orbital group as a sum of products of abstract tensor operators, summed over their multiplicity labels. The boson state is understood to be taken between basis states of $U(n)*U(n)$, and the operator $\mathfrak{M}$ is an invariant operator of $U(n)*U(n)$ whose eigenvalue is the invariant measure of the state of maximal weight in $U(n)*U(n)$. The factorization lemma (2.2) is obtained from the embedding of $U(n)*U(n)$ in totally symmetric representations of $U(n^2)$, whose couplings are multiplicity-free.

In studying matrix elements of boson states in $Sp(4)*Sp(4)$, we may make use of the factorization lemma (2.2) if we consider the $Sp(4)*Sp(4)$ states to be embedded in $U(4)*U(4)$, i.e., we take into consideration the fact that

$$B \begin{pmatrix} m_j\ m_\lambda \\ j\ \lambda \\ J_m\ \Lambda_m \\ J\ \Lambda \\ M_J\ M_\Lambda \end{pmatrix} = B \begin{pmatrix} & m_j & m_\lambda & \\ & j & \lambda & \\ J_m & & & \Lambda_m \\ 2J_m & & & 2\Lambda_m\ 0\ 0 \\ J_m & & & \Lambda_m \\ & J & \Lambda & \\ & M_J & M_\Lambda & \end{pmatrix}, \tag{2.3}$$

that the $Sp(4)*Sp(4)$ boson state is a $U(4)*U(4)$ state which has maximal $Sp(4)$ labels in both upper and lower patterns. The invariant operators $\mathfrak{M}^{-1/2}$ and $\mathfrak{M}^{1/2}$ then have as eigenvalues the invariant measures of the initial and final $U(4)*U(4)$ states.

In forming the final state

$$\left|\begin{matrix}(\max)\\ J''_m\ \Lambda''_m\\ J''\ \Lambda''\\ M''_J\ M''_\Lambda\end{matrix}\right\rangle \tag{2.4}$$

in (2.1), however, we must multiply the boson state of $Sp(4)*Sp(4)$ by a power of the symplectic invariant

$$(a^{14}_{14} + a^{23}_{14} + a^{14}_{23} + a^{23}_{23})^N, \tag{2.5}$$

where $N = \Phi + \Phi' - \Phi'' - 1$, in order to obtain the correct degree conditions. This state (2.4), multiplied by the invariant (2.5), is no longer a $U(4)*U(4)$ state, but a superposition of such states. We denote the normalized product of (2.4) and (2.5) as

$$\left|\begin{matrix}m''_j\ m''_\lambda\\ j''\ \lambda''\\ N;\ J''_m\ \Lambda''_m\\ J''\ \Lambda''\\ M''_J\ M''_\Lambda\end{matrix}\right\rangle$$

$$= \left(\frac{(2\Delta''+1)(2\Phi''+5)!}{N!(N+2\Phi''+5)!(\Phi''-\Delta''-1)!(\Phi''+\Delta'')!}\right)^{1/2}$$

$$\times \langle 0|\,(\bar a^{14}_{14} + \bar a^{23}_{14} + \bar a^{14}_{23} + \bar a^{23}_{23})^N \times \overline{B}\begin{pmatrix}m''_j\ m''_\lambda\\ j''\ \lambda''\\ J''_m\ \Lambda''_m\\ J''\ \Lambda''\\ M''_J\ M''_\Lambda\end{pmatrix}, \tag{2.6}$$

where $\overline{B}$ indicates that the state is to be constructed with conjugate bosons $\bar a^i_j$. Inserting a complete set of intermediate $U(4)*U(4)$ states, we obtain for the $Sp(4)*Sp(4)$ factorization lemma

$$\left\langle\begin{matrix}m''_j\ m''_\lambda\\ j''\ \lambda''\\ N;J''_m\ \Lambda''_m\\ J''\ \Lambda''\\ M''_J\ M''_\Lambda\end{matrix}\right|B\begin{pmatrix}m'_j\ m'_\lambda\\ j'\ \lambda'\\ J'_m\ \Lambda'_m\\ J'\ \Lambda'\\ M'_J\ M'_\Lambda\end{pmatrix}\left|\begin{matrix}m_j\ m_\lambda\\ j\ \lambda\\ J_m\ \Lambda_m\\ J\ \Lambda\\ M_J\ M_\Lambda\end{matrix}\right\rangle$$

$$= \sum_{\substack{[m'']_4(\Gamma)\\ (\gamma)(\gamma')}} \times \langle N; J''_m\Lambda''_m|\,[m'']_4\rangle\mathfrak{M}^{1/2}([m'']_4)$$

$$\times\left\langle\begin{matrix}[m'']_4\\ J''_m\ \Lambda''_m\end{matrix}\right|\left|\begin{bmatrix}(\Gamma)\\ 2J'_m\ 2\Lambda'_m\ 0\ \ 0\\ J'_m\ \ \Lambda'_m\\ (\gamma)\end{bmatrix}\right|\left.\begin{matrix}2J_m\ 2\Lambda_m\ 0\ \ 0\\ J_m\ \ \Lambda_m\end{matrix}\right\rangle$$

$$\times\left\langle\begin{matrix}[m'']_4\\ J''_m\ \Lambda''_m\end{matrix}\right|\left|\begin{bmatrix}(\Gamma)\\ 2J'_m\ 2\Lambda'_m\ 0\ \ 0\\ J'_m\ \ \Lambda'_m\\ (\gamma')\end{bmatrix}\right|\left.\begin{matrix}2J_m\ 2\Lambda_m\ 0\ \ 0\\ J_m\ \ \Lambda_m\end{matrix}\right\rangle$$

$$\times\left\langle\begin{matrix}J''_m\ \Lambda''_m\\ J''\ \Lambda''\\ M''_J\ M''_\Lambda\end{matrix}\right|\left|\begin{matrix}(\gamma)\\ J'_m\ \Lambda'_m\\ J'\ \Lambda'\\ M'_J\ M'_\Lambda\end{matrix}\right\rangle\left|\begin{matrix}J_m\ \Lambda_m\\ J\ \Lambda\\ M_J\ M_\Lambda\end{matrix}\right\rangle\left\langle\begin{matrix}J''_m\ \Lambda''_m\\ j''\ \lambda''\\ m''_j\ m''_\lambda\end{matrix}\right|\left|\begin{matrix}(\gamma')\\ J'_m\ \Lambda'_m\\ j'\ \lambda'\\ m'_j\ m'_\lambda\end{matrix}\right\rangle\left|\begin{matrix}J_m\ \Lambda_m\\ j\ \lambda\\ m_j\ m_\lambda\end{matrix}\right\rangle$$

$$\times\left(\frac{(2\Delta+1)}{(\Phi-\Delta-1)!(\Phi+\Delta)!}\right)^{1/2}$$

$$\equiv \left(\frac{(2\Delta''+1)(2\Phi''+5)!}{N!(N+2\Phi''+5)!(\Phi''-\Delta''-1)!(\Phi''+\Delta'')!}\right)^{1/2}$$

$$\times\left(\frac{(2\Delta+1)}{(\Phi-\Delta-1)!(\Phi+\Delta)!}\right)^{1/2}$$

$$\times\ \mathfrak{M}(N;J''_m\Lambda''_m)\sum_{(\gamma)}\left\langle\begin{matrix}J''_m\ \Lambda''_m\\ J''\ \Lambda''\\ M''_J\ M''_\Lambda\end{matrix}\right|\left\langle\begin{matrix}(\gamma)\\ J'_m\ \Lambda'_m\\ J'\ \Lambda'\\ M'_J\ M'_\Lambda\end{matrix}\right\rangle\left|\begin{matrix}J_m\ \Lambda_m\\ J\ \Lambda\\ M_J\ M_\Lambda\end{matrix}\right\rangle$$

$$\times\left\langle\begin{matrix}J''_m\ \Lambda''_m\\ j''\ \lambda''\\ m''_j\ m''_\lambda\end{matrix}\right|\left\langle\begin{matrix}(\gamma)\\ J'_m\ \Lambda'_m\\ j'\ \lambda'\\ m'_j\ m'_\lambda\end{matrix}\right\rangle\left|\begin{matrix}J_m\ \Lambda_m\\ j\ \lambda\\ m_j\ m_\lambda\end{matrix}\right\rangle, \tag{2.7}$$

where $\mathfrak{M}(N, J''_m\ \Lambda''_m)$ is a constant of normalization. The third member of the equation is guaranteed by the isomorphism of the $Sp(4)*Sp(4)$ states with the matrix elements of finite transformations in the corresponding irreducible representation of $Sp(4)$. The expression (2.1), then, is to be interpreted in terms of (2.7). In this paper we shall not determine the constant $\mathfrak{M}(N, J''_m\ \Lambda''_m)$, but shall consider only the techniques of factorization.

We now turn to the special case of the matrix elements of totally symmetric tensor operators in $Sp(4)$, and we shall perform the factorization indicated in (2.1). In order to do so we shall first perform the factorization for the simpler case in which the initial and final states of both matrix elements on the right of (2.1) are maximal. This case is trivial, since the coupling process is analogous to that involved in the factorization of a $U(2)*U(2)$ matrix element into two $SU(2)$ matrix elements. We shall quote only the result for the matrix element of the reduced tensor operator:

$$\left\langle\begin{matrix}J''_m\ \Lambda''_m\\ (\max)\end{matrix}\right|\left|\begin{bmatrix}q\\ J'_m\ \ 0\\ J'\ \ \Lambda'\\ J''_m - J_m\ \Lambda''_m - \Lambda_m\end{bmatrix}\right|\left.\begin{matrix}J_m\ \Lambda_m\\ (\max)\end{matrix}\right\rangle$$

$$= (-1)^{\Lambda_m+\Lambda'-\Lambda''_m}\left(\frac{(\Phi+\Delta)(\Phi+J'_m-\Phi'')!(2q+1)}{(2\Delta+1)(2\Delta''+1)\mathfrak{M}(N;J''_m\Lambda''_m)(\Phi''-\Delta'')}\right)^{1/2}$$

$$\times\left[\frac{(J''_m+J''_m-J')!(J''_m+J_m+J'+1)!}{[\tfrac12(\Phi+\Phi''-J'_m)-1-q]!}\right.$$

$$\left.\times\ \frac{(\Lambda''_m+\Lambda_m-\Lambda')!(\Lambda''_m+\Lambda_m+\Lambda'+1)!}{[\tfrac12(\Phi+\Phi''-J'_m)+q]!}\right]^{1/2}$$

$$\times\ C^{\frac12(\Phi-\Phi''+J'_m)\quad q\quad \Delta}_{\frac12(\Delta-\Delta''+\delta')\quad \frac12(\Delta+\Delta''-\delta')\quad \Delta}$$

$$\times\ C^{\frac12(\Phi''-\Phi+J'_m)\quad q\quad \Delta''}_{\frac12(\Delta''-\Delta+\delta')\quad \frac12(\Delta+\Delta''-\delta'')\quad \Delta''}, \tag{2.8}$$

where we use the abbreviations

$$\Phi = J_m + \Lambda_m + 1, \qquad \Delta = J_m - \Lambda_m,$$

$$\varphi = J + \Lambda + 1, \qquad \delta = J - \Lambda \tag{2.9}$$

and similarly for primed quantities.

Having determined the matrix element (2.8), we may now proceed to the evaluation of more general matrix elements, knowing the exact form of the first element on the right of (2.1), which is to be removed in the factorization. At this point it is instructive to examine the coupling process in detail. We form the boson state[8,10]

$$\left|\begin{pmatrix}(\max)\\ J_m\ \Lambda_m\\ J\ \ \Lambda\end{pmatrix}\right\rangle = \sum_{J_1 J_2 \Lambda_1 \Lambda_2} \sum_{\substack{m_1 m_2\\ \mu_1 \mu_2}} (-1)^z C^{\frac{1}{2}(\Phi-\varphi)\ \ \frac{1}{2}(\Phi+\varphi)\ \ \Phi}_{\frac{1}{2}(\Phi-\varphi)-z\ \ z+\delta-\frac{1}{2}(\Phi-\varphi)\ \ \delta} \times C^{J\ \ \Lambda\ \ \Delta}_{J_1-J_2\ \ \Lambda_1-\Lambda_2\ \ \Delta} C^{J_1\ J_2\ J}_{m_1\ m_2\ J} C^{\Lambda_1\ \Lambda_2\ \Lambda}_{\mu_1\ \mu_2\ \Lambda}$$

$$\times \frac{1}{[(2J_1)!(2J_2)!(2\Lambda_1)!(2\Lambda_2)!]^{1/2}} d^{J_1}_{m_1 J_1}(a^{14}_{14})\ d^{J_2}_{m_2 J_2}(a^{23}_{14})$$

$$\times d^{\Lambda_1}_{\mu_1 \Lambda_1}(a^{14}_{23}) d^{\Lambda_2}_{\mu_2 \Lambda_2}(a^{23}_{23})|0\rangle, \tag{2.10}$$

where

$$z = J_1 + J_2 - J, \quad J_1 + \Lambda_1 = J_m, \quad J_2 + \Lambda_2 = \Lambda_m, \tag{2.11}$$

and

$$d^j_{m\mu}(a^i_\ell{}^k_p) = \sum_x \frac{[(j+m)!(j-m)!(j+\mu)!(j-\mu)!]^{1/2}}{x!(j-m-x)!(j-\mu-x)!(m+\mu+x)!} \times (a^i_\ell)^{m+\mu+x}(a^i_p)^{j-m-x}(a^k_\ell)^{j-\mu-x}(a^k_p)^x. \tag{2.12}$$

Similarly,

$$B\begin{pmatrix} m'_j & m'_\lambda\\ j' & \lambda'\\ J'_m & 0\\ J' & \Lambda'\\ M'_J & M'_\Lambda\end{pmatrix} = [(2J'_m)!]^{1/2} \sum_{J'_1 J'_2 \Lambda'_1 \Lambda'_2} \sum_{\substack{m'_1 m'_2\\ \mu'_1 \mu'_2}} \sum_{\substack{n'_1 n'_2\\ \nu'_1 \nu'_2}} \times C^{J'\ \ \Lambda'\ \ J'_m}_{J'_1-J'_2\ \ \Lambda'_1-\Lambda'_2\ \ j'-\lambda'} C^{J'_1\ J'_2\ J'}_{m'_1\ m'_2\ M'_J}$$

$$\times (-1)^{\Lambda'-M'_\Lambda} C^{\Lambda'_1\ \Lambda'_2\ \Lambda'}_{\mu'_1\ \mu'_2\ M'_\Lambda} C^{J'_1\ \Lambda'_1\ j'}_{n'_1\ n'_2\ m'_j} (-1)^{\lambda'-m'_\lambda} C^{J'_2\ \Lambda'_2\ \lambda'}_{\nu'_1\ \nu'_2\ m'_\lambda}$$

$$\times \frac{1}{[(2J'_1)!(2J'_2)!(2\Lambda'_1)!(2\Lambda'_2)!]^{1/2}} d^{J'_1}_{m'_1 n'_1}(a^{14}_{14}) d^{J'_2}_{m'_2 \nu'_1}(a^{23}_{14})$$

$$\times d^{\Lambda'_1}_{\mu'_1 n'_2}(a^{14}_{23}) d^{\Lambda'_2}_{\mu'_2 \nu'_2}(a^{23}_{23}), \tag{2.13}$$

$$J'_1 + \Lambda'_1 = j', \quad J'_2 + \Lambda'_2 = \lambda', \quad J'_1 + J'_2 = J',$$

$$\Lambda'_1 + \Lambda'_2 = \Lambda', \quad J' + \Lambda' = j' + \lambda' = J'_m, \tag{2.14}$$

and

$$\left\langle\left\langle\begin{pmatrix}(\max)\\ J''_m\ \Lambda''_m\\ J''\ \ \Lambda''\\ J''\ \ \Lambda''\end{pmatrix}\right| = \left(\frac{(2\Phi''+5)!}{N!(N+2\Phi''+5)!}\right)^{1/2} \langle 0|(\bar a^{14}_{14} + \bar a^{14}_{23} + \bar a^{23}_{14} + \bar a^{23}_{23})^N \sum_{J''_1 \Lambda_1'' J''_2 \Lambda''_2} \sum_{\substack{m''_1 m''_2\\ \mu''_1 \mu''_2}} (-1)^{z''}$$

$$\times C^{\frac{1}{2}(\Phi''-\varphi'')\ \ \frac{1}{2}(\Phi''+\varphi'')\ \ \Phi''}_{\frac{1}{2}(\Phi''-\varphi'')-z''\ \ z''+\delta''-\frac{1}{2}(\Phi''-\varphi'')\ \ \delta''} C^{J''\ \ \Lambda''\ \ \Delta''}_{J''_1-J''_2\ \ \Lambda''_1-\Lambda''_2\ \ \Delta''} \times C^{J''_1\ J''_2\ J''}_{m''_1\ m''_2\ J''} C^{\Lambda''_1\ \Lambda''_2\ \Lambda''}_{\mu''_1\ \mu''_2\ \Lambda''} \frac{1}{[2J''_1)!(2\Lambda''_1)!(2J''_2)!(2\Lambda''_2)!]^{1/2}}$$

$$\times d^{J''_1}_{m''_1 J''_1}(\bar a^{14}_{14})\ d^{J''_2}_{m''_2 J''_2}(\bar a^{23}_{14})\ d^{\Lambda''_1}_{\mu''_1 \Lambda''_1}(\bar a^{14}_{23})\ d^{\Lambda''_2}_{\mu''_2 \Lambda''_2}(\bar a^{23}_{23}), \tag{2.15}$$

and the summations are to be taken with the restrictions (2.11) above. Here $N = \Phi + J'_m - \Phi''$ is the power of the $Sp(4)*Sp(4)$ invariant necessary to establish the correct degree conditions when the commutations are performed. We now evaluate

$$\left\langle\left\langle\begin{pmatrix}(\max)\\ J''_m\ \Lambda''_m\\ J''\ \ \Lambda''\\ J''\ \ \Lambda''\end{pmatrix}\right| B\begin{pmatrix} m'_j & m'_\lambda\\ j' & \lambda'\\ J'_m & 0\\ J' & \Lambda'\\ M'_J & M'_\Lambda\end{pmatrix} \left|\left\langle\begin{pmatrix}(\max)\\ J_m\ \Lambda_m\\ J\ \ \Lambda\end{pmatrix}\right\rangle\right. = (-1)^{\Lambda'+\Lambda-\Lambda''}$$

$$\times (-1)^{\Lambda_m+\lambda'-\Lambda''_m} \left(\frac{(2\Phi''+5)!(2J'_m)!(\Phi+J'_m-\Phi'')!}{(\Phi+J'_m+\Phi''+5)!}\right)^{1/2}$$

$$\times \sum_{\substack{J_i J'_i J''_i\\ \Lambda_i \Lambda'_i \Lambda''_i}} (-1)^{z+z''} C^{\frac{1}{2}(\Phi''-\varphi'')\ \ \frac{1}{2}(\Phi''+\varphi'')\ \ \Phi''}_{\frac{1}{2}(\Phi''-\varphi'')-z''\ \ z''+\delta''-\frac{1}{2}(\Phi''-\varphi'')\ \ \delta''} \times C^{\frac{1}{2}(\Phi-\varphi)\ \ \frac{1}{2}(\Phi+\varphi)\ \ \Phi}_{\frac{1}{2}(\Phi-\varphi)-z\ \ z+\delta-\frac{1}{2}(\Phi-\varphi)\ \ \delta} C^{J\ \ \Lambda\ \ \Delta}_{J_1-J_2\ \ \Lambda_1-\Lambda_2\ \ \Delta} C^{J'\ \ \Lambda'\ \ J'_m}_{J'_1-J'_2\ \ \Lambda'_1-\Lambda'_2\ \ j'-\lambda'}$$

$$\times C^{J''\ \ \Lambda''\ \ \Delta''}_{J''_1-J''_2\ \ \Lambda''_1-\Lambda''_2\ \ \Delta''} (J_1+J'_1+J''_1+1)!(\Lambda_1+\Lambda'_1+\Lambda''_1+1)!$$

$$\times (J_2+J'_2+J''_2+1)!(\Lambda_2+\Lambda'_2+\Lambda''_2+1)!$$

$$\times \left(\frac{(2J+1)(2J'+1)(2j'+1)(2J_m+1)}{(2J_1)!(2J_2)!(2J'_1)!(2J'_2)!(2J''_1)!(2J''_2)!} \times \frac{(2\Lambda+1)(2\Lambda'+1)(2\lambda'+1)(2\Lambda_m+1)}{(2\Lambda_1)!(2\Lambda_2)!(2\Lambda'_1)!(2\Lambda'_2)!(2\Lambda''_1)!(2\Lambda''_2)!}\right)^{1/2}$$

$$\times \begin{Bmatrix} J_1 & J_2 & J\\ J'_1 & J'_2 & J'\\ J''_1 & J''_2 & J''\end{Bmatrix} \begin{Bmatrix} \Lambda_1 & \Lambda_2 & \Lambda\\ \Lambda'_1 & \Lambda'_2 & \Lambda'\\ \Lambda''_1 & \Lambda''_2 & \Lambda''\end{Bmatrix} \begin{Bmatrix} J_1 & \Lambda_1 & J_m\\ J'_1 & \Lambda'_1 & j'\\ J''_1 & \Lambda''_1 & J''_m\end{Bmatrix} \begin{Bmatrix} J_2 & \Lambda_2 & \Lambda_m\\ J'_2 & \Lambda'_2 & \lambda'\\ J''_2 & \Lambda''_2 & \Lambda''_m\end{Bmatrix}$$

$$\times C^{J\ \ J'\ \ J''}_{J\ \ J''-J\ \ J''} C^{\Lambda\ \ \Lambda'\ \ \Lambda''}_{\Lambda\ \ \Lambda''-\Lambda\ \ \Lambda''} C^{J_m\ \ j'\ \ J''_m}_{J_m\ \ J''_m-J_m\ \ J''_m} C^{\Lambda_m\ \ \lambda'\ \ \Lambda''_m}_{\Lambda_m\ \ \Lambda''_m-\Lambda_m\ \ \Lambda''_m}. \tag{2.16}$$

We disregard the last four $SU(2)$ Wigner coefficients, since these are merely the coupling coefficients of the $SU(2) \times SU(2)$ subgroups of lower and upper pattern spaces. Examination of (2.16) shows us what our immediate task must be: We must separate the (9-j) coefficients into factors in which $J_1 - J_2, \Lambda_1 - \Lambda_2$, and the corresponding primed quantities appear as magnetic quantum numbers, and similarly for the quantities $J_1 + J_2, \Lambda_1 + \Lambda_2$, and their primes. The last two (9-j) symbols in (2.16) are triply stretched, and hence are monomials. The first two are only singly stretched. We make use of the following identity for the singly stretched (9-j) symbol 9:

$$\begin{Bmatrix} j_2 & j_4 & j_{24}\\ j_{12} & j_{34} & j_{12}+j_{34}\\ j_1 & j_3 & j_{13}\end{Bmatrix} = (-1)^{j_4+j_{13}-j_{34}-j_1} \left(\frac{(2j_{12})!(2j_{34})!(j_1+j_2-j_{12})!(j_3+j_4-j_{34})!(j_1+j_3+j_{13}+1)!}{(2j_{12}+2j_{34}+1)!(2j_{13})!(j_1+j_3-j_{13})!(j_1+j_2+j_{12}+1)!}\right)^{1/2}$$

$$\times\left(\frac{(j_{12}+j_{13}-j_{24}+j_{34})!(j_{12}+j_{13}+j_{24}+j_{34}+1)!(j_2+j_4+j_{12}+j_{34}-j_{13}+1)!}{(j_1+j_3-j_2-j_4+j_{12}+j_{34})!(j_3+j_4+j_{34}+1)!(j_2+j_4+j_{13}-j_{12}-j_{34})!(j_2+j_4-j_1-j_3+j_{12}+j_{34})!}\right)^{1/2}$$

$$\times\sum_{\Psi} C^{j_{24}\ \ j_{13}\ \ \Psi}_{j_2-j_4\ \ j_3-j_1\ \ j_2-j_4+j_3-j_1} C^{\frac{1}{2}(j_2+j_4-j_1-j_3+j_{12}+j_{34})\ \frac{1}{2}(j_1+j_3-j_2-j_4+j_{12}+j_{34})\ \ \Psi}_{\frac{1}{2}(j_2-j_4-j_1+j_3+j_{12}-j_{34})\ \frac{1}{2}(j_2-j_4-j_1+j_3-j_{12}+j_{34})\ j_2-j_4+j_3-j_1} \begin{Bmatrix} \frac{1}{2}(j_2+j_4+j_{12}+j_{34}-j_{13}) & \frac{1}{2}(j_2+j_4+j_{13}-j_{12}-j_{34}) & j_{24} \\ \frac{1}{2}(j_1+j_3-j_{13}) & \frac{1}{2}(j_1+j_3+j_{13}) & j_{13} \\ \frac{1}{2}(j_2+j_4-j_1-j_3+j_{12}+j_{34}) & \frac{1}{2}(j_1+j_3-j_2-j_4+j_{12}+j_{34}) & \Psi \end{Bmatrix}. \tag{2.17}$$

The $(9\text{-}j)$ symbol on the right is triply degenerate and has the form of a monomial multiplied by a ${}_3F_2$ series; it is thus proportional to the analytic continuation of an $SU(2)$ Wigner coefficient. Quantities dependent on the differences between the first two columns of the $(9\text{-}j)$ symbol on the left are now isolated in the monomial coefficient and the magnetic quantum numbers of the indicated $SU(2)$ Wigner coefficients. We apply the identity (2.17) to the first two $(9\text{-}j)$ symbols on the right of (2.16), labeling the parameter Ψ as Ψ_J and Ψ_Λ in the two cases, respectively, and perform the summation

$$\sum C^{J\ \ J''\ \ \Psi_J}_{J_1-J_2\ \ J_2''-J_1''\ \ J_1-J_2+J_2''-J_1''} C^{\Lambda\ \ \Lambda''\ \ \Psi_\Lambda}_{\Lambda_1-\Lambda_2\ \ \Lambda_2''-\Lambda_1''\ \ \Lambda_1-\Lambda_2+\Lambda_2''-\Lambda_1''}$$

$$\times\ C^{J\ \ \Lambda\ \ \Delta}_{J_1-J_2\ \ \Lambda_1-\Lambda_2\ \ \Delta} C^{J''\ \ \Lambda''\ \ \Delta''}_{J_2''-J_1''\ \ \Lambda_2''-\Lambda_1''\ \ -\Delta''}$$

$$=\sum_{\Psi}[(2\Psi_J+1)(2\Psi_\Lambda+1)(2\Delta+1)(2\Delta''+1)]^{1/2}\ C^{\Delta\ \ \Delta''\ \ \Psi}_{\Delta\ -\Delta''\ \Delta-\Delta''}$$

$$\times\ C^{\Psi_J\ \ \Psi_\Lambda\ \ \Psi}_{J_1-J_2+J_2''-J_1''\ \ \Lambda_1-\Lambda_2+\Lambda_2''-\Lambda_1''\ \ \Delta-\Delta''} \begin{Bmatrix} J & J'' & \Psi_J \\ \Lambda & \Lambda'' & \Psi_\Lambda \\ \Delta & \Delta'' & \Psi \end{Bmatrix}. \tag{2.18}$$

We note that the dependence of the Wigner coefficient

$$C^{J'\ \ \Lambda'\ \ J'_m}_{J_1'-J_2'\ \ \Lambda_1'-\Lambda_2'\ \ j'-\lambda'} = \left(\frac{(2J')!(2\Lambda')!(J'_m+j'-\lambda')!(J'_m-j'+\lambda')!}{(2J'_m)!(J'-J_1'+J_2')!(J'+J_1'-J_2')!(\Lambda'-\Lambda_1'+\Lambda_2')!(\Lambda'+\Lambda_1'-\Lambda_2')!}\right)^{1/2} \tag{2.19}$$

on its first two magnetic quantum numbers is canceled by factors represented by $[(2j_{12})!(2j_{34})!]^{1/2}$ in (2.17). The final two $(9\text{-}j)$ symbols on the right of (2.16) are monomials which cancel the factor in curly brackets in (2.16).

From monomial factors provided in (2.17) we may now form two stretched $SU(2)$ Wigner coefficients and sum:

$$\sum C^{\frac{1}{2}(J_1+J_2-J_1''-J_2''+J')\ \ \frac{1}{2}(J_1''+J_2''-J_1-J_2+J')\ \ \Psi_J}_{\frac{1}{2}(J_1-J_2-J_1''+J_2''+J_1'-J_2')\ \frac{1}{2}(J_1-J_2-J_1''+J_2''-J_1'+J_2')\ J_1-J_2+J_2''-J_1''} C^{\frac{1}{2}(\Lambda_1+\Lambda_2-\Lambda_1''-\Lambda_2''+\Lambda')\ \ \frac{1}{2}(\Lambda_1''+\Lambda_2''-\Lambda_1-\Lambda_2+\Lambda')\ \ \Psi_\Lambda}_{\frac{1}{2}(\Lambda_1-\Lambda_2-\Lambda_1''+\Lambda_2''+\Lambda_1'-\Lambda_2')\ \frac{1}{2}(\Lambda_1-\Lambda_2-\Lambda_1''+\Lambda_2''-\Lambda_1'+\Lambda_2')\ \Lambda_1-\Lambda_2+\Lambda_2''-\Lambda_1''}$$

$$\times\ C^{\frac{1}{2}(J_1+J_2-J_1''-J_2''+J')\ \ \frac{1}{2}(\Lambda_1+\Lambda_2-\Lambda_1''-\Lambda_2''+\Lambda')\ \ \frac{1}{2}(\Phi-\Phi''+J'_m)}_{\frac{1}{2}(J_1-J_2-J_1''+J_2''+J_1'-J_2')\frac{1}{2}(\Lambda_1-\Lambda_2-\Lambda_1''+\Lambda_2''+\Lambda_1'-\Lambda_2')\frac{1}{2}(\Delta-\Delta''+j'-\lambda')} C^{\frac{1}{2}(J_1''+J_2''-J_1-J_2+J')\ \ \frac{1}{2}(\Lambda_1''+\Lambda_2''-\Lambda_1-\Lambda_2+\Lambda')\ \ \frac{1}{2}(\Phi''-\Phi+J'_m)}_{\frac{1}{2}(J_1-J_2-J_1''+J_2''-J_1'+J_2')\ \frac{1}{2}(\Lambda_1-\Lambda_2-\Lambda_1''+\Lambda_2''-\Lambda_1'+\Lambda_2')\ \frac{1}{2}(\Delta-\Delta''-j'+\lambda')}$$

$$\times\ C^{\Psi_J\ \ \Psi_\Lambda\ \ \Psi}_{J_1-J_2+J_2''-J_1''\ \ \Lambda_1-\Lambda_2+\Lambda_2''-\Lambda_1''\ \ \Delta-\Delta''}$$

$$=[(2\Psi_J+1)(2\Psi_\Lambda+1)(\Phi-\Phi''+J'_m+1)(\Phi''-\Phi+J'_m+1)]^{1/2}$$

$$\times \begin{Bmatrix} \frac{1}{2}(J_1+J_2-J_1''-J_2''+J') & \frac{1}{2}(J_1''+J_2''-J_1-J_2+J') & \Psi_J \\ \frac{1}{2}(\Lambda_1+\Lambda_2-\Lambda_1''-\Lambda_2''+\Lambda') & \frac{1}{2}(\Lambda_1''+\Lambda_2''-\Lambda_1-\Lambda_2+\Lambda') & \Psi_\Lambda \\ \frac{1}{2}(\Phi-\Phi''+J'_m) & \frac{1}{2}(\Phi''-\Phi+J'_m) & \Psi \end{Bmatrix} C^{\frac{1}{2}(\Phi-\Phi''+J'_m)\ \ \frac{1}{2}(\Phi''-\Phi+J'_m)\ \ \Psi}_{\frac{1}{2}(\Delta-\Delta''+j'-\lambda')\ \frac{1}{2}(\Delta-\Delta''-j'+\lambda')\ \Delta-\Delta''}$$

$$=\left(\frac{(2\Psi_J+1)(2\Psi_\Lambda+1)(J'_m-\Psi)!(J'_m+\Psi+1)!}{(2\Psi+1)(\Phi-\Phi''+J'_m)!(\Phi''-\Phi+J'_m)!(J'-\Psi_J)!}\right)^{1/2},$$

$$\times\left(\frac{(J_1+J_2-J_1''-J_2''+J')!(J_1''+J_2''-J_1-J_2+J')!(\Lambda_1+\Lambda_2-\Lambda_1''-\Lambda_2''+\Lambda')!(\Lambda_1''+\Lambda_2''-\Lambda_1-\Lambda_2+\Lambda')!}{(J'+\Psi_J+1)!(\Lambda'-\Psi_\Lambda)!(\Lambda'+\Psi_\Lambda+1)!}\right)^{1/2}$$

$$\times\ C^{\Psi_J\ \ \Psi_\Lambda\ \ \Psi}_{J_1+J_2-J_1''-J_2''\ \ \Lambda_1+\Lambda_2-\Lambda_1''-\Lambda_2''\ \ \Phi-\Phi''} C^{\frac{1}{2}(\Phi-\Phi''+J'_m)\ \ \frac{1}{2}(\Phi''-\Phi+J'_m)\ \ \Psi}_{\frac{1}{2}(\Delta-\Delta''+j'-\lambda')\ \frac{1}{2}(\Delta-\Delta''+j'-\lambda')\ \Delta-\Delta''}. \tag{2.20}$$

as monomial factors, and we write $z_1=z-z''$, $z_2=z''$. We then treat the two triply degenerate $(9\text{-}j)$ symbols which emerge from our application of the identity (2.17) by means of the relation[11]

$$C^{\frac{1}{2}(\Phi-\varphi)\ \ \frac{1}{2}(\Phi+\varphi)\ \ \Phi}_{\frac{1}{2}(\Phi-\varphi)-z\ \ z+\delta-\frac{1}{2}(\Phi-\varphi)\ \ \delta} C^{\frac{1}{2}(\Phi''-\varphi'')\ \ \frac{1}{2}(\Phi''+\varphi'')\ \ \Phi''}_{\frac{1}{2}(\Phi''-\varphi'')-z''\ z''+\delta''-\frac{1}{2}(\Phi''-\varphi'')\ \delta''} \tag{2.21}$$

It remains to perform the sums over $z=J_1+J_2-J$ and $z''=J_1''+J_2''-J''$. We write out explicitly the degenerate Wigner coefficients

$$\begin{Bmatrix} j_1 & j_2 & j_1+j_2 \\ j_3 & j_4 & j_{34} \\ j_1+j_3 & j_{24} & j_1+j_2-j_{34} \end{Bmatrix} = (-1)^{-j_2+j_3+j_{34}+j_{24}} \left(\frac{(2j_2)!(2j_3)!(2j_{34})!(2j_1+2j_2-2j_{34})!(-j_2+j_4+j_{24})!}{(2j_1+2j_3+1)!(j_2-j_4+j_{24})!(j_2+j_4+j_{24}+1)!(j_2+j_4-j_{24})!}\right)^{1/2}$$

$$\times\left(\frac{(-j_3+j_4+j_{34})!(2j_1+j_2+j_3+j_{24}-j_{34}+1)!(2j_1+j_2+j_3-j_{24}-j_{34})!(j_3-j_2+j_{24}+j_{34})!}{(j_2-j_3+j_{24}-j_{34})!(j_3-j_4+j_{34})!(j_3+j_4+j_{34}+1)!(j_3+j_4-j_{34})!}\right)^{1/2}$$

$$\times\sum_x\frac{(-1)^x}{x!}\frac{(2j_{24}-x)!(j_2+j_4-j_{24}+x)!}{(j_2-j_3+j_{34}-j_{24}+x)!(j_4-j_2+j_{24}-x)!(2j_1+j_2+j_3-j_{34}+j_{24}+1-x)!}. \tag{2.22}$$

We may immediately perform the sum over z_2, since this is merely a ${}_2F_1$ series of unit argument. The remaining three sums, over z_1 and the two indices of summation x_1 and x_2 obtained by application of the identity (2.22) to the (9-j) symbols which result from the use of (2.17), have the form

$$\sum_{z_1x_1x_2}(-1)^{z_1+x_1+x_2}\frac{(\Psi_J-J+J''+x_1)!(2J-x_1)!(\Psi_\Lambda-\Lambda+\Lambda''+x_2)!(2\Lambda-x_2)!(\Phi''+\Phi+\varphi-\varphi''-x_1-x_2)!}{x_1!(\Psi_J+J-J''-x_1)!(2J''-2J-z_1+x_1)!x_2!(\Psi_\Lambda+\Lambda-\Lambda''-x_2)!(\Phi''-\Phi+\delta-\delta''+z_1+x_2)!}$$

$$\times\frac{1}{(\Phi+\varphi-x_1-x_2)!(\Phi''+\varphi+\delta-\varphi''+z_1-x_1)!(\Phi-\delta-z_1-x_2)!}$$

$$\times\left(\frac{(\Psi_\Lambda-\Phi+\Phi''+J-J''+z_1)!(\Psi_J-J+J''-z_1)!}{(\Psi_\Lambda+\Phi-\Phi''-J+J''-z_1)!(\Psi_J+J-J''+z_1)!}\right)^{1/2}C^{\Psi_J\ \ \Psi_\Lambda\ \ \Psi}_{J-J''+z_1\ \ \Phi-\Phi''-J+J''-z_1\ \ \Phi-\Phi''}\equiv S. \tag{2.23}$$

This sum may be compared with the expression for the general (9-j) symbol[12]

$$\begin{Bmatrix}j_1&j_2&j_{12}\\j_3&j_4&j_{34}\\j_{13}&j_{24}&j\end{Bmatrix}=\left(\frac{(j_{13}+j_{24}+j+1)!(j_{12}+j_{34}+j+1)!(j_{13}-j_{24}+j)!(j_1+j_3-j_{13})!(j_{12}+j_{34}-j)!}{(j_1-j_3+j_{13})!(-j_1+j_3+j_{13})!(j_1+j_3+j_{13}+1)!(j_{12}-j_{34}+j)!(-j_{12}+j_{34}+j)!}\right)^{1/2}$$

$$\times\frac{\Delta(j_1j_2j_{12})\Delta(j_3j_4j_{34})}{(j_{12}+j_1-j_2)!(j_{12}-j_1+j_2)!(j_{34}+j_3-j_4)!(j_{34}-j_3+j_4)![(2j_{24}+1)]^{1/2}}\sum_{xym}(-1)^{j_1-j_{12}+m+x+y}C^{j_2\ \ j_4\ \ j_{24}}_{m\ \ j-j_{13}-m\ \ j-j_{13}}$$

$$\times\left(\frac{(j_2-m)!(j_4-j+j_{13}+m)!}{(j_2+m)!(j_4+j-j_{13}-m)!}\right)^{1/2}\frac{(2j_1-x)!(j_{12}+j_2-j_1+x)!(j_{34}+j-j_1-m+x)!}{x!(j_{12}-j_1-m+x)!(j_1+j_2-j_{12}-x)!y!}$$

$$\times\frac{(2j_3-y)!(j_{34}+j_4-j_3+y)!}{(j_{34}-j_3-j+j_{13}+m+y)!(j_3+j_4-j_{34}-y)!}\frac{(j_{12}+j_{13}-j_3+m+y)!}{(j_1+j_3-j_{13}-x-y)!(-j_1-j_3+j_{13}+j_{12}+j_{34}+j+1+x+y)} \tag{2.24}$$

it is evident that the sum (2.23) is proportional to (2.24) once we make the analytic continuations $\Phi\to-\Phi-1$, $\Phi''\to-\Phi''-1$, i.e., (2.23) is proportional to an analytically continued (9-j) symbol. Specifically,

$$i(-1)^{\frac{1}{2}(\Phi''-\Phi-\varphi''+\varphi)}\begin{Bmatrix}J&\Psi_J&J''\\\Lambda&\Psi_\Lambda&\Lambda''\\-\Phi-1&\Psi-\Phi''-1&\end{Bmatrix}=(-1)^{2J-2J''}\left(\frac{(\Phi-\varphi)!(\Phi+\varphi)!(\Phi-\delta)!(\Phi+\delta)!(\Phi''+\varphi'')!}{(\Phi+\Phi''-\Psi)!(\Phi+\Phi''+\Psi+1)!(\Phi''-\varphi'')!(2\Psi+1)}\right)^{1/2}$$

$$\times\frac{[(\Phi''-\delta'')!(\Phi''+\delta'')!]^{1/2}\Delta(JJ''\Psi_J)\Delta(\Lambda\Lambda''\Psi_\Lambda)}{(\Psi_J+J''-J)!(J+J''-\Psi_J)!(\Psi_\Lambda+\Lambda''-\Lambda)!(\Lambda+\Lambda''-\Psi_\Lambda)!}S, \tag{2.25}$$

where S is given by (2.23).

We must factor out the dependence on j' and λ' into a factor of the form (2.8). From (2.18) and (2.20), respectively, we remove the Wigner coefficients

$$C^{\Delta\ \Delta''\ \ \Psi}_{\Delta\ -\Delta''\ \Delta-\Delta''}C^{\frac{1}{2}(\Phi-\Phi''+J'_m)\ \ \frac{1}{2}(\Phi''-\Phi+J'_m)\ \ \Psi}_{\frac{1}{2}(\Delta-\Delta''+j'-\lambda')\ \frac{1}{2}(\Delta-\Delta''-j'+\lambda')\ \Delta-\Delta''}=\sum_q(-1)^{\Psi+\Delta+\Delta''}(-1)^{\Delta+\frac{1}{2}(\Phi-\Phi''+J'_m)-q}(2\Psi+1)\left(\frac{(2q+1)}{(2\Delta''+1)}\right)^{1/2}$$

$$\times\begin{Bmatrix}\Delta&\frac{1}{2}(\Phi-\Phi''+J'_m)&q\\\frac{1}{2}(\Phi''-\Phi+J'_m)&\Delta''&\Psi\end{Bmatrix}C^{\frac{1}{2}(\Phi-\Phi''+J'_m)\ \ \Delta\ q}_{\frac{1}{2}(\Delta''-\Delta-j'+\lambda')\ \Delta\ \ \frac{1}{2}(\Delta+\Delta''-j'+\lambda')}C^{q\ \ \frac{1}{2}(\Phi''-\Phi+J'_m)\ \ \Delta''}_{\frac{1}{2}(\Delta+\Delta''-j'+\lambda')\ \frac{1}{2}(\Delta''-\Delta+j'-\lambda')\ \Delta''}. \tag{2.26}$$

All remaining dependence on j' and λ' is contained in monomial factors prescribed by (2.8), so we can use q as our index of multiplicity. Removing the Wigner coefficient of upper pattern space and inserting normalization factors as prescribed by (2.7), we find the general matrix element of a totally symmetric unit tensor operator in $Sp(4)$:

$$\left\langle\begin{matrix}J''_m&\Lambda''_m\\J''&\Lambda''\\M''_J&M''_\Lambda\end{matrix}\middle|\middle\langle\begin{matrix}q\\J'_m\ 0\\J'\ \ \Lambda'\\M'_J\ M'_\Lambda\end{matrix}\middle\rangle\middle|\begin{matrix}J_m&\Lambda_m\\J&\Lambda\\M_J&M_\Lambda\end{matrix}\right\rangle=(-1)^{\Lambda''+\Lambda'-\Lambda+\frac{1}{2}(\Phi''-\Phi+J''_m)+q-\Delta''+2\Delta+2J'_m}\left(\frac{(2J+1)(2\Lambda+1)(\Phi+J'_m-\Phi'')!}{\mathfrak{N}(N;J''_m\Lambda''_m)(2\Phi)!(2\Phi'')!}\right)^{1/2}$$

$$\times \left(\frac{(2q+1)[\tfrac{1}{2}(\Phi+\Phi''-J'_m)-1-q]![\tfrac{1}{2}(\Phi+\Phi''-J'_m)+q]!(J+J'-J'')!}{(J-J'+J'')!} \right.$$

$$\left. \times \frac{(-J+J'+J'')!(J+J'+J''+1)!(\Lambda+\Lambda'-\Lambda'')!(-\Lambda+\Lambda'+\Lambda'')!}{(\Lambda-\Lambda'+\Lambda'')!} \right)^{1/2} [(\Lambda+\Lambda'+\Lambda''+1)!]^{1/2} \sum_{\Psi_J \Psi_\Lambda \Psi} (-1)^{\Psi_J+\Psi_\Lambda+\Psi}$$

$$\times (2\Psi_J+1)(2\Psi_\Lambda+1)(2\Psi+1) \frac{[(J'_m-\Psi)!(J'_m+\Psi+1)!(\Phi+\Phi''-\Psi)!(\Phi+\Phi''+\Psi+1)!]^{1/2}}{(J'-\Psi_J)!(J'+\Psi_J+1)!(\Lambda'-\Psi_\Lambda)!(\Lambda'+\Psi_\Lambda+1)!}$$

$$\times i(-1)^{\frac{1}{2}(\Phi''-\varphi''-\Phi+\varphi)} \begin{Bmatrix} J & \Psi_J & J'' \\ \Lambda & \Psi_\Lambda & \Lambda'' \\ -\Phi-1 & \Psi & -\Phi''-1 \end{Bmatrix} \begin{Bmatrix} J & J'' & \Psi_J \\ \Lambda & \Lambda'' & \Psi_\Lambda \\ \Delta & \Delta'' & \Psi \end{Bmatrix} \begin{Bmatrix} \Delta & & \tfrac{1}{2}(\Phi-\Phi''+J'_m) & q \\ \tfrac{1}{2}(\Phi''-\Phi+J'_m) & & \Delta'' & \Psi \end{Bmatrix} C^{J\ J'\ J''}_{M_J\ M'_J\ M''_J} C^{\Lambda\ \Lambda'\ \Lambda''}_{M_\Lambda\ M'_\Lambda\ M''_\Lambda}. \quad (2.27)$$

We note also the degenerate case

$$\left\langle \begin{matrix} J''_m & \Lambda''_m \\ (\max) & \end{matrix} \middle| \left\langle \begin{matrix} J'_m & 0 \\ J' & \Lambda' \\ M'_J & M'_\Lambda \end{matrix} \middle| \begin{matrix} J_m & \Lambda_m \\ J & \Lambda \\ M_J & M_\Lambda \end{matrix} \right\rangle \right\rangle = (-1)^{\Delta''+\frac{1}{2}(\Phi''-\Phi+J'_m)-q+\Lambda''_m+\Lambda'-\Lambda} \left(\frac{(2J+1)(2\Lambda+1)(\Phi+J'_m-\Phi'')!}{\mathfrak{N}(N;J''_m\Lambda''_m)(\Phi''-\Delta'')} \right)^{1/2}$$

$$\times \left(\frac{(\varphi+\Delta)!(\varphi-\Delta-1)!(\Phi-\Phi''+J'_m+1)!(\Phi''+\varphi+J'_m+1)!(2\Phi+1)(J''_m-J+J')!}{(\Phi''-\Phi+J'_m)!(\varphi-\Phi''+J'_m)!(\Phi+\varphi)!} \right.$$

$$\left. \times \frac{(J''_m+J-J')!(\Lambda''_m-\Lambda+\Lambda')!(\Lambda''_m+\Lambda-\Lambda')!}{(\Phi-\varphi)!} \right)^{1/2} [(2q+1)[\tfrac{1}{2}(\Phi+\Phi''-J'_m)-1-q]![\tfrac{1}{2}(\Phi+\Phi''-J'_m)+q]!]^{1/2}$$

$$\times \sum_{\Psi_1 \Psi_2} \frac{(-1)^{2\Psi_1}[(2\Psi_1+1)]^{1/2}}{[\tfrac{1}{2}(\Phi''+\varphi-J'_m)-1-\Psi_1]![\tfrac{1}{2}(\Phi''+\varphi-J'_m)+\Psi_1]!} C^{\frac{1}{2}(\varphi-\Phi''+J'_m)\ \Delta\ \Psi_1}_{\frac{1}{2}(\Delta''-\delta-\delta')\ \delta\ \frac{1}{2}(\Delta''+\delta-\delta')} C^{\Psi_1\ \Phi\ \Psi_2}_{\frac{1}{2}(\Delta''+\delta-\delta')\ -\delta\ \frac{1}{2}(\Delta''-\delta-\delta')}$$

$$\times C^{\Delta''\ \frac{1}{2}(\Phi''+\varphi+J'_m)\ \Psi_2}_{\Delta''\ -\frac{1}{2}(\Delta''+\delta+\delta')\ \frac{1}{2}(\Delta''-\delta-\delta')} \begin{Bmatrix} \Delta & \Psi_1 & \frac{1}{2}(\varphi-\Phi''+J'_m) \\ \frac{1}{2}(\Phi-\varphi) & \frac{1}{2}(\Phi-\Phi''+J'_m) & q \end{Bmatrix} \begin{Bmatrix} \Psi_2 & \Psi & \Phi \\ \frac{1}{2}(\Phi-\varphi) & \frac{1}{2}(\Phi+\varphi) & q \end{Bmatrix} \begin{Bmatrix} \frac{1}{2}(\Phi''+\varphi+J'_m) & \frac{1}{2}(\Phi''-\Phi+J'_m) & \frac{1}{2}(\Phi+\varphi) \\ q & \Psi_2 & \Delta'' \end{Bmatrix}$$

$$\times C^{J\ J'\ J''_m}_{M_J\ M'_J\ J''_m} C^{\Lambda\ \Lambda'\ \Lambda''_m}_{M_\Lambda\ M'_\Lambda\ \Lambda''_m} \quad (2.28)$$

It is easily determined that our multiplicity parameter q has the correct dimensionality to label the multiplicity space of the totally symmetric tensor operator. This dimensionality, as prescribed by the theorem on the isomorphism of the state-labeling problem and the multiplicity-labeling problem,[13] is given by

$$J'_m - |J''_m - J_m| - |\Lambda''_m - \Lambda_m| + 1. \quad (2.29)$$

It should be noted that our factorization method allows us to determine the matrix element (2.27) only up to an orthogonal transformation in multiplicity space. The particular representation which we have chosen has the property that in the degenerate case (2.8) the reduced matrix element becomes a monomial. The theorem on the dimensionality of multiplicity space given in Ref. (13) is proved for tensor operators in $U(n)$ alone; the proof, however, also holds for $Sp(4)$, since the only hypothesis of the proof is the existence of a cyclic vector in the carrier space of an irreducible representation.

3. FACTORIZATION OF THE GENERAL $Sp(4) * Sp(4)$ MATRIX ELEMENT

We now wish to extract the most general Wigner coefficient of $Sp(4)$ by factorization of the $Sp(4)*Sp(4)$ matrix element of the operator

$$B\begin{pmatrix} m'_j & m'_\lambda \\ j' & \lambda' \\ J'_m & \Lambda'_m \\ M'_J & M'_\Lambda \end{pmatrix} = \left(\frac{(\Phi'+\Delta')!(\Phi'-\Delta'-1)!}{(2\Delta'+1)} \right)^{1/2} \sum_{\substack{J'_1+\Lambda'_1=j' \\ J'_2+\Lambda'_2=\lambda'}}$$

$$(-1)^{J'_1+J'_2-J} C^{\frac{1}{2}(\Phi'-\varphi')\ \frac{1}{2}(\Phi'+\varphi')\ \Phi'}_{\frac{1}{2}(\Lambda'_1+\Lambda'_2-J'_1-J'_2+J'-\Lambda')\ \frac{1}{2}(J'_1+J'_2-\Lambda'_1-\Lambda'_2+J'-\Lambda')\ J'-\Lambda'}$$

$$\times C^{J'_1\ \Lambda'_1\ j'}_{J'_2\ \Lambda'_2\ \lambda'} \frac{1}{[(2J'_1)!(2J'_2)!(2\Lambda'_1)!(2\Lambda'_2)!]^{1/2}}$$

$$\sum_{\substack{m'_1 m'_2 \\ \mu'_1 \mu'_2}} \sum_{\substack{n'_1 n'_2 \\ \nu'_1 \nu'_2}} C^{J'_1\ J'_2\ J'}_{m'_1\ m'_2\ M'_J} (-1)^{\Lambda'-M'_\Lambda} C^{\Lambda'_1\ \Lambda'_2\ \Lambda'}_{\mu'_1\ \mu'_2\ M'_\Lambda}$$

$$\times C^{J'_1\ \Lambda'_1\ j'}_{n'_1\ n'_2\ m'_j} (-1)^{\lambda'-m'_\lambda} C^{J'_2\ \Lambda'_2\ \lambda'}_{\nu'_1\ \nu'_2\ m'_\lambda} d^{J'_1}_{m'_1 n'_1}(a^{14}_{14})$$

$$d^{\Lambda'_1}_{\mu'_1 n'_2}(a^{14}_{23}) d^{J'_2}_{m'_2 \nu'_1}(a^{23}_{14})\, d^{\Lambda'_2}_{\mu'_2 \nu'_2}(a^{23}_{23}), \quad (3.1)$$

where we have imposed the restriction $j' + \lambda' = J'_m + \Lambda'_m = \Phi' - 1$ in upper pattern space. We wish to take the matrix element of this operator between the states (2.10) and (2.15). We find, as in case (2.16), that this matrix element contains four (9-j) symbols, but now the first two contain no degeneracies, while the last two are triply stretched as before. We treat the first two by means of the identity[9]

$$\begin{Bmatrix} j_1 & j_3 & j_{13} \\ j_2 & j_4 & j_{24} \\ j_{12} & j_{34} & j \end{Bmatrix} = (-1)^{j_{13}+j_{24}+j} \left(\frac{(j_1+j_3-j_{13})!(j_1+j_3+j_{13}+1)!(j_2+j_4-j_{24})!(j_2+j_4+j_{24}+1)!}{(2j_{13}+1)(2j_{24}+1)(2j+1)} \right)^{1/2}$$

$$\times [(j_{12}+j_{34}-j)!(j_{12}+j_{34}+j+1)!]^{1/2} \Delta(j_1 j_2 j_{12}) \Delta(j_3 j_4 j_{34}) \sum_{\Psi_1 \Psi_2 \Psi_3} (2\Psi_1+1)(2\Psi_2+1)(2\Psi_3+1)$$

$$\times \{[\tfrac{1}{2}(-j_1-j_3+j_2+j_4+j_{12}+j_{34})-\Psi_1]![\tfrac{1}{2}(-j_1-j_3+j_2+j_4+j_{12}+j_{34})+\Psi_1+1]!$$

$$\times [\tfrac{1}{2}(j_1+j_3-j_2-j_4+j_{12}+j_{34})-\Psi_2]![\tfrac{1}{2}(j_1+j_3-j_2-j_4+j_{12}+j_{34})+\Psi_2+1]!$$

$$\times [\tfrac{1}{2}(j_1+j_3+j_2+j_4-j_{12}-j_{34})-\Psi_3]![\tfrac{1}{2}(j_1+j_3+j_2+j_4-j_{12}-j_{34})+\Psi_3+1]!\}^{-1}$$

$$\times \begin{Bmatrix} j_{13} & j_{24} & j \\ \Psi_1 & \Psi_2 & \Psi_3 \end{Bmatrix} C^{\Psi_3}_{\frac{1}{2}(j_1-j_3+j_2-j_4-j_{12}+j_{34})\ \frac{1}{2}(j_1-j_3-j_2+j_4+j_{12}-j_{34})\ j_1-j_3} C^{\Psi_2}_{\frac{1}{2}(-j_1+j_3+j_2-j_4+j_{12}-j_{34})\ \frac{1}{2}(j_1-j_3+j_2-j_4-j_{12}+j_{34})\ j_2-j_4}$$

$$\times C^{\Psi_1}_{\frac{1}{2}(j_1-j_3-j_2+j_4+j_{12}-j_{34})\ \frac{1}{2}(-j_1+j_3+j_2-j_4+j_{12}-j_{34})\ j_{12}-j_{34}}, \tag{3.2}$$

in terms of which the couplings become straightforward. We obtain the factorization

$$\frac{1}{\mathfrak{M}(N;J''_m\Lambda''_m)} \left(\frac{N!(N+2\Phi''+5)!(\Phi''-\Delta''-1)!(\Phi''+\Delta'')!(\Phi-\Delta-1)!(\Phi+\Delta)!}{(2\Delta''+1)(2\Delta+1)(2\Phi''+5)!} \right)^{1/2}$$

$$\times \left\langle \begin{pmatrix} (\max) \\ J''_m\ \Lambda''_m \\ J''\ \Lambda'' \end{pmatrix} \middle| B \begin{pmatrix} J''_m & -J_m & \Lambda''_m - \Lambda_m \\ j' & & \lambda' \\ J'_m & & \Lambda'_m \\ J' & & \Lambda' \\ J''-J & & \Lambda''-\Lambda \end{pmatrix} \middle| \begin{pmatrix} (\max) \\ J_m\ \Lambda_m \\ J\ \Lambda \end{pmatrix} \right\rangle = C^{J_m\ j'\ J''_m}_{J_m\ J''_m-J_m\ J''_m} C^{\Lambda_m\ \lambda'\ \Lambda''_m}_{\Lambda_m\ \Lambda''_m-\Lambda_m\ \Lambda''_m} C^{J\ J'\ J''}_{J\ J''-J\ J''} C^{\Lambda\ \Lambda'\ \Lambda''}_{\Lambda\ \Lambda''-\Lambda\ \Lambda''}$$

$$\times \sum_{q_1 q_2 q_3} F_{q_1 q_2 q_3} \begin{pmatrix} J''_m & \Lambda''_m; & J'_m & \Lambda'_m; & J_m & \Lambda_m \\ J_m & \Lambda_m; & j' & \lambda'; & J_m & \Lambda_m \end{pmatrix} F_{q_1 q_2 q_3} \begin{pmatrix} J''_m & \Lambda''_m; & J'_m & \Lambda'_m; & J_m & \Lambda_m \\ J'' & \Lambda''; & J'_m & \Lambda'_m; & J & \Lambda \end{pmatrix} = C^{J_m\ j'\ J''_m}_{J_m\ J''_m-J_m\ J''_m} C^{\Lambda_m\ \lambda'\ \Lambda''_m}_{\Lambda_m\ \Lambda''_m-\Lambda_m\ \Lambda''_m}$$

$$\times C^{J\ J'\ J''}_{J\ J''-J\ J''} C^{\Lambda\ \Lambda'\ \Lambda''}_{\Lambda\ \Lambda''-\Lambda\ \Lambda''} \sum_{q_J q_\Lambda} \left\langle \begin{matrix} J''_m & \Lambda''_m \\ J''_m & \Lambda''_m \end{matrix} \middle| \begin{bmatrix} J''_m & -J_m & \Lambda''_m & -\Lambda_m \\ q_J & & q_\Lambda & \\ J'_m & & \Lambda'_m & \\ j' & & \lambda' & \\ J''_m & -J_m & \Lambda''_m & -\Lambda_m \end{bmatrix} \middle| \begin{matrix} J_m & \Lambda_m \\ J_m & \Lambda_m \end{matrix} \right\rangle \left\langle \begin{matrix} J''_m & \Lambda''_m \\ J'' & \Lambda'' \end{matrix} \middle| \begin{bmatrix} J''_m & J_m & \Lambda''_m & -\Lambda_m \\ q_J & & q_\Lambda & \\ J'_m & & \Lambda'_m & \\ J' & & \Lambda' & \\ J''-J & & \Lambda''-\Lambda & \end{bmatrix} \middle| \begin{matrix} J_m & \Lambda_m \\ J & \Lambda \end{matrix} \right\rangle \tag{3.3}$$

where

$$F_{q_1 q_2 q_3} \begin{pmatrix} J''_m & \Lambda''_m; & J'_m & \Lambda'_m; & J_m & \Lambda_m \\ J'' & \Lambda''; & J'_m & \Lambda'_m; & J & \Lambda \end{pmatrix} = (-1)^{\Lambda''_m-\Lambda_m+\Lambda+\Lambda''} \left(\frac{(\Phi+\Phi'-\Phi''-1)!}{\mathfrak{M}(N;J''_m\Lambda''_m)(2\Phi)!(2\Phi')!(2\Phi'')!} \right)^{1/2}$$

$$\times [(2J+1)(2\Lambda+1)(2J'+1)(2\Lambda'+1)(2q_1+1)(2q_2+1)(2q_3+1)(\Phi+\varphi)!(\Phi-\varphi)!(\Phi+\delta)!(\Phi-\delta)!(\Phi'+\varphi')!]^{1/2}$$

$$\times [(\Phi'-\varphi')!(\Phi'+\delta')!(\Phi'-\delta')!(\Phi''+\varphi'')!(\Phi''-\varphi'')!(\Phi''+\delta'')!(\Phi''-\delta'')![\tfrac{1}{2}(\Phi'+\Phi''-\Phi-1)-q_1]!]^{1/2}$$

$$\times \{[\tfrac{1}{2}(\Phi'+\Phi''-\Phi+1)+q_1]![\tfrac{1}{2}(\Phi''+\Phi-\Phi'-1)-q_2]![\tfrac{1}{2}(\Phi''+\Phi-\Phi'+1)+q_2]![\tfrac{1}{2}(\Phi+\Phi'-\Phi''-1)-q_3]!\}^{1/2}$$

$$\times \{[\tfrac{1}{2}(\Phi+\Phi'-\Phi''+1)+q_3]!\}^{1/2} (\Phi+\Phi'-\Phi''+1)!(\Phi'+\Phi''-\Phi+1)!(\Phi''+\Phi-\Phi'+1)! \sum_{\Psi_1 \cdots \Psi_6} (-1)^{2\Psi_1+2\Psi_2+2\Psi_3}$$

$$\times (2\Psi_1+1)(2\Psi_2+1)(2\Psi_3+1)(2\Psi_4+1)(2\Psi_5+1)(2\Psi_6+1) \begin{Bmatrix} J & J' & J'' \\ \Psi_1 & \Psi_2 & \Psi_3 \end{Bmatrix} \begin{Bmatrix} \Lambda & \Lambda' & \Lambda'' \\ \Psi_4 & \Psi_5 & \Psi_6 \end{Bmatrix}$$

$$\times \begin{Bmatrix} \Psi_3 & \Psi_2 & J \\ \Psi_6 & \Psi_5 & \Lambda \\ q_3 & q_2 & \Delta \end{Bmatrix} \begin{Bmatrix} \Psi_1 & \Psi_3 & J' \\ \Psi_4 & \Psi_6 & \Lambda' \\ q_1 & q_3 & \Delta' \end{Bmatrix} \begin{Bmatrix} \Psi_2 & \Psi_1 & J'' \\ \Psi_5 & \Psi_4 & \Lambda'' \\ q_2 & q_1 & \Delta'' \end{Bmatrix} \frac{1}{[\frac{1}{2}(\Phi+\Phi'-\Phi''-1)-\Psi_3-\Psi_6]!}$$

$$\times \frac{1}{[\frac{1}{2}(\Phi+\Phi'-\Phi''-1)-\Psi_3+\Psi_6+1]![\frac{1}{2}(\Phi+\Phi'-\Phi''-1)+\Psi_3-\Psi_6+1]![\frac{1}{2}(\Phi+\Phi'-\Phi''-1)+\Psi_3+\Psi_6+2]!}$$

$$\times \frac{1}{[\frac{1}{2}(\Phi'+\Phi''-\Phi-1)-\Psi_1-\Psi_4]![\frac{1}{2}(\Phi'+\Phi''-\Phi-1)-\Psi_1+\Psi_4+1]![\frac{1}{2}(\Phi'+\Phi''-\Phi-1)+\Psi_1-\Psi_4+1]!}$$

$$\times \frac{1}{[\frac{1}{2}(\Phi'+\Phi''-\Phi-1)+\Psi_1+\Psi_4+2]![\frac{1}{2}(\Phi''+\Phi-\Phi'-1)-\Psi_2-\Psi_5]![\frac{1}{2}(\Phi''+\Phi-\Phi'-1)-\Psi_2+\Psi_5+1]!}$$

$$\times \frac{1}{[\frac{1}{2}(\Phi''+\Phi-\Phi'-1)+\Psi_2-\Psi_5+1]![\frac{1}{2}(\Phi''+\Phi-\Phi'-1)+\Psi_2+\Psi_5+2]!}. \tag{3.4}$$

The F coefficient (3.4), however, cannot be identified with the matrix element of a reduced canonical unit tensor operator in $Sp(4)$ because the dimensionality of the parameters q_1, q_2, q_3 is greater than the dimensionality of the multiplicity space as prescribed by the theorem of Baird and Biedenharn.[13] Hence the F coefficients (3.4) do not constitute an orthonormal set in the multiplicity space, and we must follow a Schmidt orthogonalization procedure in order to extract the canonical Wigner coefficients. The $Sp(4)*Sp(4)$ matrix element treats the three representations being coupled symmetrically, whereas the canonical Wigner coefficient singles out one of the representations of $Sp(4)$ as the carrier of the multiplicity labels; hence the coefficients (3.4) possess a higher degree of symmetry than the canonical Wigner coefficients, and their usefulness is not exhausted when we have derived from them a general canonical Wigner coefficient by the Schmidt process. The F coefficients still prescribe the symmetries by means of which one set of canonical Wigner coefficients is related to another, i.e., how a Wigner coefficient may be expressed in terms of others which select a different representation as the carrier of the multiplicity labels. In this connection we note that the labels q_1, q_2, q_3 satisfy the inequalities

$$\begin{aligned} &\Phi - 1 \geq q_2 + q_3 \geq \Delta \geq |q_2 - q_3|, \\ &\Phi' - 1 \geq q_1 + q_3 \geq \Delta' \geq |q_1 - q_3|, \\ &\Phi'' - 1 \geq q_1 + q_2 \geq \Delta'' \geq |q_1 - q_2|, \end{aligned} \tag{3.5}$$

which the two canonical multiplicity parameters must satisfy when the corresponding representation is selected as the carrier of the multiplicity labels. However, (3.4) does become the canonical Wigner coefficient when one of the three representations is totally symmetric or in the case that the coupling is multiplicity-free, e.g., when one of the quantities

$$(\Phi + \Phi' - \Phi'' - 1), (\Phi' + \Phi'' - \Phi - 1), (\Phi'' + \Phi - \Phi' - 1) \tag{3.6}$$

is equal to zero. It is not yet known whether such "symmetrical" parametrization or a canonical parametrization emerges from the direct factorization of the $U(n)*U(n)$ matrix element, but we may speculate that for $n \geq 4$, at least, the symmetries of the $U(n)*U(n)$ matrix element will yield, under factorization, F coefficients whose multiplicity space is "symmetrically" parametrized, and from which canonical Wigner coefficients must be obtained by the Schmidt process. The $U(3)*U(3)$ case is special in that only a single multiplicity parameter is involved, its maximal compact subgroup is multiplicity-free, and an operator algorithm is available[6] for the extraction of the Wigner coefficients by induction. This remains speculation since as yet no direct factorization of $U(n)*U(n)$ matrix elements has been performed for $n \geq 3$.

ACKNOWLEDGMENT

The author would like to thank Dr. Wesley Brittin for his kind hospitality at the University of Colorado.

---

*Present address: Physics Department, University of North Carolina, Chapel Hill, N.C., 27514.

[1]K. T. Hecht, Nucl. Phys. **63**, 177 (1965).
[2]K. T. Hecht and S. C. Pang, J. Math. Phys. **10**, 1571 (1969).
[3]M. K. F. Wong, J. Math. Phys. **12**, 1530 (1971).
[4]S. J. Ališauskas and A. P. Jucys, J. Math. Phys. **10**, 2227 (1969).
[5]S. J. Ališauskas and A. P. Jucys, J. Math. Phys. **12**, 594 (1971).
[6]L. C. Biedenharn, A. Giovannini, and J. D. Louck, J. Math. Phys. **8**, 691 (1967).
[7]L. C. Biedenharn and J. D. Louck, Commun. Math. Phys. **8**, 89 (1968).
[8]W. J. Holman III, Nuovo Cimento A **4**, 904 (1971).
[9]W. J. Homan III, J. Math. Phys. **13**, ,(1972).
[10]W. J. Homan III, J. Math. Phys. **10**, 1710 (1969).
[11]R. T. Sharp, Nucl. Phys. **95**, 222 (1967).
[12]A. P. Jucys, A. A. Bandzaitis, *Teoriya Momenta Kolichestva Dvizheniya v Kvantovoi Mekhanike* (Vilnius, Lithuanian S.S.R., 1965), p. 170, Eq. (25.5).
[13]G. E. Baird and L. C. Biedenharn, J. Math. Phys. **5**, 1730 (1964).

The representation theory of $SU(4)$ and $Sp(4)$

Wayne J. Holman III*

Department of Physics and Astrophysics, University of Colorado, Boulder, Colorado 80302
(Received 21 August 1972)

Basis states of irreducible representations of $SU(4)$ in its $Sp(4)$ basis are explicitly constructed in the boson calculus by the method of lowering operators. Expressions are derived for the Wigner coefficients of internal coupling and for the most general matrix which transforms an irreducible representation of $SU(4)$ from its $Sp(4)$ basis to a $U(3)$ basis.

1. INTRODUCTION

The representation theory of $SU(4)$ in the boson calculus has been studied by Ciftan for the $U(3)$ basis[1,2] from the standpoint of combinatorial analysis, and by the present author from that of the internal coupling structure of the general basis state.[3] It remains, then, to examine the boson states of $SU(4)$ in its $Sp(4) \supset U(2) \times U(2)$ basis; the more limited problem of the $SU(4) \supset U(2) \times U(2)$ reduction has been investigated by Hecht and Pang[4] and Draayer.[5] In the present work we shall use the general lowering operators for the orthogonal groups constructed by Pang and Hecht,[6] apply them to the state of maximal weight of a general irreducible representation of $SU(4)$, and so determine the general state of this representation in the boson calculus. This induction is carried out in Sec. 2, and in Sec. 3 we shall use the result to determine the $SU(4)$ Wigner coefficient of the internal coupling and the matrix which transforms an irreducible representation of $SU(4)$ from its $U(3)$ basis to its $Sp(4) \supset U(2) \times U(2)$ basis.

2. CONSTRUCTION OF THE GENERAL BASIS STATE

The generators of $SU(4)$, locally isomorphic to $SO(6)$, are given in the table. The generators E_{ij} have the commutation relations

$$[E_{ij}, E_{k\ell}] = \delta_{jk} E_{i\ell} - \delta_{i\ell} E_{kj} \tag{2.1}$$

and are represented by 4×4 matrices with unity in the (ij) place and zeros elsewhere. The $SO(6)$ generators L_{ij} have the commutation relations

$$[L_{ij}, L_{k\ell}] = i\delta_{ik} L_{j\ell} - i\delta_{jk} L_{i\ell} + i\delta_{j\ell} L_{ik} - i\delta_{i\ell} L_{jk} \tag{2.2}$$

and are given opposite the $SU(4)$ generators to which they are isomorphic.

TABLE OF GENERATORS.

$\frac{1}{2}(E_{11} - E_{44})$	$\frac{1}{2}(L_{12} + L_{34})$
E_{14}	$\frac{1}{2}[(L_{23} + L_{14}) + i(L_{31} + L_{24})]$
E_{41}	$\frac{1}{2}[(L_{23} + L_{14}) - i(L_{31} + L_{24})]$
$\frac{1}{2}(E_{22} - E_{33})$	$\frac{1}{2}(L_{12} - L_{34})$
E_{23}	$\frac{1}{2}[(L_{14} - L_{23}) + i(L_{24} - L_{31})]$
E_{32}	$\frac{1}{2}[(L_{14} - L_{23}) - i(L_{24} - L_{31})]$
$\frac{1}{2}(E_{13} + E_{24})$	$\frac{1}{2}(L_{52} + iL_{15})$
$\frac{1}{2}(E_{31} + E_{24})$	$\frac{1}{2}(L_{52} - iL_{15})$
$\frac{1}{2}(E_{12} - E_{34})$	$\frac{1}{2}(L_{45} + iL_{53})$
$\frac{1}{2}(E_{21} - E_{43})$	$\frac{1}{2}(L_{45} - iL_{53})$
$\frac{1}{2}(E_{11} - E_{22} - E_{33} + E_{44})$	L_{56}
$\frac{1}{2}(E_{13} - E_{24})$	$\frac{1}{2}(L_{61} + iL_{62})$
$\frac{1}{2}(E_{31} - E_{42})$	$\frac{1}{2}(L_{61} - iL_{62})$
$\frac{1}{2}(E_{12} + E_{34})$	$\frac{1}{2}(L_{36} + iL_{46})$
$\frac{1}{2}(E_{21} + E_{43})$	$\frac{1}{2}(L_{36} - iL_{46})$

The first six of these generators are those of the $SO(4)$ subgroup of $SO(6)$ [the $SU(2) \times SU(2)$ subgroup of $SU(4)$] and the first ten are those of the $SO(5)$ subgroup of $SO(6)$ [the $Sp(4)$ subgroup of $SU(4)$].

The most general irreducible representation of $SU(4)$ is described by three positive integers which represent the number of nodes in the three rows of the corresponding Young frame $[m_{14}m_{24}m_{34}]$, and a state of this representation in the canonical decomposition $SU(4) \supset U(3) \supset U(2) \supset U(1)$ is given by the Gel'fand pattern

$$\left| \begin{pmatrix} m_{14} & & m_{24} & & m_{34} & & 0 \\ & m_{13} & & m_{23} & & m_{33} & \\ & & m_{12} & & m_{22} & & \\ & & & m_{11} & & & \end{pmatrix} \right|. \tag{2.3a}$$

An irreducible representation of $SO(6) \supset SO(5) \supset SO(4) \supset SO(3) \supset SO(2)$ is labeled by the Gel'fand pattern

$$\left| \begin{pmatrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51} & & m_{52} & \\ & m_{41} & & m_{42} & \\ & & m_{31} & & \\ & & m_{21} & & \end{pmatrix} \right|, \tag{2.3b}$$

with the conditions

$$\begin{gathered} m_{61} \geq m_{51} \geq m_{62} \geq m_{52} \geq |m_{63}| \\ m_{51} \geq m_{41} \geq m_{52} \geq |m_{42}| \\ m_{41} \geq m_{31} \geq |m_{42}| \\ m_{31} \geq |m_{21}|. \end{gathered} \tag{2.3'}$$

The relations between the $SU(4)$ and $SO(6)$ invariant labels are given by

$$\begin{aligned} m_{61} &= \tfrac{1}{2}(m_{14} + m_{24} - m_{34}), \\ m_{62} &= \tfrac{1}{2}(m_{14} - m_{24} + m_{34}), \\ m_{63} &= \tfrac{1}{2}(m_{14} - m_{24} - m_{34}). \end{aligned} \tag{2.4}$$

The state of maximal weight of an irreducible representation of $SU(4)$ is given in the boson calculus as

$$\left| \begin{pmatrix} m_{14} & & m_{24} & & m_{34} & & 0 \\ & m_{14} & & m_{24} & & m_{34} & \\ & & m_{14} & & m_{24} & & \\ & & & m_{14} & & & \end{pmatrix} \right|$$

$$= \mathfrak{N}_{\max}^{-1/2} (a_{123}^{123})^{m_{34}} (a_{12}^{12})^{m_{24} - m_{34}} (a_1^1)^{m_{14} - m_{24}} |0\rangle, \tag{2.5}$$

where

$$a_{j_1 \cdots j_k}^{i_1 \cdots i_k} = \sum_p \epsilon(j_{l_1} \cdots j_{l_k}) a_{j_{l_1}}^{i_1} \cdots a_{j_{l_k}}^{i_k}, \tag{2.5'}$$

the sum being taken over all permutations of the indices (j_ℓ), and $\epsilon((j_\ell))$ denoting the signature of the permutation. We shall abbreviate $a^{123}_{ijk} = a_{ijk}$ and $a^1_i = a_i$. The bosons obey the commutation relations

$$[\bar{a}^i_j, \bar{a}^{i'}_{j'}] = [a^i_j, a^{i'}_{j'}] = 0,$$
$$[\bar{a}^i_j, a^{i'}_{j'}] = \delta_{ii'}\delta_{jj'},$$

and the condition

$$\bar{a}^i_j|0\rangle = \langle 0|a^i_j = 0.$$

The generators are realized by the mapping

$$E_{ij} \to \sum_{p=1}^{4} a^p_i \bar{a}^p_j. \tag{2.6}$$

The normalization constant for the maximal state is given by

$$\mathfrak{N}_{\max} = \frac{(m_{14}+2)!(m_{24}+1)!m_{34}!}{(m_{14}-m_{34}+2)(m_{14}-m_{24}+1)(m_{24}-m_{34}+1)}. \tag{2.7}$$

We now wish to obtain an expression for the general state of such a representation in the $Sp(4) \supset U(2) \times U(2)$ basis by means of the application of suitable lowering operators to the maximal state (2.5). I.e., we wish to find operators $\mathfrak{L}^5_i$ with the property that

$$\begin{aligned}
\mathfrak{L}^5_1 \left|\begin{pmatrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51} & & m_{52} & \end{pmatrix}\right| &= c_1 \left|\begin{pmatrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51}-1 & & m_{52} & \end{pmatrix}\right|, \\
\mathfrak{L}^5_2 \left|\begin{pmatrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51} & & m_{52} & \end{pmatrix}\right| &= c_2 \left|\begin{pmatrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51} & & m_{52}-1 & \end{pmatrix}\right|,
\end{aligned} \tag{2.8}$$

where the states indicated in (2.8) are all maximal in their $Sp(4)$ subgroups. Such operators have been found by Pang and Hecht.[6] In terms of the $SU(4)$ generators we may write them

$$\begin{aligned}
\mathfrak{L}^5_1 = &-(E_{31}-E_{42})(E_{22}-E_{33}+1)(E_{11}+E_{22}-E_{33}-E_{44}+2)(E_{11}-E_{44}+2)(E_{11}+E_{22}-E_{33}-E_{44}+3) \\
&+E_{32}(E_{21}+E_{43})(E_{11}+E_{22}-E_{33}-E_{44}+2)(E_{11}-E_{44}+2)(E_{11}+E_{22}-E_{33}-E_{44}+3) \\
&+(E_{31}+E_{42})(E_{11}-E_{22}-E_{33}+E_{44})(E_{22}-E_{33}+1)(E_{11}-E_{44}+2)(E_{11}+E_{22}-E_{33}-E_{44}+3) \\
&-E_{32}(E_{21}-E_{43})(E_{11}-E_{22}-E_{33}+E_{44})(E_{11}-E_{44}+2)(E_{11}+E_{22}-E_{33}-E_{44}+3) \\
&-E_{41}(E_{12}+E_{34})(E_{22}-E_{33}+1)(E_{11}+E_{22}-E_{33}-E_{44}+2)(E_{11}+E_{22}-E_{33}-E_{44}+3) \\
&+(E_{31}+E_{42})(E_{21}-E_{43})(E_{12}+E_{34})(E_{22}-E_{33}+1)(E_{11}+E_{22}-E_{33}-E_{44}+3) \\
&-E_{32}(E_{21}-E_{43})(E_{21}-E_{43})(E_{12}+E_{34})(E_{11}+E_{22}-E_{33}-E_{44}+3) \\
&-E_{32}E_{41}(E_{13}-E_{24})(E_{11}+E_{22}-E_{33}-E_{44}+2)(E_{11}-E_{44}+2) \\
&+(E_{31}+E_{42})(E_{31}+E_{42})(E_{13}-E_{24})(E_{22}-E_{33}+1)(E_{11}-E_{44}+2) \\
&-E_{32}(E_{21}-E_{43})(E_{31}+E_{42})(E_{13}-E_{24})(E_{11}-E_{44}+2) \\
&-E_{41}E_{32}(E_{13}-E_{24})(E_{22}-E_{33}+1)(E_{11}+E_{22}-E_{33}-E_{44}+2) \\
&+(E_{31}+E_{42})(E_{21}-E_{43})E_{32}(E_{13}-E_{24})(E_{22}-E_{33}+1) \\
&-E_{32}(E_{21}-E_{43})(E_{21}-E_{43})E_{32}(E_{13}-E_{24}),
\end{aligned}$$

$$\begin{aligned}
\mathfrak{L}^5_2 = &(E_{21}+E_{43})(E_{11}-E_{22}+E_{33}-E_{44})(E_{11}-E_{22}+E_{33}-E_{44}+1)(E_{11}-E_{44}+2) \\
&-(E_{21}-E_{43})(E_{11}-E_{22}-E_{33}+E_{44})(E_{11}-E_{22}+E_{33}-E_{44}+1)(E_{11}-E_{44}+2) \\
&-(E_{21}-E_{43})(E_{21}-E_{43})(E_{12}+E_{34})(E_{11}-E_{44}+2) \\
&-E_{41}(E_{13}-E_{24})(E_{11}-E_{22}+E_{33}-E_{44})(E_{11}-E_{22}+E_{33}-E_{44}+1) \\
&-(E_{21}-E_{43})(E_{31}+E_{42})(E_{13}-E_{24})(E_{11}-E_{22}+E_{44}+1) \\
&-(E_{21}-E_{43})(E_{21}-E_{43})E_{32}(E_{13}-E_{24}).
\end{aligned} \tag{2.9}$$

Applying these operators to the maximal state (2.5), we obtain the normalized semimaximal state

$$\left|\begin{pmatrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51} & & m_{52} & \\ & m_{51} & & m_{52} & \\ & & m_{51} & & \\ & & m_{51} & & \end{pmatrix}\right\rangle$$

$$\begin{aligned}
&= \left(\frac{(m_{61}+m_{63}+2)(m_{62}+m_{63}+1)(m_{61}-m_{62}+1)}{(m_{61}+m_{62}+2)!(m_{61}-m_{63}+1)!(m_{62}-m_{63})!}\right. \\
&\times \frac{(2m_{51}+3)!(m_{61}+m_{62}+2)!}{(m_{61}+m_{51}+3)!(m_{51}+m_{62}+2)!}\Bigg)^{1/2} \\
&\times \left(\frac{(m_{61}-m_{63}+1)!(m_{61}+m_{63}+1)!(m_{51}-m_{52}+1)!}{(m_{51}-m_{63}+1)!(m_{61}-m_{51})!(m_{51}+m_{63}+1)!}\right. \\
&\times \frac{(m_{51}+m_{52}+2)!(m_{61}-m_{62})!(2m_{52}+1)!(m_{62}-m_{63})!}{(m_{61}-m_{52}+1)!(m_{61}+m_{52}+2)!(m_{51}-m_{62})!} \\
&\times \frac{(m_{62}+m_{63})!}{(m_{62}+m_{52}+1)!}\Bigg)^{1/2} \\
&\times \left(\frac{1}{(m_{52}-m_{63})!(m_{52}+m_{63})!(m_{62}-m_{52})!}\right)^{1/2} \\
&\times (a_{123})^{m_{52}-m_{63}}(a_{124}a_1+a_{123}a_2)^{m_{62}-m_{52}}(a^{12}_{12})^{m_{51}-m_{62}} \\
&\times (a^{12}_{14}+a^{12}_{23})^{m_{61}-m_{51}}(a_1)^{m_{52}+m_{63}}|0\rangle.
\end{aligned} \tag{2.10}$$

We note that application of the operator $\mathfrak{L}_1^5$ merely removes one power of a_{12}^{12} and replaces it with $(a_{14}^{12} + a_{23}^{12})$, the invariant of the $Sp(4)$ subgroup. The operator $\mathfrak{L}_2^5$, on the other hand, removes one power each of a_{123} and a_1, replacing them with the factor

$$a_{124}a_1 + a_{123}a_2 = a_{12}^{12}(a_{14}^{13} + a_{23}^{13}) - a_{12}^{13}(a_{14}^{12} + a_{23}^{12}). \tag{2.11}$$

We now proceed to apply the lowering operators $\mathfrak{L}_1^4$ and $\mathfrak{L}_2^4$, which have the properties, when operating on a semimaximal state of $Sp(4)$, that

$$\mathfrak{L}_1^4 \left|\left(\begin{matrix} m_{51} & m_{52} \\ m_{41} & m_{42} \end{matrix}\right)\right\rangle = c_1' \left|\left(\begin{matrix} m_{51} & m_{52} \\ m_{41}-1 & m_{42} \end{matrix}\right)\right\rangle ,$$
$$\mathfrak{L}_2^4 \left|\left(\begin{matrix} m_{51} & m_{52} \\ m_{41} & m_{42} \end{matrix}\right)\right\rangle = c_2' \left|\left(\begin{matrix} m_{51} & m_{52} \\ m_{41} & m_{42}-1 \end{matrix}\right)\right\rangle . \tag{2.12}$$

These operators have also been determined by Pang and Hecht[6] and by Hecht[7] in a previous publication. They may be written

$$\mathfrak{L}_1^4 = E_{32}\mathfrak{L}_2^4 + E_{41}(E_{12} - E_{34})(E_{22} - E_{33} + 1)$$
$$- (E_{31} + E_{42})(E_{11} - E_{44} + 1)(E_{22} - E_{33} + 1),$$
$$\mathfrak{L}^4 = E_{41}(E_{13} + E_{24}) + (E_{21} - E_{43})(E_{11} - E_{44} + 1). \tag{2.13}$$

At this point we shall abbreviate

$$\begin{aligned} m_{62} - m_{63} &= A, & m_{61} - m_{51} &= m, \\ m_{61} - m_{62} &= B, & m_{62} - m_{52} &= n, \\ m_{62} + m_{63} &= C, & m_{51} - m_{41} &= m', \\ & & m_{52} - m_{42} &= n'. \end{aligned} \tag{2.14}$$

We may simplify the application of the lowering operators $\mathfrak{L}_i^4$ to the state (2.10) by the following procedure. We write the state (2.10) as

$$\begin{aligned} &\mathfrak{N}^{1/2}(a_{123})^{A-n}[a_{12}^{12}a_I^{13} - a_{12}^{13}a_I^{12}]^n(a_{12}^{12})^{B-m} \\ &\quad\times(a_I^{12})^m(a_1)^{C-n}|0\rangle \\ &= \mathfrak{N}^{1/2}\sum_x \frac{n!(-1)^x}{x!(n-x)!}(a_I^{12})^{m+x}(a_I^{13})^{n-x}(a_{123})^{A-n} \\ &\quad\times(a_{12}^{12})^{B-m+n-x}(a_{12}^{13})^x(a_1)^{C-n}|0\rangle \\ &= \mathfrak{N}^{1/2}\sum_x \frac{n!(B-m+n-x)!(-1)^x}{x!(n-x)!(B-m+n)!}(a_I^{12})^{m+x}(a_I^{13})^{n-x} \\ &\quad\times(E^{32})^x(a_{123})^{A-n}(a_{12}^{12})^{B-m+n}(a_1)^{C-n}|0\rangle, \end{aligned} \tag{2.15}$$

where

$$a_I^{ij} = a_{14}^{ij} + a_{23}^{ij} \tag{2.16}$$

is an $Sp(4)$ invariant. Here, the generator E^{32} is defined by

$$E^{ij} = \sum_p a_p^i \bar{a}_p^j. \tag{2.17}$$

It is a generator of the group of "upper pattern space" in the orbital group $SU(4)*SU(4)$ and commutes with all the generators (2.6) of "lower pattern space." Hence the lowering operators $\mathfrak{L}_i^4$ may be commuted through the powers of the symplectic invariants a_I^{ij} and the operator $(E^{32})^x$ immediately, and we need evaluate only

$$\begin{aligned} &(\mathfrak{L}_1^4)^{m'}(\mathfrak{L}_2^4)^{n'}(a_{123})^{A-n}(a_{12}^{12})^{B-m+n}(a_1)^{C-n}|0\rangle \\ &= (-1)^{m'}\frac{(A+B-m+1)!}{(A+B-m-m'+1)!} \\ &\times\frac{(A+B+C-n-m+1)!}{(A+B+C-n-m-n'-m'+1)!} \\ &\times\frac{(B-m+n)!}{(B-m-m'+n)!}\frac{(C-n)!}{(C-n-n')!} \\ &\times {}_2F_1\left(-n', -A+n \,|\, C-n-n'+1 \,|\, -\frac{a_{124}a_1}{a_{123}a_2}\right)\left(\frac{a_2}{a_1}\right)^{n'} \\ &\times {}_2F_1\left(-m', -B-C+m-1 \,|\, A+B-m-m'+2 \,|\, -\frac{a_{23}^{12}}{a_{14}^{12}}\right) \\ &\times\left(\frac{a_{14}^{12}}{a_{12}^{12}}\right)^{m'}(a_{123})^{A-n}(a_{12}^{12})^{B-m+n}(a_1)^{C-n}|0\rangle. \end{aligned} \tag{2.18}$$

Adding the normalization factors of the three lowering operators we find the general state of $SU(4)$ which is maximal in its $SU(2) \times SU(2)$ subgroup:

$$\left|\left(\begin{matrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51} & & m_{52} & \\ & m_{41} & & m_{42} & \\ & & m_{41} & & \\ & & m_{41} & & \end{matrix}\right)\right\rangle$$
$$= \mathfrak{N}^{1/2}\left(\frac{(A+B+C-m-n-n'+1)!}{n'!m'!(A+B+C-m-n+1)!}\right.$$
$$\times\frac{(A+B+C-m-n-m'+1)!(A+C-2n-n')!}{(A+B+C-m-n-m'-n'+1)!}$$
$$\times\frac{(B-m+n)!}{(A+C-2n)!}\Bigg)^{1/2}\left(\frac{(A+2B+C-2m-m'+2)!}{(B-m-m'+n)!}\right.$$
$$\times\frac{(B-m+n-m'+n')!}{(A+2B+C-2m+2)!(B-m+n+n'+1)!}\Bigg)^{1/2}$$
$$\times\frac{(A+B-m+1)!(C-n)!}{(A+B-m-m'+1)!(C-n-n')!}\frac{n!(-1)^{m'}}{(B-m+n)!}$$
$$\times\sum_x\frac{(-1)^x(B-m+n-x)!}{x!(n-x)!}(a_I^{12})^{m+x}(a_I^{13})^{n-x}(E^{32})^x$$
$$\times {}_2F_1\left(-n', -A+n \,|\, C-n-n'+1 \,|\, -\frac{a_{124}a_1}{a_{123}a_2}\right)\left(\frac{a_2}{a_1}\right)^{n'}$$
$$\times {}_2F_1\left(-m', -B-C+m-1 \,|\, A+B-m-m'+2 \,|\, -\frac{a_{23}^{12}}{a_{14}^{12}}\right)$$
$$\times\left(\frac{a_{14}^{12}}{a_{12}^{12}}\right)^{m'}(a_{123})^{A-n}(a^{12})^{B-m+n}(a_1)^{C-n}|0\rangle, \tag{2.19}$$

where the coefficient $\mathfrak{N}$ is merely the coefficient under the squareroot sign in (2.10). We may easily apply the operator $(E^{32})^x$ according to the formula

$$(E^{32})^x(a_{14}^{12})^\alpha(a_{23}^{12})^\beta(a_{12}^{12})^\gamma|0\rangle = \sum_{tu}\frac{x!}{t!u!(x-t-u)!}$$
$$\times\frac{\alpha!}{(\alpha-u)!}\frac{\beta!}{(\beta-t)!}\frac{\gamma!}{(\gamma-x+t+u)!}$$
$$\times(a_{14}^{12})^{\alpha-u}(a_{14}^{13})^u(a_{23}^{12})^{\beta-t}(a_{23}^{13})^t(a_{12}^{12})^{\gamma-x+t+u}(a_{12}^{13})^{x-t-u}|0\rangle. \tag{2.20}$$

We note that in the first Gauss function in (2.19) we may replace

$$a_{124}a_1 = a_{12}^{12}a_{14}^{13} - a_{12}^{13}a_{14}^{12},$$
$$a_{123}a_2 = a_{12}^{12}a_{23}^{13} - a_{12}^{13}a_{23}^{12}. \tag{2.21}$$

3. COUPLING OF TOTALLY SYMMETRIC REPRESENTATIONS INTO A GENERAL REPRESENTATION

For the purpose of forming matrix elements with the boson states of $SU(4)$ derived in the preceding section we must expand our general states into powers of the individual bosons a_j^i. In doing so we obtain an expansion of the general boson state into totally symmetric representations of $SU(4)$ which are coupled by appropriate Wigner coefficients of $SU(4)$. The first step is that of coupling two such symmetric representations into a third representation whose Young diagram has two rows which are equal in length to those of the two symmetric representations being coupled. This coupling is straightforward, since the resulting state is simply equal to the general $Sp(4)$ state studied in Ref. 8, multiplied by a power of the symplectic invariant a_I^{12}, and normalized. The result is

$$\left|\begin{pmatrix} m_{61} & & m_{62} & & m_{62} \\ & m_{51} & & m_{62} & \\ & m_{41} & & m_{42} & \\ & & m_{41} & & \\ & & m_{41} & & \end{pmatrix}\right\rangle$$

$$= \sum_{\substack{J_1+\Lambda_1+J_2+\Lambda_2=m_{61} \\ J_1-J_2+\Lambda_1-\Lambda_2=m_{62} \\ Z=J_1+J_2-\frac{1}{2}(m_{41}+m_{42})}} \sum_{m_1 m_2 \mu_1 \mu_2} (-1)^z$$

$$\times C^{\frac{1}{2}(m_{61}-m_{41})\quad \frac{1}{2}(m_{61}+m_{41})+1\quad m_{51}+1}_{\frac{1}{2}(m_{61}-m_{41})-z\quad z+m_{42}-\frac{1}{2}(m_{61}-m_{41})\quad m_{42}}$$

$$\times C^{\frac{1}{2}(m_{41}+m_{42})\quad \frac{1}{2}(m_{41}-m_{42})\quad m_{62}}_{J_1-J_2\qquad \Lambda_1-\Lambda_2\qquad m_{62}}$$

$$\times C^{J_1\quad J_2\quad \frac{1}{2}(m_{41}+m_{42})}_{m_1\quad m_2\quad \frac{1}{2}(m_{41}+m_{42})} C^{\Lambda_1\quad \Lambda_2\quad \frac{1}{2}(m_{41}-m_{42})}_{\mu_1\quad \mu_2\quad \frac{1}{2}(m_{41}-m_{42})}$$

$$\times \frac{(a_1^1)^{J_1+m_1}(a_4^1)^{J_1-m_1}(a_1^2)^{J_2+m_2}(a_4^2)^{J_2-m_2}}{[(J_1+m_1)!(J_1-m_1)!(J_2+m_2)!(J_2-m_2)!}$$

$$\times \frac{(a_2^1)^{\Lambda_1+\mu_1}(a_3^1)^{\Lambda_1-\mu_1}(a_2^2)^{\Lambda_2+\mu_2}(a_3^2)^{\Lambda_2-\mu_2}}{(\Lambda_1+\mu_1)!(\Lambda_1-\mu_1)!(\Lambda_2+\mu_2)!(\Lambda_2-\mu_2)!]^{1/2}} |0\rangle. \tag{3.1}$$

We may obtain a general state in the $SU(2) \times SU(2)$ basis by lowering the magnetic quantum numbers of the last two $SU(2)$ Wigner coefficients; the normalized lowering operators are given by

$$\left(\frac{(J+M_J)!}{(2J)!(J-M_J)!}\right)^{1/2}(E_{41})^{J-M_J},$$
$$\left(\frac{(\Lambda+M_\Lambda)!}{(2\Lambda)!(\Lambda-M_\Lambda)!}\right)^{1/2}(-E_{32})^{\Lambda-M_\Lambda}, \tag{3.2}$$

where

$$J = \tfrac{1}{2}(m_{41}+m_{42}), \qquad \Lambda = \tfrac{1}{2}(m_{41}-m_{42}). \tag{3.3}$$

Note that when the magnetic quantum number M_Λ is lowered from its maximal to general values we introduce into the expression (3.1) the phase $(-1)^{\Lambda-M_\Lambda}$. We may then form the Gel'fand state of the $SO(4) \supset SO(3) \supset SO(2)$ chain by multiplying the resultant expression by

$$C^{J\quad \Lambda\quad m_{31}}_{M_J\quad M_\Lambda\quad m_{21}} \tag{3.4}$$

and summing over $M_J + M_\Lambda = m_{21}$.

We shall use the following notation for the representations of $SU(4) \supset Sp(4) \supset SU(2) \times SU(2)$:

$$\left|\left(\begin{matrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51} & & m_{52} & \\ & J & & \Lambda & \\ & M_J & & M_\Lambda & \end{matrix}\right)\right\rangle, \tag{3.5}$$

which is self-explanatory in terms of (3.3) and (3.4).

Our task is now to form the general basis state

$$\left|\left(\begin{matrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51} & & m_{52} & \\ & J & & \Lambda & \\ & M_J & & M_\Lambda & \end{matrix}\right)\right\rangle$$

$$= \sum_{\substack{q, n_{51} \\ J_3+\Lambda_3=\frac{1}{2}(m_{62}-m_{63}) \\ J_i, \Lambda_i}} \left\langle \begin{matrix} m_{51} & m_{52} \\ J & \Lambda \end{matrix} \middle| \begin{bmatrix} q & \\ \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) \\ J_3 & \Lambda_3 \\ J-J_i & \Lambda-\Lambda_i \end{bmatrix} \middle| \begin{matrix} n_{51} & \frac{1}{2}(m_{62}+m_{63}) \\ J_i & \Lambda_i \end{matrix} \right\rangle$$

$$\times \left\langle \begin{matrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51} & & m_{52} & \end{matrix} \middle| \begin{bmatrix} \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) \\ & \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) \\ & & q \end{bmatrix} \middle| \begin{matrix} m_{61}+\frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}+m_{63}) & \frac{1}{2}(m_{62}+m_{63}) \\ & n_{51} & \frac{1}{2}(m_{62}+m_{63}) \end{matrix} \right\rangle$$

$$\times \sum_{\substack{m_i+m_3=M_J \\ \mu_i+\mu_3=M_\Lambda}} C^{J_i\quad J_3\quad J}_{m_i\quad m_3\quad M_J} C^{\Lambda_i\quad \Lambda_3\quad \Lambda}_{\mu_i\quad \mu_3\quad M_\Lambda} \left|\left(\begin{matrix} m_{61}+\frac{1}{2}(m_{62}-m_{63}) & & \frac{1}{2}(m_{62}+m_{63}) & & \frac{1}{2}(m_{62}+m_{63}) \\ & n_{51} & & \frac{1}{2}(m_{62}+m_{63}) & \\ & J_i & & \Lambda_i & \\ & m_i & & \mu_i & \end{matrix}\right)\right\rangle$$

$$\times \left| \begin{pmatrix} \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) \\ & \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) \\ & J_3 & \Lambda_3 \\ & m_3 & \mu_3 \end{pmatrix} \right\rangle, \tag{3.6}$$

where the totally symmetric representation is given by

$$\left| \begin{pmatrix} \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) \\ & \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) \\ & J_3 & \Lambda_3 \\ & m_3 & \mu_3 \end{pmatrix} \right\rangle$$

$$= (-1)^{\Lambda_3-\mu_3} \times \frac{(a_1^3)^{J_3+m_3}(a_4^3)^{J_3-m_3}(a_2^3)^{\Lambda_3+\mu_3}(a_3^3)^{\Lambda_3-\mu_3}}{[(J_3+m_3)!(J_3-m_3)!(\Lambda_3+\mu_3)!(\Lambda_3-\mu_3)!]^{1/2}} |0\rangle \tag{3.7}$$

under the restriction

$$J_3 + \Lambda_3 = \tfrac{1}{2}(m_{62} - m_{63}). \tag{3.8}$$

In (3.6) the operator pattern of the reduced $Sp(4)$ tensor operator is abbreviated by the single index of multiplicity q, and the operator pattern of the reduced $SU(4)$ operator is omitted. The omission is possible since the coupling of totally symmetric tensor operators is multiplicity-free in $SU(4)$. The matrix elements of totally symmetric tensor operators in $Sp(4)$ have already been studied,[9] so the task remaining to us is to determine the reduced matrix element in $SU(4)$ alone. The expression (2.9) yields

$$\left\langle \begin{matrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51} & & m_{52} & \end{matrix} \middle\| \begin{bmatrix} \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) \\ & \frac{1}{2}(m_{62}-m_{63}) & \frac{1}{2}(m_{62}-m_{63}) \\ & & q \end{bmatrix} \middle\| \begin{matrix} m_{61}+\frac{1}{2}(m_{62}+m_{63}) & & \frac{1}{2}(m_{62}+m_{63}) & & \frac{1}{2}(m_{62}+m_{63}) \\ & n_{51} & & \frac{1}{2}(m_{62}+m_{63}) & \end{matrix} \right\rangle$$

$$= \left(\frac{(m_{61}+m_{63}+2)!(m_{61}-m_{62}+1)!(2m_{51}+3)!(m_{51}-m_{52}+1)(m_{51}+m_{52}+2)}{(m_{61}+m_{51}+3)!(m_{62}+m_{51}+2)!(m_{61}-m_{51})!(m_{51}+m_{63}+1)!} \right.$$

$$\left. \times \frac{(m_{51}-m_{63}+1)!(m_{52}-m_{63})!}{(m_{61}-m_{52}+1)!(m_{61}+m_{52}+2)!} \right)^{1/2}$$

$$\times \left(\frac{(m_{62}-m_{52})![m_{61}+\frac{1}{2}(m_{62}-m_{63})+n_{51}+3]![m_{61}+\frac{1}{2}(m_{62}-m_{63})-n_{51}]!(2n_{51}+3)!(2m_{52}+1)}{(m_{51}-m_{62})!(m_{62}+m_{52}+1)!(m_{52}+m_{63})![n_{51}+\frac{1}{2}(m_{62}-m_{63})-m_{51}]!} \right)^{1/2}$$

$$\times (\{\tfrac{1}{2}[n_{51}+m_{51}-\tfrac{1}{2}(m_{62}-m_{63})]-q\}!\{\tfrac{1}{2}[n_{51}+m_{51}-\tfrac{1}{2}(m_{62}-m_{63})]+1+q\}!)^{1/2}$$

$$\times \frac{\mathfrak{M}^{1/2}(n_{51}+\frac{1}{2}(m_{62}-m_{63})-m_{51};\ \frac{1}{2}(m_{51}+m_{52}),\ \frac{1}{2}(m_{51}-m_{52}))}{[n_{51}+\frac{1}{2}(m_{62}-m_{63})+m_{51}+2]!}$$

$$\times \sum_{\Psi} (-1)^{m_{62}+n_{51}+\frac{1}{2}(m_{62}-m_{63})} (-1)^{\frac{1}{2}[m_{51}-n_{51}+\frac{1}{2}(m_{62}-m_{63})]+q-m_{52}} (2\Psi+1)(2q+1)^{1/2}$$

$$\times \left(\frac{[\Psi+m_{52}-\frac{1}{2}(m_{62}+m_{63})]![\Psi-m_{52}+\frac{1}{2}(m_{62}+m_{63})]!(\Psi+n_{51}-m_{51})![m_{52}+\frac{1}{2}(m_{62}+m_{63})+1+\Psi]!}{[\frac{1}{2}(m_{62}-m_{63})+\Psi+1]![\frac{1}{2}(m_{62}-m_{63})-\Psi]!(\Psi+m_{51}-n_{51})!} \right)^{1/2}$$

$$\times \{[m_{52}+\tfrac{1}{2}(m_{62}+m_{63})-\Psi]!\}^{1/2}$$

$$\times \frac{(m_{61}+m_{62}-m_{51}+m_{52}+1)![m_{61}-m_{51}+n_{51}+m_{52}+\frac{1}{2}(m_{62}-m_{63})+2]!}{[m_{61}+m_{62}-m_{51}+m_{52}-\frac{1}{2}(m_{62}-m_{63})+1+\Psi]![n_{51}-m_{51}+m_{52}-\frac{1}{2}(m_{62}-m_{63})]!}$$

$$\times \frac{1}{[\Psi-m_{52}+\frac{1}{2}(m_{62}+m_{63})]![n_{51}-m_{63}+1-\Psi)!}$$

$$\times \begin{Bmatrix} \frac{1}{2}[m_{51}-n_{51}+\frac{1}{2}(m_{62}-m_{63})] & \frac{1}{2}[n_{51}-m_{51}+\frac{1}{2}(m_{62}-m_{63})] & \Psi \\ \frac{1}{2}(m_{62}+m_{63}) & m_{52} & q \end{Bmatrix}$$

$$\times {}_3F_2(m_{52}-\tfrac{1}{2}(m_{62}+m_{63})-\Psi, n_{51}-m_{51}-\Psi, m_{61}-m_{51}+n_{51}+m_{52}+\tfrac{1}{2}(m_{62}-m_{63})+3 \mid$$

$$n_{51}-m_{63}-\Psi+2,\ n_{51}-m_{51}+m_{52}-\tfrac{1}{2}(m_{62}+m_{63})+1 \mid 1), \tag{3.9}$$

where we have adopted the phase conventions and parametrization of multiplicity space of Ref. 9. The undetermined normalization constant $\mathfrak{M}^{1/2}$ corresponds to the same undetermined constant in the matrix element of the reduced $Sp(4)$ operator given in Eq. (2.27) in Ref. 9, so that the product of (3.9) and Eq. (2.27) in Ref. 9 is completely determined.

We note that we may perform summations over q and n_{51} in (3.6) and obtain an expression for the general state which contains fewer indices of summation; direct coupling in (2.18) gives us the expression

$$\left|\begin{pmatrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51} & & m_{52} & \\ & m_{41} & & m_{42} & \\ & & m_{41} & & \\ & & m_{41} & & \end{pmatrix}\right\rangle$$

$$= \mathfrak{N}^{1/2}\left(\frac{(A+B+C-m-n-n'+1)!(A+B+C-m-n-m'+1)!(A+C-2n-n)!n'!m'!}{(A+B+C-n-m+1)!(A+B+C-n-m-n'-m'+1)!(A+C-2n)!(B-m+n)!}\right)^{1/2}$$

$$\times\left(\frac{(B-m-m'+n)!(A+2B+C-2m-m'+2)!(B-m-m'+n+n')!}{(A+2B+C-2m+2)!(B-m+n+n'+1)!}\right)^{1/2}$$

$$\times (A+B-m+1)!(A-n)!(C-n)!n!(B+C-m+1)!$$

$$\times \sum_{\substack{J_1J_2\\ \Lambda_1\Lambda_2}} \sum_{\substack{J_3+\Lambda_3=\frac{1}{2}A\\ J_i\Lambda_i}} \sum_{\substack{xyz\\ uvw\\ kt}} (-1)^{x+v+w+y+z+J_1+J_2-J_i+2J_3+m+n} \frac{(B-m+n-x)!(A+B-m-m'-t)!t!(m+x)!}{w!(n-x-w)!(n'+n-A+\Lambda_1+\Lambda_2+\Lambda_3-\Lambda-k+u-w)!}$$

$$\times \frac{(J_1+J_2+J_3-J-m-m'+k)!}{(C-m-m'-n'-2n+J_1+J_2+J_3-J+k-u+w)!(\Lambda_1+\Lambda_2+\Lambda_3-\Lambda-k-y+u-w)!}$$

$$\times \frac{(\Lambda_1+\Lambda_2+\Lambda_3-\Lambda-k-t)!}{(J_1+J_2+J_3-J-m-m'-n+k-u-z+\omega)!} \frac{1}{v!(m+x-v)!(k+t-y-v)!(m'-k-x-z+v)!}$$

$$\times \frac{1}{(B+C-m+1-k+u+v)!(A+B-m-m'+1+k-u-v)!y!z!}$$

$$\times \frac{1}{(y-t-u)!(u+x+z)!(t-y-z)!(B-m-m'+n-t+y+z)!(J+J_3-J_1-J_2+m+m'-k)!}$$

$$\times \frac{1}{(\Lambda+\Lambda_3-\Lambda_1-\Lambda_2+k+t)!(J_1+J_2-J_i-m-m'+k)!(J_1+J_2+J_i+1-m-m'+k)!}$$

$$\times \frac{1}{(\Lambda_1+\Lambda_2-\Lambda_i-k-t)!(\Lambda_1+\Lambda_2+\Lambda_i+1-k-t)!} \nabla(JJ_iJ_3)\nabla(\Lambda\Lambda_i\Lambda_3)\nabla(\tfrac{1}{2}CJ_i\Lambda_i)$$

$$\times \left(\frac{(2J_i+1)(2\Lambda_i+1)(J_1+J_2-J_i)!(J_1+J_2+J_i+1)!(\Lambda_1+\Lambda_2-\Lambda_i)!(\Lambda_1+\Lambda_2+\Lambda_i+1)!}{(2J+1)!(2\Lambda+1)!(C+1)!}\right)^{1/2}$$

$$\times C\begin{matrix} J_i & \Lambda_i & \frac{1}{2}C \\ J_1-J_2 & \Lambda_1-\Lambda_2 & \frac{1}{2}C \end{matrix} \sum_{m_i,\mu_i} C\begin{matrix} J_1 & J_2 & J_i \\ m_1 & m_2 & m_{12} \end{matrix} C\begin{matrix} \Lambda_1 & \Lambda_2 & \Lambda_i \\ \mu_1 & \mu_2 & \mu_{12} \end{matrix} C\begin{matrix} J_i & J_3 & J \\ m_{12} & m_3 & J \end{matrix} C\begin{matrix} \Lambda_i & \Lambda_3 & \Lambda \\ \mu_{12} & \mu_3 & \Lambda \end{matrix}$$

$$\times \frac{(a_1^1)^{J_1+m_1}(a_4^1)^{J_1-m_1}(a_1^2)^{J_2+m_2}(a_4^2)^{J_2-m_2}(a_1^3)^{J_3+m_3}(a_4^3)^{J_3-m_3}}{[(J_1+m_1)!(J_1-m_1)!(J_2+m_2)!(J_2-m_2)!(J_3+m_3)!(J_3-m_3)!]^{1/2}}$$

$$\times \frac{(a_2^1)^{\Lambda_1+\mu_1}(a_3^1)^{\Lambda_1-\mu_1}(a_2^2)^{\Lambda_2+\mu_2}(a_3^2)^{\Lambda_2-\mu_2}(a_2^3)^{\Lambda_3+\mu_3}(a_3^3)^{\Lambda_3-\mu_3}}{[(\Lambda_1+\mu_1)!(\Lambda_1-\mu_1)!(\Lambda_2+\mu_2)!(\Lambda_2-\mu_2)!(\Lambda_3+\mu_3)!(\Lambda_3-\mu_3)!]^{1/2}}|0\rangle, \tag{3.10}$$

where

$$J=\tfrac{1}{2}(A+B+C-n-m-n'-m'),\quad \Lambda=\tfrac{1}{2}(B-m-m'+n+n'),\quad J_1+\Lambda_1=\tfrac{1}{2}(A+B+C),$$

$$J_2+\Lambda_2=\tfrac{1}{2}(A+B),\quad J_3+\Lambda_3=\tfrac{1}{2}A,\quad \nabla(a\,b\,c)=\left(\frac{(a+b-c)!(a+c-b)!(a+b+c+1)!}{(b+c-a)!}\right)^{1/2}. \tag{3.11}$$

In the case of the expression for the semimaximal state (3.10) reduces to

$$\left|\begin{pmatrix} m_{61} & & m_{62} & & m_{63} \\ & m_{51} & & m_{52} & \\ & m_{51} & & m_{52} & \\ & & m_{51} & & \\ & & m_{52} & & \end{pmatrix}\right\rangle$$

$$= \mathfrak{N}^{1/2}n!m!(-1)^{B-m} \sum_{\substack{J_1+\Lambda_1=\frac{1}{2}(A+B+C)\\ J_2+\Lambda_2=\frac{1}{2}(A+B)}} \sum_{\substack{J_3+\Lambda_3=\frac{1}{2}A\\ J_i\Lambda_i}} \sum_{r\rho} \nabla(\Lambda\Lambda_3\Lambda_i)\nabla(JJ_3J_i) \frac{\nabla|\Lambda_iJ_i\tfrac{1}{2}C)}{(\Lambda+\Lambda_3+\Lambda_i+1)!}$$

$$\times\left(\frac{(2J_i+1)(2\Lambda_i+1)(J_1+J_2-J_i)!(J_1+J_2+J_i+1)!(\Lambda_1+\Lambda_2-\Lambda_i)!(\Lambda_1+\Lambda_1+\Lambda_2+\Lambda_i+1)!}{(2J+1)!(2\Lambda+1)!(C+1)!}\right)^{1/2}$$

$$\times\frac{(\Lambda+\Lambda_3+\Lambda_i+1+\rho)!(\frac{1}{2}C+J+J_3-\Lambda_i-\rho)!}{r!(m-r)!\rho!(J_1+J_2-J-J_3-r+\rho)!(J+\Lambda+J_3+\Lambda_3+\Lambda_i-J_1-J_2+1+r)!}$$

$$\times\frac{(\Lambda_1+\Lambda_2+\Lambda_3-\Lambda-m+r)!}{(J+\Lambda+J_3+\Lambda_3-\Lambda_i-J_1-J_2+r-\rho)!}\;\frac{1}{(J+J_3-J_i-\rho)!(\frac{1}{2}C-J-J_3+\Lambda_i+\rho)!(J+J_3+J_i+1-\rho)!}$$

$$\times C\begin{matrix}J_i & \Lambda_i & \frac{1}{2}C\\ J_1-J_2 & \Lambda_1-\Lambda_2 & \frac{1}{2}C\end{matrix}\sum_{m_i\mu_i} C\begin{matrix}J_1 & J_2 & J_i\\ m_1 & m_2 & m_{12}\end{matrix} C\begin{matrix}\Lambda_1 & \Lambda_2 & \Lambda_i\\ \mu_1 & \mu_2 & \mu_{12}\end{matrix} C\begin{matrix}J_i & J_3 & J\\ m_{12} & m_3 & J\end{matrix} C\begin{matrix}\Lambda_i & \Lambda_3 & \Lambda\\ \mu_{12} & \mu_3 & \Lambda\end{matrix}$$

$$\times\frac{(a_1^1)^{J_1+m_1}(a_4^1)^{J_1-m_1}(a_1^2)^{J_2+m_2}(a_4^2)^{J_2-m_2}}{[(J_1+m_1)!(J_1-m_1)!(J_2+m_2)!(J_2-m_2)!]^{1/2}}\;\frac{(a_1^3)^{J_3+m_3}(a_4^3)^{J_3-m_3}(a_2^1)^{\Lambda_1+\mu_1}(a_3^1)^{\Lambda_1-\mu_1}(a_2^2)^{\Lambda_2+\mu_2}(a_3^2)^{\Lambda_2-\mu_2}}{[(J_3+m_3)!(J_3-m_3)!(\Lambda_1+\mu_1)!(\Lambda_1-\mu_1)!(\Lambda_2+\mu_2)!(\Lambda_2-\mu_2)!}$$

$$\times\frac{(a_2^3)^{\Lambda_3+\mu_3}(a_3^3)^{\Lambda_3-\mu_3}}{(\Lambda_3+\mu_3)!(\Lambda_3-\mu_3)!]^{1/2}}\,|0\rangle, \tag{3.12}$$

where the sums are taken under the restrictions (3.11) with the condition $m'=n'=0$. The notation of (3.10) and (3.12) is explained by (2.14) and (3.6). Observe that the sum over ρ in (3.12) is a Saalchützian ${}_4F_3$ series, an analytically continued Racah coefficient. The parameter r is the index of summation which emerges upon the binomial expansion of $(a_I^{12})^m=(a_{14}^{12}+a_{23}^{12})^m$.

In terms of the boson states which we have constructed we may now easily determine the matrix which transforms an irreducible representation of $SU(4)$ from its basis in the $U(3)\supset U(2)$ chain to its basis in $Sp(4)\supset SU(2)\times SU(2)$. We recall the boson state for the canonical basis, which has been studied by Ciftan,[1,2] and the present author[3]:

$$\left|\begin{pmatrix} m_{14} & m_{24} & m_{34} & 0\\ m_{13} & m_{23} & m_{33} & \\ m_{12} & m_{22} & & \\ m_{11} & & & \end{pmatrix}\right\rangle=\sum_{[\beta]_2}\sum_{(\gamma)(\gamma')}\sum_{[\alpha]_2}\sum_{(\alpha_{11})}\mathfrak{M}^{-1/2}([\beta]_2)D^{[\beta]_2}_{(\gamma)(\gamma')}(a_{12}^{12})$$

$$\times C\begin{matrix}\frac{1}{2}(\beta_{12}-\beta_{22}) & \frac{1}{2}(m_{12}+m_{22}-\beta_{12}-\beta_{22}) & \frac{1}{2}(m_{12}-m_{22})\\ \gamma-\frac{1}{2}(\beta_{12}+\beta_{22}) & m_{11}-\gamma-\frac{1}{2}(m_{12}+m_{22}-\beta_{12}-\beta_{22}) & m_{11}-\frac{1}{2}(m_{12}+m_{22})\end{matrix}$$

$$\times\left\langle\begin{pmatrix} m_{13} & m_{23} & m_{33}\\ \alpha_{12} & \alpha_{22} & \\ \alpha_{11} & & \end{pmatrix}\middle|\begin{pmatrix} m_{13}+m_{23}+m_{33}-m_{12}-m_{32} & 0 & 0\\ \alpha_{12}+\alpha_{22}-\beta_{12}-\beta_{22} & 0 & \\ \alpha_{11}-\gamma' & & \end{pmatrix}\right\rangle\left\langle\begin{pmatrix} m_{12} & m_{22} & 0\\ \beta_{12} & \beta_{22} & \\ \gamma' & & \end{pmatrix}\right\rangle$$

$$\times\left\langle\begin{pmatrix} m_{14} & m_{24} & m_{34}\\ m_{14} & m_{24} & \\ m_{14} & & \end{pmatrix}\middle|\begin{pmatrix} m_{14}+m_{24}+m_{34}-m_{13}-m_{23}-m_{33} & 0 & 0\\ m_{14}+m_{24}-\alpha_{12}-\alpha_{22} & 0 & \\ m_{14}-\alpha_{11} & & \end{pmatrix}\right\rangle\left\langle\begin{pmatrix} m_{13} & m_{23} & m_{33}\\ \alpha_{12} & \alpha_{22} & \\ \alpha_{11} & & \end{pmatrix}\right\rangle$$

$$\times\frac{(a_1^3)^{m_{11}-\gamma}(a_2^3)^{m_{12}+m_{22}-m_{11}+\gamma-\beta_{12}-\beta_{22}}(a_3^1)^{\alpha_{11}-\gamma'}(a_3^2)^{\alpha_{12}+\alpha_{22}-\alpha_{11}+\gamma'-\beta_{12}-\beta_{22}}}{[(m_{11}-\gamma)!(m_{12}+m_{22}-m_{11}+\gamma-\beta_{12}-\beta_{22})!(\alpha_{11}-\gamma')!(\alpha_{12}+\alpha_{22}-\alpha_{11}+\gamma'-\beta_{12}-\beta_{22})!]^{1/2}}$$

$$\times\frac{(a_3^3)^{m_{13}+m_{23}+m_{33}-m_{12}-m_{22}-\alpha_{12}-\alpha_{22}+\beta_{12}+\beta_{22}}(a_4^1)^{m_{14}-\alpha_{11}}(a_4^2)^{m_{24}-\alpha_{12}-\alpha_{22}+\alpha_{11}}}{[(m_{13}+m_{23}+m_{33}-m_{12}-m_{22}-\alpha_{12}-\alpha_{22}+\beta_{12}+\beta_{22})!(m_{14}-\alpha_{11})!(m_{24}-\alpha_{12}-\alpha_{22}+\alpha_{11})!]^{1/2}}$$

$$\times\frac{(a_4^3)^{m_{34}-m_{13}-m_{23}-m_{33}+\alpha_{12}+\alpha_{22}}}{[(m_{34}-m_{13}-m_{23}-m_{33}+\alpha_{12}+\alpha_{22})!]^{1/2}}\,|0\rangle \tag{3.13}$$

in the notation of Ref. 3. Forming the inner product with the state (3.6) we obtain

$$\left\langle\begin{pmatrix} m_{61} & m_{62} & m_{63}\\ m_{51} & m_{52} & \\ J & \Lambda & \\ M_J & M_\Lambda & \end{pmatrix}\middle|\begin{pmatrix} m_{14} & m_{24} & m_{34} & 0\\ m_{13} & m_{23} & m_{33} & \\ m_{12} & m_{22} & & \\ m_{11} & & & \end{pmatrix}\right\rangle$$

$$=\sum_{\substack{J_i\Lambda_i\\ j\lambda}}\sum_{\substack{\alpha_{i2}\beta_{i2}\\ J_3+\Lambda_3=\frac{1}{2}(m_{62}-m_{63})}}\sum_{\substack{m_i\mu_i m_3\mu_3\\ n_{51}}}[(2J_i+1)(2\Lambda_i+1)]^{1/2}[(\beta_{12}-\beta_{22}+1)(\alpha_{12}-\alpha_{22}+1)]^{1/2}(-1)^{\frac{1}{2}(m_{14}+m_{24})+\lambda+\Lambda_3}$$

$$\times \begin{Bmatrix} \tfrac{1}{2}(m_{14}-m_{24}) & J_i & \Lambda_i \\ \tfrac{1}{2}(j+m_i) & \tfrac{1}{2}(\alpha_{12}-\alpha_{22}) & \tfrac{1}{2}(j-m_i) \end{Bmatrix} \begin{Bmatrix} \tfrac{1}{2}(\beta_{12}-\beta_{22}) & \tfrac{1}{2}(j+m_i) & \tfrac{1}{2}(\lambda+\mu_i) \\ \Lambda_i & \tfrac{1}{2}(\lambda-\mu_i) & \tfrac{1}{2}(\alpha_{12}-\alpha_{12}) \end{Bmatrix}$$

$$C \begin{matrix} \tfrac{1}{2}(\beta_{12}-\beta_{22}) & \tfrac{1}{2}[m_3+\mu_3+\tfrac{1}{2}(m_{62}-m_{63})] & \tfrac{1}{2}(m_{12}-m_{22}) \\ \tfrac{1}{2}(j-\lambda+m_i-\mu_i) & \tfrac{1}{2}(J_3-\Lambda_3+m_3-\mu_3) & m_{11}-\tfrac{1}{2}(m_{12}+m_{22}) \end{matrix}$$

$$\times C \begin{matrix} \tfrac{1}{2}[m_{61}+\tfrac{1}{2}(m_{62}-m_{63})-J_i-\Lambda_i] & \tfrac{1}{2}[m_{61}+\tfrac{1}{2}(m_{62}-m_{63})+J_i+\Lambda_i]+1 & n_{51}+1 \\ \tfrac{1}{2}(\lambda-j+J_i-\Lambda_i) & \tfrac{1}{2}(j-\lambda+J_i-\Lambda_i) & J_i-\Lambda_i \end{matrix}$$

$$\times \left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ \alpha_{12} & \alpha_{22} \end{matrix} \middle| \begin{bmatrix} \sum_{i=1}^{3} m_{i3} - \sum_{i=1}^{2} m_{i2} & 0 & 0 \\ \sum_{i=1}^{2} (\alpha_{i2}-\beta_{i2}) & 0 \end{bmatrix} \middle| \begin{matrix} m_{12} & m_{22} & 0 \\ \beta_{12} & \beta_{22} \end{matrix} \right\rangle$$

$$\times \left\langle \begin{matrix} m_{14} & m_{24} & m_{34} \\ m_{14} & m_{24} \end{matrix} \middle| \begin{bmatrix} \sum_{i=1}^{3} (m_{i4}-m_{i3}) & 0 & 0 \\ \sum_{i=1}^{2} (m_{i4}-\alpha_{i2}) & 0 \end{bmatrix} \middle| \begin{matrix} m_{13} & m_{23} & m_{33} \\ \alpha_{12} & \alpha_{22} \end{matrix} \right\rangle$$

$$\times \left\langle \begin{matrix} m_{61} & m_{62} & m_{63} \\ m_{51} & m_{52} & \\ J & \Lambda & \\ M_J & M_\Lambda & \end{matrix} \middle| \left\langle \begin{matrix} \tfrac{1}{2}(m_{62}-m_{63}) & \tfrac{1}{2}(m_{62}-m_{63}) & \tfrac{1}{2}(m_{62}-m_{63}) \\ \tfrac{1}{2}(m_{62}-m_{63}) & \tfrac{1}{2}(m_{62}-m_{63}) & \\ J_3 & \Lambda_3 & \\ m_3 & \mu_3 & \end{matrix} \middle| \begin{matrix} m_{61}+\tfrac{1}{2}(m_{62}-m_{63}) & \tfrac{1}{2}(m_{62}+m_{63}) & \tfrac{1}{2}(m_{62}+m_{63}) \\ n_{51} & \tfrac{1}{2}(m_{62}+m_{63}) & \\ J_i & \Lambda_i & \\ m_i & \mu_i & \end{matrix} \right\rangle \right\rangle . \quad (3.14)$$

Of the final three Wigner coefficients indicated in (3.14) the first two are reduced matrix elements in $U(3)$, which have been given explicitly by Chacón, Ciftan, and Biedenharn[10]; the operators are totally symmetric, hence multiplicity-free. The last Wigner coefficient is simply the total coupling coefficient indicated in (3.6). The following restrictions hold in (3.14) in addition to (2.4):

$$\begin{aligned} m_{13}+m_{23}+m_{33} &= 2\Lambda_3+2\lambda+j+J_3+M_J \\ &= m_{61}+m_{62}-m_{63}+\lambda+\Lambda_3+M_J, \\ m_{12}+m_{22} &= m_{61}+m_{62}-m_{63}+M_J+M_\Lambda, \\ m_{11} &= j+J_3+M_J, \\ j+\lambda &= m_{61}+\tfrac{1}{2}(m_{62}-m_{63}), \\ J_3+\Lambda_3 &= \tfrac{1}{2}(m_{62}-m_{63}). \end{aligned} \quad (3.15)$$

We have taken the inner product of the state (3.6) with a state of that $U(3)$ subgroup of $SU(4)$ which is simultaneously maximal with it, i.e., the matrix (3.14) has the property

$$\left\langle \begin{matrix} m_{61} & m_{62} & m_{63} \\ (\max) & & \end{matrix} \middle| \begin{matrix} m_{14} & m_{24} & m_{34} & 0 \\ & (\max) & & \end{matrix} \right\rangle = 1. \quad (3.16)$$

The inner product of (3.6) with $SU(4)$ states based in other $U(3)$ subgroups is easily found by means of appropriate Weyl transformations.

ACKNOWLEDGMENT

The author would like to thank Dr. Wesley Brittin for his kind hospitality at the University of Colorado.

*Present address: Physics Department, Duke University, Durham, North Carolina 27706.

[1] M. Ciftan, J. Math. Phys. **10**, 221 (1969).
[2] M. Ciftan, J. Math. Phys. **10**, 1635 (1969).
[3] W. J. Holman III, Nuovo Cimento A **4**, 904 (1971).
[4] K. T. Hecht and S. C. Pang, J. Math. Phys. **10**, 1571 (1969).
[5] J. P. Draayer, J. Math. Phys. **11**, 3225 (1970).
[6] S. C. Pang and K. T. Hecht, J. Math. Phys. **8**, 1233 (1967).
[7] K. T. Hecht, Nucl. Phys. **63**, 177 (1965).
[8] W. J. Holman III, J. Math. Phys. **10**, 1710 (1969).
[9] W. J. Holman III, J. Math. Phys. **14**, 440, (1973).
[10] E. Chacon, M. Ciftan, and L. C. Biedenharn, "On the Evaluation of the Multiplicity Free Wigner Coefficients of $U(n)$," Duke University, Durham, preprint (1971).

Conformally invariant orthogonal decomposition of symmetric tensors on Riemannian manifolds and the initial-value problem of general relativity*

James W. York Jr.

Joseph Henry Laboratories, Princeton University, Princeton, New Jersey 08540
(Received 25 July 1972)

It is shown that an arbitrary symmetric tensor ψ^{ab} (or ψ_{ab}) of any weight can be covariantly decomposed on a Riemannian manifold (M,g) into a unique sum of transverse-traceless, longitudinal, and pure trace parts. The summands involve only linear operators and are mutually orthogonal in the global scalar product on (M,g). Each summand transforms separately into itself if the decomposition is carried out properly in a conformally related space $(M,\bar{g})$. The decomposition is therefore determined by a conformal equivalence class of Riemannian manifolds. This property makes the decomposition ideally suited to the initial-value problem of general relativity, which becomes, as a result, a well-defined system of elliptic equations. Three of the four initial-value equations are linear and determine the decomposition of a symmetric tensor. The fourth equation is quasilinear and determines the conformal factor. The decomposition applied to the space of symmetric tensors on (M,g) can be written in terms of a direct sum of orthogonal linear spaces and gives a framework for treating and classifying deformations of Riemannian manifolds pertinent to the theory of gravitation and to pure geometry.

1. INTRODUCTION AND SUMMARY

The purpose of this paper is essentially twofold. Firstly, we give a conformally invariant, orthogonal, covariant decomposition of symmetric tensors on positive-definite Riemannian manifolds into transverse-traceless, longitudinal, and pure trace parts. Secondly, we show that this decomposition enables one to set the initial-value problem of general relativity as a system of four second-order elliptic equations for four unknown functions. Three of these equations are linear with a conformally-invariant vector field as the solution. The other is quasilinear and determines the conformal factor. Physically, the *unconstrained* fields correspond to pure spin-two transverse traceless dynamical variables. Geometrically, they describe the *anisotropy of space*. This description is of course equivalent to the conventional treatment of space-time in terms of a hyperbolic four-metric ${}^{(4)}g_{\mu\nu}$ satisfying Einstein's field equations, but is more useful for a number of purposes.

It is well known[1] that in flat or in curved Riemannian spaces one can decompose an arbitrary vector or one-form into its transverse and longitudinal parts. Physically, this procedure leads to the identification of the true canonical degrees of freedom of the electromagnetic field and to the identification of the gauge, or non-dynamical, variables. Thus, if $\mathbf{A}$ is the magnetic vector potential and $\mathbf{E}$ the electric field, then the transverse fields $\mathbf{A}_t$ and $\mathbf{E}_t (\nabla \cdot \mathbf{A}_t = \nabla \cdot \mathbf{E}_t = 0)$ are the dynamical or pure spin-one variables, while the longitudinal (l) part of $\mathbf{A}$ is determined by a choice of gauge. Moreover, this decomposition is not only covariant with respect to arbitrary coordinate transformations, it is also orthogonal in the natural global scalar product. That is, for any two vectors $\mathbf{v}$ and $\mathbf{w}$, we have

$$\int_M v_g V_t{}^a W_l{}^b g_{ab} \equiv (V_t, W_l) = 0, \tag{1}$$

where v_g denotes the volume element which makes the integral invariant and the integration extends over the entire manifold M. Geometrically, the decomposition of 1-forms, and more generally p-forms, leads via de Rham's theorem to a characterization of topological invariants of M (i.e., Betti Numbers).[2]

In this work, we shall consider *three*-dimensional Riemannian spaces because this case is of the greatest interest for physical applications. However, the results are easily generalized to n-spaces, provided they are Riemannian, i.e., have positive-definite metric.[3] Also, it is assumed that the three-spaces are *closed* (compact manifolds without boundary) and complete. The choice of closed spaces is made for mathematical convenience, and because closed three-spaces are of greatest interest in recent studies of the classical and quantum dynamics of general relativity.[4] However, the decomposition is also valid for open, asymptotically flat 3-spaces, certain assumptions being made as to the decay of the metric and other tensor fields as one approaches infinity.[5] In the pioneering work of Deser[6] on the covariant decomposition of tensors, only the case of asymptotic flatness was treated. Deser's procedure is also satisfactory for *closed* spaces; however, his decomposition is neither conformally invariant nor completely orthogonal in general.

There are two compelling physical motivations for decomposing symmetric tensors, both having to do with gravitation. One is the desire to separate gravitational variables into irreducible (spin) parts, so as to distinguish the dynamical (spin-2) variables from gauge variables and constrained variables. This was done in general relativity by Arnowitt, Deser, and Misner[1] using a *non*covariant decomposition and was treated covariantly by Deser[6] in later works. Another strong reason is that a covariant procedure should lead in the initial-value problem to a well-defined system of equations determining the constrained variables (Sec. 4).

From a mathematical point of view, the decomposition of tensor fields is also significant. A certain "T-decomposition" [6] (T = transverse) of symmetric tensors characterizes possible deformations of Riemannian manifolds, as shown by Ebin.[7] This is also important in general relativity, for the dynamics of gravity may be viewed as a continuous (time-dependent) deformation of spacelike hypersurfaces in enveloping spacetimes satisfying Einstein's field equations. This deformation process may also be viewed in superspace $\mathcal{S}$, which is the collection of Riemannian metrics *modulo* diffeomorphisms of M. Each point of $\mathcal{S}$ is a 3-*geometry* ${}^3\mathcal{G} = (M,g)$. The structure of superspace was examined in detail by Fischer[8] and was discussed in a helpful article by B. DeWitt.[9] In the Appendix, we show that for metrics possess-

ing symmetries, one can readily construct from one transverse tensor others which will also be transverse. Similar results hold for the TT decomposition.

In a certain sense the superspace picture and the associated T-decomposition of symmetric tensors are not sufficiently "fine" for all purposes in treating gravitation. This is the case because, using Wheeler's phrase,[10] "3-geometry is the carrier of information about time". Thus, ${}^3\mathcal{G}$ may be regarded as being specified by three functions (g_{ab} *modulo* diffeomorphisms of M) and its associated momentum, a transverse tensor density, is also specified by three functions. If these six functions were subject to no further constraints on a spacelike hypersurface, we would conclude that the gravitational field has three, not two, canonical degrees of freedom at each point of space. However, the metric and momentum *are* subject to an additional constraint. (The transversality of the momentum already comprises three of the four equations of constraint). We may regard, therefore, one of the metric variables and one of the momentum variables as describing a nondynamical pair of variables subject to a constraint. To what physical quantities could they correspond? Not surprisingly, they correspond to "time" and the "Hamiltonian."[11] Qualitatively similar arguments to this effect have been given in analyses of the dynamics of gravity by Dirac,[12] by Misner,[4] by Arnowitt, Deser, and Misner,[1] by Wheeler,[10] by Kuchař,[13] by the author,[11] and by others. Therefore, because there is *one* constraint on these *two* variables, one of them may be specified and the other must be determined by satisfying the remaining initial-value equation.

Details of the procedure followed differ among different students of the problem. Some prefer simply to carry along the extra pair, letting them, in effect, be governed implicitly by the fourth constraint written as a "Hamilton-Jacobi" type functional differential equation. This is what may be called roughly the "superspace approach."[14] Others make a choice of space and time coordinates based on some "preferred" background metric and select dynamical and nondynamical variables according to the criterion of convenience in treating the problem at hand. The essence of the latter formulation is found in the "mini-superspace" approach,[15,16] where only metrics with certain specified symmetries are considered. Thus, all degrees of freedom of the field are assumed to be frozen except those with the given symmetry. This procedure involves an element of risk because the gravitational field is nonlinear: degrees of freedom with different symmetry, or no symmetry, *do* interact; thus, ignoring some of them could lead to misleading conclusions. Moreover, in the quantum version of this approach, freezing certain degrees of freedom manifestly violates the uncertainty principle. These points are well known and have been discussed in the literature.[15,16] Of course, there are a number of other schemes for dealing with the problems engendered by the constraints, but I will not discuss them here.

One notices the "dual" nature of this problem: (1) picking variables in order to satisfy ultimately the initial-value equations on a spacelike hypersurface; (2) treating the dynamics of gravity based on this choice. In reality, however, there is only *one* problem because if the constraints are satisfied at one moment, they automatically continue to hold at succeeding moments by virtue of the field equations at that moment and the contracted Bianchi identities ${}^{(4)}\nabla_\mu {}^{(4)}G^{\mu\nu} \equiv 0$. [We use a prefix (4) and Greek indices when referring to space–time, as opposed to space.] Moreover, it has been shown that from the initial-value equations one can recover the entire dynamical content of general relativity.[17]

There is, however, a specific choice of variables which lead to a refinement of the superspace approach.[11,18] To proceed further, I make the only apparent simple three-dimensionally covariant refinement of variables, namely, the scale or conformal factor of the metric and the trace of the momentum tensor (actually the *scalar* part of this trace) are identified as separate entities. These two variables are selected to comprise the "extra" pair. There are compelling physical reasons for this choice. Firstly, I have previously shown that the 3-geometry *modulo* conformal factor specifies the pure spin-2 (TT) representation of space geometry.[18] Secondly, the scalar trace of the momentum is just the "volume Hubble parameter", i.e., the specific rate of volume expansion of the 3-space as it evolves in space–time.[11] Therefore, this scalar is naturally identifiable with "time," in that it identifies the expansion or contraction epoch of the 3-space in its history. Thirdly, if we use this scalar as $x^0 = t$ = time, then on the surfaces t = const., the trace-free part of the momentum tensor is transverse. Hence, the dynamical part of the momentum tensor is TT, and a TT tensor is determined only by the underlying *conformal* geometry.[18] (See Sec. 3). Mathematically, this procedure is fruitful in the initial-value problem (Sec. 4). Earlier, Dirac[12] made a related choice of variables, in that he treated the scale factor of the geometry as a nondynamical variable. However, he did not show that the conformal geometry has the properties we have mentioned above, nor did he use the scalar trace of the momentum as time. Rather, he set this trace equal to *zero* for all time as an implicit condition on the time coordinate. It is by now well known that time variables compatible with his procedure do not exist in closed universes, i.e., for closed 3-spaces, in general.[14] This defect is not present in our case; in fact, the scalar trace need not necessarily be a constant (Sec. 4). In any case, our time variable is local and identifies the surface in spacetime to which one is referring.

Because the 3-geometry contains as an "extra" variable the scale factor, it does in a sense "carry information about time," though somewhat indirectly. The fact that the 3-geometry must carry extra information of some kind led Wheeler[19] to ask, in effect, "What is two-thirds of superspace?" An answer to this question is given in Sec. 6, namely, "conformal superspace."

The "TT-decomposition" described in this paper characterizes deformations of *conformal* Riemannian manifolds $(M, \tilde{g}) \equiv {}^3\mathcal{K}$, where $\tilde{g}_{ab} \equiv (\det g)^{-1/3} g_{ab}$. Thus ${}^3\mathcal{K}$ is determined by a Riemannian metric g_{ab} *modulo* diffeomorphisms of M and *modulo* conformal mappings $\bar{g}_{ab} = \phi^4 g_{ab}$, where $\phi(x)$ is an arbitrary, real, nonvanishing scalar function on M. (We may say ${}^3\mathcal{K}$ is determined by g_{ab} *modulo* "conformeomorphisms"[20] for short.) The collection of conformal 3-geometries may be called "conformal superspace" $\tilde{\mathcal{S}}$. Each point of $\tilde{\mathcal{S}}$ is a ${}^3\mathcal{K}$. We sometimes find it convenient to regard ${}^3\mathcal{K}$ also as a set of conformally equivalent Riemannian 3-geometries. In all respects, the TT-decomposition bears the same relation to $\tilde{\mathcal{S}}$ and to the dynamics of ${}^3\mathcal{K}$ in space–time that the T-decomposition bears to $\mathcal{S}$ and to the dynamics of ${}^3\mathcal{G}$ in space–time.

However, the TT-decomposition has features that the T-decomposition does not possess. Notably, it is a "finer" splitting of a symmetric tensor than the T-decomposition in that it has more independent pieces. This fineness leads to its conformal invariance, which,

in turn, makes it quite useful in analysis of the initial-value equations. Moreover, since conformally invariant objects described the pure spin-2 aspects of gravity.[18] the TT-decomposition is ideally suited to the construction of such objects.

In Sec. 2, the TT-decomposition of a symmetric tensor ψ^{ab} is defined by

$$\psi^{ab} = \psi^{ab}_{\rm TT} + \psi^{ab}_{\rm L} + \psi^{ab}_{\rm Tr}, \tag{2}$$

where the longitudinal part is

$$\psi^{ab}_{\rm L} \equiv \nabla^a W^b + \nabla^b W^a - \tfrac{2}{3} g^{ab} \nabla_c W^c \equiv (LW)^{ab} \tag{3}$$

and the trace part is

$$\psi^{ab}_{\rm Tr} \equiv \tfrac{1}{3} \psi g^{ab}, \quad \psi \equiv g_{cd}\psi^{cd}. \tag{4}$$

We use ∇_a to denote covariant differentiation and ∂_a ordinary differentiation. Our conventions are such that

$$\nabla_{[b}\nabla_{c]} V_a = \tfrac{1}{2} V_d R^d{}_{acb}, \tag{5}$$

$$R_{ab} = R^c{}_{acb}. \tag{6}$$

The transversality requirement leads to the only equations that need be solved in this procedure. This can be done, uniquely, with W^a the solution. The TT, L, and TR parts are mutually orthogonal. Upon conformal mapping of g_{ab} and ψ^{ab}, (2) maintains the same form and the equations guaranteeing transversality have the *same* solution W^a after the mapping as they did before the mapping. These conformal properties are demonstrated in Sec. 3.

In Sec. 4, we discuss application of the TT-decomposition to the initial-value problem. In Sec. 5, the decomposition is written in terms of orthogonal projection operators on the space of symmetric tensors. Deformations of conformal Riemannian manifolds are treated in the final section. In the Appendix, we show that when symmetries are present, one may readily construct from a given TT tensor others that are automatically TT.

2. TRANSVERSE–TRACELESS DECOMPOSITION

We define $\psi^{ab}_{\rm TT}$ in accordance with (2) by

$$\psi^{ab}_{\rm TT} \equiv \psi^{ab} - \tfrac{1}{3}\psi g^{ab} - (LW)^{ab}. \tag{7}$$

Let us suppose that both ψ^{ab} and g_{ab} are C^∞ tensor fields. For concreteness, we work here with tensors rather than with tensor densities; one need only multiply through by an appropriate power of $g^{1/2}$ for the densities. We note that the trace condition

$$g_{ab}\psi^{ab}_{\rm TT} = 0 \tag{8}$$

is satisfied by construction. The transversality requirement

$$\nabla_b \psi^{ab}_{\rm TT} = 0 \tag{9}$$

leads to covariant equations for the vector field W^a:

$$(DW)^a \equiv -\nabla_b (LW)^{ab} = -\nabla_b(\psi^{ab} - \tfrac{1}{3}\psi g^{ab}). \tag{10}$$

Notice that only the divergence of the *trace-free* part of ψ^{ab} enters (10). As a result, it is helpful to introduce an abbreviated notation. Define the algebraic operator Λ that projects any symmetric tensor into its trace-free part:

$$\Lambda^{ab}_{cd}\psi^{cd} \equiv (\delta^{(a}_c \delta^{b)}_d - \tfrac{1}{3} g^{ab} g_{cd})\psi^{cd} \equiv \psi^{ab} - \tfrac{1}{3}\psi g^{ab}. \tag{11}$$

Also, let the divergence of ψ^{ab} be denoted by $\nabla\cdot\psi$. Then (10) assumes the abbreviated form

$$DW \equiv -\nabla\cdot LW = -\nabla\cdot\Lambda\psi. \tag{10'}$$

Let us now discuss the basic properties of (10). The operator D is linear and second order as we see by inspection. Moreover, as we show below, this operator is positive-definite, Hermitian,[21] and its "harmonic"[22] functions are always orthogonal to the source (right-hand side) in (10). Thus (10) will always possess solutions[2] W unique up to conformal Killing vectors (see below). These solutions can be obtained by the eigenfunction method, assuming that D possesses a complete set of orthogonal eigenfunctions. We expect the spectrum of D to be discrete for closed spaces. If not, we assume its eigenvalues do not have zero as an accumulation point.[21] The operator inverse to D, D^{-1}, therefore exists and we may write

$$W = D^{-1}[-\nabla\cdot\Lambda\psi]. \tag{12}$$

To show that D is positive definite, we multiply $(DW)^a$ by W_a, form the global scalar product, and integrate by parts to show

$$(W, DW) = \tfrac{1}{2}(LW, LW), \tag{13}$$

where

$$(LW, LW) \equiv \int_M v_g (LW)_{ab}(LW)^{ab} \geq 0. \tag{14}$$

Thus D is positive unless $LW = 0$, a case discussed below. That D is Hermitian follows from a similar argument in which one integrates by parts twice to find

$$(V, DW) = (DV, W) \tag{15}$$

for any vectors V and W.

The right-hand side of (13) can vanish only if $LW = 0$. This means either $W = 0$ or $W =$ conformal Killing vector (CKV) of the metric. The condition for a CKV is, of course, not satisfied for an arbitrary ("conformally wild") metric. The condition for a CKV is given by $\pounds_W \tilde{g}_{ab} = 0$ or

$$\pounds_W g_{ab} = \lambda g_{ab} \tag{16}$$

for some scalar function λ, where $\pounds_W$ denotes the Lie derivative along W. Equation (16) is just

$$\nabla_a W_b + \nabla_b W_a = \lambda g_{ab}. \tag{17}$$

Taking the trace of both sides, we find

$$\lambda = \tfrac{2}{3}\nabla_c W^c. \tag{18}$$

Therefore, W is a CKV if and only if

$$\nabla^a W^b + \nabla^b W^a - \tfrac{2}{3} g^{ab}\nabla_c W^c \equiv (LW)^{ab} = 0. \tag{19}$$

It follows that the only nontrivial solutions of $DW = 0$ are CKV's, if they exist. Hence the nontrivial "harmonic" functions of D are CKV's. We shall now show that even if these "harmonic" solutions exist, they are always orthogonal to the right-hand side of (10) and, hence, can cause no difficulties in solving equation (10) by an eigenfunction expansion.

Denote the CKV's by $W^a = C^a$, where by definition

$LC = 0$. Form the scalar product of the right-hand side of (10) with C and integrate by parts to find

$$(-\nabla\cdot\Lambda\psi, C) = \tfrac{1}{2}(\Lambda\psi, LC) = 0. \tag{20}$$

Hence the source is in the domain of D^{-1} and D^{-1} gives the solution of (10) even in the presence of conformal symmetries.

The above results also show that the solution of (10) must be unique up to CKV's. However, since only $(LW)^{ab}$ enters in the definition (7) of $\psi^{ab}_{\rm TT}$, CKV's cannot affect $\psi^{ab}_{\rm TT}$.

The orthogonality of $\psi^{ab}_{\rm TT}$, $(LW)^{ab}$, and $\frac{1}{3}\psi g^{ab}$ is easily demonstrated. We see readily that $\frac{1}{3}\psi g^{ab}$ is pointwise orthogonal to $(LW)^{ab}$ and to $\psi^{ab}_{\rm TT}$, as $(LW)^{ab}$ and $\psi^{ab}_{\rm TT}$ are both trace-free. To show that $\psi^{ab}_{\rm TT}$ and $(LV)^{ab}$ are orthogonal for any vector V and any TT tensor, we have only to show that

$$(LV, \psi_{\rm TT}) = 0, \tag{21}$$

which follows readily using integration by parts, Gauss's theorem, and $\nabla\cdot\psi_{\rm TT} = 0$. We conclude, therefore, that the decomposition defined by (7) exists, is unique, and is orthogonal. The properties of (7) under conformal mappings of g_{ab} and ψ^{ab} are the subject of the next section.

One can further decompose the vector W^a uniquely into *its* transverse (t) and longitudinal (l) parts with respect to the metric g_{ab}. This splitting is orthogonal, as in (1). But W_t and W_l themselves "mix" under the conformal transformations defined in the next section. In any event, since this further splitting is well defined, we see that an arbitrary symmetric tensor field can be split into a sum of pure spin-two (TT), pure spin-one (W_t) and spin-zero ($g_{ab}\psi^{ab}$ and $\nabla\cdot W_l$) parts.

Also, it is easily verified in this procedure that a given tensor that is already TT has no L or Tr parts; a pure L tensor has no TT or Tr parts; and a pure Tr tensor has no TT or L parts.

3. CONFORMAL TRANSFORMATIONS

Understanding the conformal properties of (7) is of great interest in itself and is essential in the application of these results to the gravity initial-value problem (Sec. 4). A space conformally related to (M,g) is $(M,\bar g)$, where

$$\bar g_{ab} = \phi^4 g_{ab}. \tag{22}$$

Therefore, we have for the connection coefficients

$$\bar\Gamma_{bc}{}^a = \Gamma_{bc}{}^a + 2(\delta^a_b\nabla_c\ln\phi + \delta^a_c\nabla_b\ln\phi - g_{bc}\nabla^a\ln\phi), \tag{23}$$

with $\phi(x)$ an arbitrary real positive scalar function. The freely given tensor ψ^{ab}, which is to be decomposed, will *also* be mapped conformally by the transformation

$$\bar\psi^{ab} = \phi^{-10}\psi^{ab}. \tag{24}$$

Thus, on $(M,\bar g)$ we will decompose $\phi^{-10}\psi^{ab}$, not ψ^{ab} itself. We shall prove that decomposing $\phi^{-10}\psi^{ab}$ on $(M,\bar g)$ is completely equivalent to decomposing ψ^{ab} on (M,g), for an *arbitrary* choice of $\phi(x)$. This fact is of essential importance in the application of (7) to the gravitational initial-value problem. The choice of (24) is not arbitrary, but is dictated by the form of (7), as we shall see below. If ψ^{ab} were a tensor density of weight one, we would use in place of (24), $\bar\psi^{ab} = \phi^{-4}\psi^{ab}$. If it were of weight 5/3, we would have $\bar\psi^{ab} = \psi^{ab}$. This follows from (24) and $\bar g^{1/2} = \phi^6 g^{1/2}$. These cases are also important in the initial-value problem.[18]

Returning to the tensor case, we rewrite (7), using Eqs. (22), (23), (24), to obtain

$$\psi^{ab}_{\rm TT} = \phi^{10}(\bar\psi^{ab} - \tfrac{1}{3}\bar\psi\bar g^{ab}) - (LW)^{ab}, \tag{25}$$

where $\bar\psi \equiv \bar g_{cd}\bar\psi^{cd}$. We now wish to transform the longitudinal part, $(LW)^{ab}$. Substitution of (22) and (23) into (3) gives

$$(LW)^{ab} = \phi^4(\bar\nabla^a W^b + \bar\nabla^b W^a - \tfrac{2}{3}\bar g^{ab}\bar\nabla_c W^c) \equiv \phi^4(\bar LW)^{ab}, \tag{26}$$

or

$$(\bar LW)^{ab} = \phi^{-4}(LW)^{ab}. \tag{27}$$

By using (27) and multiplying through by ϕ^{-10}, (25) becomes

$$\phi^{-10}\psi^{ab}_{\rm TT} = (\bar\psi^{ab} - \tfrac{1}{3}\bar g^{ab}\bar\psi) - \phi^{-6}(\bar LW)^{ab}. \tag{28}$$

Let us put $\bar T^{ab} \equiv \phi^{-10}\psi^{ab}_{\rm TT}$. We note that

$$\bar g_{ab}\bar T^{ab} = 0 \tag{29}$$

by construction. Furthermore, $\bar T^{ab}$ will also be transverse on $(M,\bar g)$ provided $\bar\nabla_b\bar T^{ab} = 0$, or

$$-\bar\nabla_b[\phi^{-6}(\bar LW)^{ab}] = -\bar\nabla_b(\bar\psi^{ab} - \tfrac{1}{3}\bar\psi\bar g^{ab}). \tag{30}$$

However, for any choice of ϕ, this equation for the "vector potention" W^a is already satisfied by the *same* vector W^a that satisfied (10)! This statement is proved as follows: The left-hand side of (30) is simply

$$\phi^{-6}[-\bar\nabla_b(\bar LW)^{ab} + 6(\bar LW)^{ab}\bar\nabla_b\ln\phi]. \tag{31}$$

Since $\ln\phi$ is a scalar, we have $\bar\nabla_b\ln\phi = \nabla_b\ln\phi = \partial_b\ln\phi$. Moreover, using (22), (23), and (27), we have

$$-\bar\nabla_b(\bar LW)^{ab} = -\phi^{-4}[\nabla_b(LW)^{ab} + 6(LW)^{ab}\nabla_b\ln\phi]. \tag{32}$$

By using (31) and (32), the left-hand side of (30) becomes

$$-\phi^{-10}\nabla_b(LW)^{ab}. \tag{33}$$

From (22), (23), and (24), the right-hand side of (30) becomes

$$-\phi^{-10}\nabla_b(\psi^{ab} - \tfrac{1}{3}\psi g^{ab}). \tag{34}$$

It is just this last result that uniquely dictates the conformal transformation (24) on ψ^{ab}. From (33) and (34) we see that (30) implies

$$-\nabla_b(LW)^{ab} = -\nabla_b(\psi^{ab} - \tfrac{1}{3}\psi g^{ab}), \tag{10}$$

which is just Eq. (10). Conversely, Eq. (10) implies (30), so that (10) and (30) are completely equivalent for any choice of ϕ and, therefore, have precisely the same solution $W^a(x)$. The "harmonic" functions of the operator $-\bar\nabla_b[\phi^{-6}(\bar LW)^{ab}]$ are precisely the same as those of $-\nabla_b(LW)^{ab} \equiv (DW)^a$, because $-\bar\nabla_b[\phi^{-6}(\bar LW)^{ab}]$ is a positive-definite operator which can only vanish if $W^a = 0$ or if $W^a =$ CKV of $\bar g_{ab}$. Since $(\bar LW)^{ab} = \phi^{-4}(LW)^{ab}$, we see that $(\bar LW)^{ab} = 0$ if and only if $(LW)^{ab} = 0$. This is not surprising, as it only says that if W^a is a CKV of g_{ab}, then it is also a CKV of any conformally related

metric, as one might expect. That is, the condition for a CKV is conformally invariant.

Hence, we identify $\bar{T}^{ab}$ with $\bar{\psi}^{ab}_{TT}$, yielding the result

$$\bar{\psi}^{ab}_{TT} = \phi^{-10}\psi^{ab}_{TT}. \tag{35}$$

Equation (21) may now be written

$$\bar{\psi}^{ab}_{TT} = (\bar{\psi}^{ab} - \tfrac{1}{3}\bar{g}^{ab}\bar{\psi}) - \phi^{-6}(\bar{L}W)^{ab}. \tag{36}$$

We see that the longitudinal part of $\bar{\psi}^{ab}$ is simply

$$\bar{\psi}^{ab}_{L} = (\bar{L}\bar{W})^{ab} = \phi^{-6}(\bar{L}W)^{ab} = \phi^{-10}(LW)^{ab}. \tag{37}$$

In summary, given ψ^{ab} on (M,g), we decompose it by means of

$$\psi^{ab} = \psi^{ab}_{TT} + \psi^{ab}_{L} + \psi^{ab}_{Tr}, \tag{2}$$

with ψ^{ab}_{L} and ψ^{ab}_{Tr} given by (3), (4), and the solution of (10). On a conformally related manifold $(M,\bar{g})$, the tensor $\bar{\psi}^{ab} = \phi^{-10}\psi^{ab}$ decomposes in the same way:

$$\bar{\psi}^{ab} = \bar{\psi}^{ab}_{TT} + \bar{\psi}^{ab}_{L} + \bar{\psi}^{ab}_{Tr}, \tag{38}$$

where

$$\bar{\psi}^{ab}_{TT} = \phi^{-10}\psi^{ab}_{TT}, \tag{35}$$

$$\bar{\psi}^{ab}_{L} = \phi^{-10}\psi^{ab}_{L}, \tag{37'}$$

$$\bar{\psi}^{ab}_{Tr} = \phi^{-10}\psi^{ab}_{Tr} \tag{39}$$

with the vector W^a determining the longitudinal part being the same for $\bar{\psi}^{ab}_{L}$ as for ψ^{ab}_{L}.

One can see from the above that the form of the decomposition may be viewed as being *determined* by conformal invariance, for we know *without* performing a decomposition that if a tensor is TT with respect to a given metric g_{ab}, that ϕ^{-10} times the tensor will also be TT with respect to the conformally transformed metric $\phi^4 g_{ab}$.[18] It is actually this latter observation whose significance led to the present method of decomposition. In particular, one can see that the form for the longitudinal part is crucial because of (27).

4. INITIAL-VALUE PROBLEM OF GENERAL RELATIVITY

Because of its conformal properties, the decomposition (7) is ideally suited for use in the gravitational initial-value problem. For simplicity, we shall first describe principally the case involving vacuum gravity fields ${}^{(4)}R_{\mu\nu} = 0$.

The initial-value problem is to construct a spacelike Riemannian three-manifold (M,g) and a symmetric tensor density of weight one, π^{ab}, such that

$$\nabla_b \pi^{ab} = 0, \tag{40}$$

$$g^{-1/2}(\pi_{ab}\pi^{ab} - \tfrac{1}{2}\pi^2) - g^{1/2}R = 0, \tag{41}$$

where R is the scalar curvature of (M,g). The conformal approach to this problem is to solve (40) in a conformally invariant manner, then to choose the conformal factor ϕ in such a way as to satisfy (41). Equations (40) and (41) are the Gauss–Codazzi equations giving necessary and sufficient conditions for the embedding of (M,g) with second fundamental tensor

$$K_{ab} = g^{-1/2}(\tfrac{1}{2}\pi g_{ab} - \pi_{ab}) \tag{42}$$

in a spacetime satisfying Einstein's equations ${}^{(4)}R_{\mu\nu} = 0$. For global analysis of this problem, it is convenient to convert (40) and (41) into second-order elliptic partial differential equations.

Here let us treat the case where (M,g) is to be embedded in such a way that its volume "Hubble parameter" is constant on the surface. The specific variable[11] which has proven to be of fundamental significance for this and other purposes is

$$\tau \equiv \tfrac{2}{3} g^{-1/2}\pi = \tfrac{4}{3}K, \tag{43}$$

which measures the rate of change of the local volume elements of (M,g) per unit volume, per unit proper time, i.e., per unit proper distance orthogonal to the surface. The "maximal" case is $\tau = 0$, but in general we simply assume that $\tau = \text{const.}$, i.e., $\nabla_a \tau = \partial_a \tau = 0$ on (M,g).

In this case (40) can be written as

$$\nabla_b(\pi^{ab} - \tfrac{1}{3}\pi g^{ab}) = 0 \tag{40'}$$

since by hypothesis $\nabla^a \pi = 0$. Thus (36) requires that we construct a TT tensor density $\sigma^{ab} \equiv \pi^{ab} - \frac{1}{3}\pi g^{ab}$. For this purpose, we give arbitrarily a tensor ψ^{ab} and construct its TT part as above by solving (10) for W^a. We then set $\sigma^{ab} = g^{1/2}\psi^{ab}_{TT}$.

Of course, the variables g_{ab}, σ^{ab} will not in general satisfy (41). However, we now map them conformally onto new variables which do satisfy (41), i.e., such that

$$\bar{g}^{-1/2}(\bar{\pi}_{ab}\bar{\pi}^{ab} - \tfrac{1}{2}\bar{\pi}^2) - \bar{g}^{1/2}\bar{R} = 0. \tag{44}$$

We first note that

$$\pi^{ab} = \sigma^{ab} + \tfrac{1}{2}g^{1/2}g^{ab}\tau, \tag{45}$$

which follows from the definition of σ^{ab} and τ. We know that the transformation $\bar{\sigma}^{ab} = \phi^{-4}\sigma^{ab}$ preserves the TT character of σ^{ab}, so that

$$\bar{\pi}^{ab} = \bar{\sigma}^{ab} + \tfrac{1}{2}\bar{g}^{1/2}\bar{g}^{ab}\tau = \phi^{-4}\sigma^{ab} + \tfrac{1}{2}\phi^2 g^{1/2}g^{ab}\tau \tag{46}$$

will satisfy (40) in the form $\bar{\nabla}_b\bar{\pi}^{ab} = 0$. Substituting (39) into (37) and using the well-known formula

$$\bar{R} = \phi^{-4}R + 8\phi^{-5}\Delta\phi, \tag{47}$$

with $\Delta\phi \equiv -g^{ab}\nabla_a\nabla_b\phi$, we obtain the equation determining ϕ:

$$(8\Delta + R)\phi = \mathfrak{M}\phi^{-7} - \tfrac{3}{8}\tau^2\phi^5, \tag{48}$$

where $\mathfrak{M} \equiv g^{-1}g_{ac}g_{bd}\sigma^{ab}\sigma^{cd}$. Equation (48) is a quasilinear elliptic equation determining $\phi(x)$. All of its coefficients involve only known or given functions, so it is not coupled back to the momentum constraint $\nabla_b\bar{\pi}^{ab} = 0$, the solution of which, as has been pointed out, is a conformally invariant problem, when $\tau = \text{const.}$

Quasilinear elliptic equations such as (48) are discussed, for example, in the treatise of Ladyzhenskaya and Ural'tseva.[23] The existence of solutions, particularly real positive[24] solutions $0 < \phi < \infty$, depends to a large extent on the detailed nature of the nonlinear terms and the possible values of their coefficients. For the particular form exhibited by (48), N. O'Murchadha and the author[25] have classified all cases of physical interest. For example, we have shown that solutions ϕ, such that $0 < \phi < \infty$ on M, exist for any $\mathfrak{M} > 0, \tau \neq 0$, and for any choice of the initial metric g_{ab} on a closed C^∞ manifold.

The case $\tau = 0$ is discussed below. It is exceptional only for *closed* manifolds. For asymptotically flat (open) spaces, existence of solutions has also been demonstrated. Moreover, for (48), we have shown that the solution is *unique* for closed *or* open spaces. (In the asymptotically flat case, the value of ϕ at infinity must also be specified, of course.) Together with the TT-decomposition, these results give a highly satisfactory description of the initial-value problem for the vacuum gravitational field. The same conclusion holds for the gravitational field with sources, described below.

The case $\tau = 0$ is exceptional for closed spaces and it is worthwhile to see why this is true. We shall see that if one chooses an initial metric with $R < 0$ everywhere on M, no solution of (48) exists if $\tau = 0$. Of course, the set of configurations with $\tau \equiv 0$ is of "measure zero" in the full configuration space $\tilde{S} \times \{\text{possible choices of } \tau\}$. Nevertheless, the discussion of this case sheds light on the meaning of the conformal method in general. Let us first, then, examine the conformal properties of (48) itself.

Had we started not from g_{ab}, but from any other metric $g'_{ab} = \nu^4 g_{ab}$ in the same conformal equivalence class as g_{ab}, we would of course have $\sigma'^{ab}_{TT} = \nu^{-4}\sigma^{ab}_{TT}$. The analog of (48) would be of the same form:

$$(8\Delta' + R')\phi' = \mathfrak{M}'(\phi')^{-7} - \tfrac{2}{3}\tau^2(\phi')^5, \tag{49}$$

where $\mathfrak{M}' = \nu^{-12}\mathfrak{M}$, $\tau' = \tau$, and $\phi' = \phi\nu^{-1}$. Note that $(8\Delta + R)$ is just the conformally invariant scalar Laplacian, so the fact that (48) is itself conformally form-invariant is not surprising. From this we see that the solution of (48) admits the conformal "gauge" transformation $\phi' = \phi\nu^{-1}$. Therefore, the uniqueness of solutions to (48) for given *conformal equivalence classes* of initial data is only uniqueness *modulo* this gauge behavior. It is clear, however, that the final metric $\bar{g}_{ab}$ and momentum $\bar{\sigma}^{ab}_{TT}$ are themselves unique with respect to the given conformal class of initial data. Thus,

$$\bar{g}_{ab} = \phi^4 g_{ab} = (\phi')^4 g'_{ab}, \tag{50}$$

$$\bar{\sigma}^{ab}_{TT} = \phi^{-4}\sigma^{ab}_{TT} = (\phi')^{-4}\sigma'^{ab}_{TT}. \tag{51}$$

The conclusion is that *the complete set of initial value equations for W^a and ϕ is conformally covariant. Their solutions transform by the rules $W^{a'} = W^a, \phi' = \phi\nu^{-1}$. From these solutions one obtains a unique initial-data set $\bar{g}_{ab}, \bar{\sigma}^{ab}_{TT}, \tau$.*

We now return to the case $\tau = 0, \mathfrak{M} > 0$. We see from (44) that $\bar{R}$ must be positive on all of M. If we choose an initial metric such that $R(g) < 0$ everywhere on M, then if a positive ϕ satisfying (48) exists, it must map from a space with $R < 0$ everywhere to one with $\bar{R} > 0$ everywhere. However, it is easy to show that no such mapping can exist for M *closed*. From (47), we have

$$8\Delta\phi = -R\phi + \bar{R}\phi^5. \tag{52}$$

Integrating (47) over a closed M gives

$$0 = \int_M v_g(-R\phi + \bar{R}\phi^5), \tag{53}$$

which cannot be satisfied in the present case.

However, the argument in the case $\tau = 0$ does not apply to asymptotically flat (open) spaces, as one cannot there *discard* the gradient of ϕ at infinity. In fact, this very boundary integral determines the mass-at-infinity of the gravitational configuration.[26]

What this simple argument also demonstrates is that, whereas the scalar curvature is not a conformally invariant object in a local sense, *globally* a *uniform sign* of R over a closed M is conformally significant. That is, if R has one sign only on M and if $\bar{R}$ also has only one sign (which is not true, in general, if $\tau \neq 0$), these signs must be the same.

This argument is closely related to *Yamabe's theorem:*[27] Every compact C^∞ Riemannian manifold of dimension ≥ 3 can be conformally deformed to a C^∞ Riemannian structure of *constant* scalar curvature. Let the constant be called k. Then $\text{sgn}(k)$ provides a convenient conformally invariant "index" for conformal equivalence classes of (closed) Riemannian manifolds. For the exceptional case $\tau = 0$, the conformal classes with $\text{sgn}(k) = 0$ and $\text{sgn}(k) = -1$ are here ruled out; whereas, if $\tau \neq 0$, they are not. For asymptotically flat spaces, no classification such as that provided by Yamabe's theorem is possible.

For a closed universe, $\tau = 0$ corresponds to the moment of maximum expansion. According to the argument above, we have the result that a closed vacuum gravitational configuration with $\text{sgn}(k) = -1$ or $\text{sgn}(k) = 0$ cannot correspond to a moment of maximum expansion. Examples of closed vacuum universes with no moment of maximum expansion are known.[28] These universes expand for all time at an ever-slowing rate which only approaches $\tau = 0$ asymptotically as the volume approaches infinity, which means the closed space is becoming effectively "open".

Conversely, at a moment of maximum expansion, $\text{sgn}(k) = +1$ if $\mathfrak{M} > 0$. If $\tau = 0$ *and* $\mathfrak{M} = 0$ (vanishing shear), we have a "moment of time symmetry",[29,30] for which $\text{sgn}(k) = 0$ is the only permissible case if M is closed.

To conclude this part of the discussion, we repeat that for $\mathfrak{M} > 0$, only cases involving $\tau = 0$ everywhere on M lead to any restrictions on the choice of conformal equivalence classes of initial data in the construction of solutions to the constraint equations.

In general, we see that when $\tau = \text{const.}$, (40) and (41) split into two separate problems. We give freely (M, g) and ψ^{ab}. We solve the linear elliptic Eqs. (10) for W^a, and thereby construct σ^{ab}_{TT}. Substituting g_{ab}, σ^{ab}_{TT}, and τ into (41), we find ϕ. The final initial data set satisfying the complete set of constraints is therefore $(M, \bar{g})$ with $\bar{g}_{ab} = \phi^4 g_{ab}$, and $\bar{\pi}^{ab}$ given by (39). This means that the initial-value problem on surfaces $\tau = \text{const.}$ is an uncoupled elliptic second-order system of four Eqs. (10) and (48) for four functions W^a and ϕ. Of these, the three Eqs. (10) are linear and (48) is quasilinear.

The conformal treatment of the initial-value equations can be generalized to cases where $\tau(x) \neq \text{const.}$ is a prescribed function. In place of the momentum conditions $\nabla_b\sigma^{ab} = 0$ for $\tau = \text{const.}$, one has

$$\nabla_b\sigma^{ab} = -\tfrac{1}{2}g^{1/2}\nabla^a\tau, \tag{54}$$

where σ^{ab} is still trace-free. We can view this problem as requiring the construction of a traceless tensor in purely *longitudinal* form $(LZ)^{ab}$. Equation (54) determines $(LZ)^{ab}$; however, *this* procedure is not independent of ϕ. Equations (54) and (48) are now coupled. On such surfaces the gravitational initial variables are not pure spin-two objects. The presence of $\nabla^a\tau$ introduces effectively a vector part Z to the complete set of initial variables. Of course, to the solutions σ^{ab} of (54) itself, may always be added a free (unconstrained) field in the form of some TT-variable. So, even in this case, TT

momenta describe free gravitational fields ("waves"). Moreover, regardless of the presence of $\nabla^a\tau \neq 0$, the *intrinsic* conformal geometry of the surface still corresponds to the pure spin-two part of the field. This is true because the three-dimensional conformal curvature tensor is *identically* TT regardless of the value of τ on the surface. Thus, the vacuum gravitational field behaves somewhat like an electromagnetic field with sources when $\tau \neq$ const.

We now wish to point out how matter or other field sources enter (48). The form of (44) is[32,33]

$$\bar{g}^{-1/2}(\bar{\pi}_{ab}\bar{\pi}^{ab} - \tfrac{1}{2}\bar{\pi}^2) - \bar{g}^{1/2}\bar{R} = -16\pi\bar{g}^{1/2}\bar{T}^*_*, \tag{55}$$

where $\bar{T}^*_* = \bar{T}^\mu{}_\nu u^\nu u_\mu$, $T^\mu{}_\nu$ = matter tensor, and u^μ is the unit timelike four-vector normal to the surface. Physically, $\bar{T}^*_*$ is the positive-definite scalar measuring the mass or energy per unit proper three-volume on the surface. The key question here is: Can $\bar{T}^*_*$ itself be freely specified, or only specified up to some conformal factor? Both dimensional arguments and arguments based on the free electromagnetic field[32] as source point to the fact that one can only give $\bar{T}^*_*$ *up to* a conformal factor. Thus, for example, the conformal properties of the electromagnetic field coupled to gravity using a "3 + 1" formalism identical to that of the present paper show that one may freely give T^*_*, where $\bar{T}^*_* = \phi^{-8}\bar{T}^*_*$. This choice is consistent with the decoupling of the momentum and energy constraints (see below) *and* with the fact that the *electromagnetic*-initial-value problem on the surface must simultaneously be satisfied.

In place of (48), one finds[32,33]

$$(8\Delta + R)\phi = \mathfrak{M}\phi^{-7} - \tfrac{3}{8}\tau^2\phi^5 + 16\pi T^*_*\phi^{-3}, \tag{56}$$

where all coefficients are known (see below), and where the sign of each coefficient on the right-hand-side is known. Again, one can show[25,32] the existence of a unique solution for all conformal equivalence classes of initial data, except those with $\tau = 0$ have to be treated separately, as above. Choquet-Bruhat[34] treated the case with $\tau = 0$ and $\bar{T}^*_* = T^*_*$, i.e., the energy density scalar *completely* specified in advance. However, this is not consistent with massless, integral spin sources nor with dimensional analysis. It is an attempt to specify *more* data about the sources that is consistent with the physics of gravity coupled to other fields. Not surprisingly, she found a number of serious restrictions on the existence and uniqueness of solutions for certain gravitational-matter configurations. Understanding these restrictions more deeply could be important in further elucidating the physical content of initial-value problems with sources.

We have already indicated how matter sources enter (48). Likewise, sources may be inserted into the momentum constraints in the form[11,32,33]

$$\nabla_b\sigma^{ab} = 8\pi g^{1/2}g^{ab}T^*_b \tag{57}$$

when $\tau =$ const. Here $G = c = 1$, T^μ_ν is the matter tensor, and $T^*_b = B^\nu_b u_\mu T^\mu_\nu$. The factor B^ν_b projects onto the surface, for which u^μ is the unit normal four-vector. The construction of a trace-free σ^{ab} satisfying (57) is again a conformally invariant problem *not* coupled to (48) (just as before) if we conformally map T^*_b by the transformation[32,35]

$$T^*_b \to \bar{T}^*_b = \phi^{-6}T^*_b. \tag{58}$$

This means that here $g^{1/2}T^*_b$ is freely specifiable, not T^*_b itself. Effectively, one may only prescribe the directional properties of the matter current T^*_b but not its absolute magnitude. Observe that T^*_b does *not* enter (56); only the solution σ^{ab} of (54) enters (56) and, of course, T^*_*. So none of the discussion of (56) has to be altered in this generalized procedure. Therefore, the inclusion of matter currents occasions no difficulties. Again, we may allow $\tau(x) \neq$ const., but the equations then become coupled, introducing, however, no difficulties in principle.

Lastly, we mention that there is an elliptic "sandwich" version of the initial value problem.[36] Here, one gives freely, on two nearby slices $\tau_1 =$ const., $\tau_2 =$ const., and two infinitesimally different conformal metrics. One finds ϕ and W^a by a method similar to our discussion above. The proper orthogonal distance N between the two surfaces is determined by an elliptic equation[11] for N, resulting from the demand that the τ's be two infinitesimally differing constants on the two surfaces. The "sandwich" problem becomes five elliptic equations for the five variables ϕ, W^a, and N. The W^a equations are linear and the ϕ and N equations are quasilinear. These five equations are coupled. Coupled elliptic systems are very hard to analyse, no sufficiently powerful mathematical theorems being readily available. However, this *particular* set has simplifying features that permit this "conformal thin-sandwich" result to be analyzed.

5. PROJECTION OPERATORS FOR THE TT-DECOMPOSITION

In this section, following a notation close to that of Berger and Ebin,[7] I shall denote the operators relevant to the TT-decomposition by

$$(\tilde{\delta}\psi)^a \equiv -\nabla_b(\psi^{ab} - \tfrac{1}{3}\psi g^{ab}) \equiv -[\nabla\cdot(\Lambda\psi)]^a, \tag{59}$$

where ψ^{ab} is a symmetric tensor. Omitting indices, we write the solution of (10) as

$$W = D^{-1}[\tilde{\delta}\psi]. \tag{60}$$

The longitudinal part of the tensor ψ is written

$$\psi_{\mathrm{L}} = LW = LD^{-1}[\tilde{\delta}\psi]. \tag{61}$$

The trace part is given by

$$\psi_{\mathrm{Tr}} = (I - \Lambda)\psi, \tag{62}$$

where I denotes the identity operator. Therefore,

$$\psi_{\mathrm{TT}} = (\Lambda - LD^{-1}\tilde{\delta})\psi. \tag{63}$$

The appropriate projection operators acting on ψ are thus

$$P_{\mathrm{TT}} = \Lambda - LD^{-1}\tilde{\delta}, \tag{64}$$

$$P_{\mathrm{L}} = LD^{-1}\tilde{\delta}, \tag{65}$$

$$P_{\mathrm{Tr}} = I - \Lambda, \tag{66}$$

where $I = P_{\mathrm{TT}} + P_{\mathrm{L}} + P_{\mathrm{Tr}}$.

The space of pure trace tensors may be written[37] $\Lambda^{-1}(0)$. The space of longitudinal tensors may be written $L(V^1)$, where V^1 denotes the space of vectors on M. We now seek a further characterization of the spaces of TT-tensors as defined in this paper.

First we note that the operator L^+ adjoint to L is $L^+ = \frac{1}{2}\tilde{\delta}$. The vector "Laplacian" D of equation (10) is given by

$$D = \tilde{\delta}L. \tag{67}$$

The dimension of the kernel $D^{-1}(0)$ for this operator equals the number of linearly independent conformal Killing vectors admitted by (M,g). This dimension is conformally invariant. Its maximum value in an N-space is $\frac{1}{2}(N+1)(N+2)$. The maximum is achieved if and only if the space is conformally flat.

Now we define the *tensor* "Laplacian" $\tilde{\tilde{\Delta}}$ by

$$\tilde{\tilde{\Delta}} = L\tilde{\delta}. \tag{68}$$

This is a linear, second-order, elliptic, positive-definite operator which vanishes if and only if the tensor it acts on is TT as defined in this paper. To see this, note that

$$(\psi, \tilde{\tilde{\Delta}}\psi) = 2(\tilde{\delta}\psi, \tilde{\delta}\psi), \tag{69}$$

which vanishes if and only if $\tilde{\delta}\psi = 0$. We can have $\tilde{\delta}\psi = 0$ if ψ is TT or if the tracefree part of ψ is transverse. However, according to our definition of TT tensors, if the tracefree part is transverse, then the longitudinal part must vanish and we have $\psi_{\mathrm{TT}} = \Lambda\psi$, i.e., the tracefree part *is* the TT part in this case. Hence, it follows that the space of TT tensors can be written as $\tilde{\delta}^{-1}(0)$ or $\tilde{\tilde{\Delta}}^{-1}(0)$. It follows also that these kernels are conformally invariant because the TT-property is preserved under the conformal mappings we have defined. Therefore, one sees by virtue of their conformal properties an interesting "duality" between the vector operator $D = \tilde{\delta}L$ and the tensor operator $\tilde{\tilde{\Delta}} = L\tilde{\delta}$.[38]

Again following the notation of Berger and Ebin,[7] we write the present splitting of the space of C^∞ symmetric tensor fields on (M,g) as

$$C^\infty(S^2) = \tilde{\tilde{\Delta}}^{-1}(0) \oplus L(V^1) \oplus \Lambda^{-1}(0), \tag{70}$$

where the summands are orthogonal.

Using different operators than those we have employed here, Berger and Ebin[7] achieved an orthogonal TT decomposition in the special case that $R = \text{const}$. Yamabe's theorem[27] should provide a link between their decomposition and the one defined in this paper.

6. DEFORMATIONS OF CONFORMAL RIEMANNIAN MANIFOLDS

Any infinitesimal deformation of a Riemannian geometry may be represented by a symmetric tensor δg_{ab}, where δ now stands for "variation," not for the negative divergence as in Sec. 5. Since any symmetric tensor may be decomposed by our procedure, we may write

$$\delta g_{ab} = \delta g^{\mathrm{TT}}_{ab} + (LW)_{ab} + \tfrac{1}{3} g_{ab} g^{cd} \delta g_{cd}. \tag{71}$$

On the other hand, by definition, any small variation of a conformal metric $\tilde{g}_{ab} = (\det g)^{-1/3} g_{ab}$ must be tracefree. Hence, the last term of (71) represents an infinitesimal conformal transformation which cannot affect the underlying conformal geometry.

Consider an infinitesimal shift of coordinates $x^a \to x^a - W^a$. The change induced in $\tilde{g}_{ab}$ is just

$$\pounds_W \tilde{g}_{ab} = (\det g)^{-1/3}(LW)_{ab}. \tag{72}$$

Hence, the second term on the right of (71) denotes a relabeling of coordinates in the underlying conformal geometry and thus to no change in the intrinsic conformal geometry itself. Thus, the only term of (71) corresponding to a variation of the intrinsic conformal geometry is $\delta g^{\mathrm{TT}}_{ab}$. Therefore, every deformation in conformal superspace $\tilde{\mathcal{S}}$ is represented by a TT tensor only, and conversely. Thus, in passing from one point of $\tilde{\mathcal{S}}$ to a neighboring point, only TT-tensors need be considered.

There is a method of constructing TT tensors that does not rely on any TT-decomposition as such. In three dimensions, the conformal curvature tensor density[18] is defined by

$$\tilde{\beta}^{ab} \equiv \tfrac{1}{2} (\det g)^{1/3} (\epsilon^{efa} g^{bm} + \epsilon^{efb} g^{am}) \nabla_e R_{fm}, \tag{73}$$

where ϵ^{efa} is the unit alternating tensor density. If $\delta g_{ab} = \lambda g_{ab}$, then one can readily show that $\delta\tilde{\beta}^{ab} = 0$. Let $\beta^{ab} = (\det g)^{-1/3}\tilde{\beta}^{ab}$. Then we also have $\delta\beta^a_b = 0$. $\tilde{\beta}^{ab}$ is identically symmetric and TT, giving rise to the conclusion that the conformal geometry represents the pure spin-2 part of the full Riemannian 3-geometry, as we described earlier.

If we construct from $\tilde{\beta}^{ab}$ a conformally invariant scalar density of weight one and integrate it over M, then the functional derivative of this integral with respect to g_{ab} will be automatically a TT density of a weight 1. The weight of this tensor can be changed at will, of course. As an example,

$$\mu^{ab} = \delta / \delta g_{ab} \int_M (\beta^c_a \beta^d_c)^{1/2} d^3x \tag{74}$$

is a TT tensor density of weight 1.[39] It is worthwhile to note that the functional derivation of a TT tensor density of coordinate weight 1, as in (74), may be readily used to prove that such an object *must* transform by the rule $\bar{\mu}^{ab} = \phi^{-4}\mu^{ab}$ under a conformal mapping $\bar{g}_{ab} = \phi^4 g_{ab}$. Thus the *tensor* form $\psi^{ab}_{\mathrm{TT}} = g^{-1/2}\mu^{ab}$ must transform as $\bar{\psi}^{ab}_{\mathrm{TT}} = \phi^{-10}\psi^{ab}_{\mathrm{TT}}$, as shown in Sec. 2.

ACKNOWLEDGMENTS

I thank V. Bargmann, S. Deser, R. Gowdy, D. Markiewicz, and especially N. O'Murchadha and R. Wald for discussions and helpful remarks.

APPENDIX: TT TENSORS AND SYMMETRIES

Let X^a be a CKV of g_{ab} and consider the tensor ψ^{TT}_{ab}. By using the well-known rules for interchanging $\pounds_X$ and ∇_b, one can easily verify that

$$\psi'^{\mathrm{TT}}_{ab} = \psi^{\mathrm{TT}}_{ab} + \pounds_X \psi^{\mathrm{TT}}_{ab} + \tfrac{1}{3}\psi^{\mathrm{TT}}_{ab}\nabla_c X^c \tag{A1}$$

is also TT with respect to g_{ab}. Similar results hold for ψ^{ab}_{TT} and for different weights. Therefore, although CKV's do not show up directly in the construction of ψ^{TT}_{ab}, as we explained, they do give automatically other TT tensors.

The *transverse* decomposition[6,7] of a tensor T_{ab} is defined by

$$T_{ab} = T^T_{ab} + \nabla_a V_b + \nabla_b V_a, \tag{A2}$$

for some unique V_a. If the metric g_{ab} possesses Killing vectors $Y(\pounds_Y g_{ab} = 0)$, then, similarly to the above,

$$T'^T_{ab} = T^T_{ab} + \pounds_Y T^T_{ab} \tag{A3}$$

is also automatically transverse with respect to g_{ab}. These results relate, respectively, to the "stratification" of $\tilde{\mathcal{S}}$, and to the "stratification" of $\mathcal{S}$.[8]

*Supported in part by National Science Foundation Grant GP30799X to Princeton University.

[1]See, for example, the article by R. Arnowitt, S. Deser, and C. W. Misner in *Gravitation,* edited by L. Witten (Wiley, New York, 1962).

[2]G. deRham, *Varietés Differentiables* (Hermann, Paris, 1955).

[3]For manifolds with indefinite metrics, such as spacetime, the decomposition treated in this paper may also be used. However, the equations to be solved become hyperbolic (for spacetime) instead of elliptic, necessitating a different choice of boundary conditions. The hyperbolic case will be treated elsewhere.

[4]See, for example, C. W. Misner, Phys. Rev. **186,** 1319 (1969); R. Gowdy, Phys. Rev. Lett. **27,** 826 (1971).

[5]The behavior of tensors at spacelike infinity may also be treated as done by R. Geroch, J. Math. Phys. **13,** 956 (1972), by adding a conformal "point at infinity", thereby compactifying the asymptotically flat space. This "conformal compactification" appears to be quite compatible with the present techniques.

[6]S. Deser, Ann. Inst. Henri Poincaré 7, 149 (1967). The present decomposition is similar to one of Deser, but has a different trace term. Our procedure is similar to one of Deser's transverse (not TT) decompositions [his equation (10)], but ours is applied to the *trace-free* part of a tensor T_{ab} whereas his is applied to T_{ab} itself. I thank Professor Deser for correspondence concerning this point.

[7]Cf. M. Berger and D. Ebin, J. Diff. Geom. 3, 379 (1969).

[8]A. E. Fischer, in *Relativity,* edited by M. Carmeli, S. I. Ficklin, and L. Witten (Plenum, New York, 1970).

[9]B. DeWitt, article in Ref. 8.

[10]R. F. Baierlein, D. Sharp, and J. A. Wheeler, Phys. Rev. **126,** 1864 (1962).

[11]J. W. York, Phys. Rev. Lett. **28,** 1082 (1972).

[12]P. A. M. Dirac, Phys. Rev. **114,** 924 (1959).

[13]K. Kuchař, J. Math. Phys. **13,** 768 (1972).

[14]Cf. B. DeWitt, Phys. Rev. **160,** 1113 (1967).

[15]Cf. C. W. Misner, article in *Magic without Magic,* edited by J. Klauder (Freeman, San Francisco, 1972).

[16]Cf. K. Kuchař, Phys. Rev. D **4,** 955 (1971).

[17]U. Gerlach, Phys. Rev. **177,** 1929 (1969).

[18]J. W. York, Phys. Rev. Lett. **26,** 1656 (1971).

[19]J. A. Wheeler, Rev. Mod. Phys. **34,** 873 (1962).

[20]I am grateful to R. Geroch for suggesting this term.

[21]Strictly speaking, one needs "self-adjointness." However, on closed C^∞ Riemannian manifolds, there is no necessary distinction for our purposes. See the results on elliptic operators in Ref. 7. For the relation in general, cf. F. Riesz and B. Sz.-Nagy, *Functional Analysis* (Ungar, New York, 1955).

[22]By "harmonic$_T$" we mean here simply the kernel of D, $D^{-1}(0)$, consisting of the linearly independent conformal Killing vectors (if any) on $(M,\tilde{g})$.

[23]O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasi-Linear Elliptic Equations* (Academic, New York, 1968).

[24]Vanishing ϕ means the Riemannian structure is singular, e.g., "pinched off." Here we are considering only nonsingular manifolds without boundary (or nonsingular manifolds with boundary only "at infinity").

[25]N. O'Murchadha and J. W. York, "Existence and Uniqueness of Solutions of the Hamiltonian Constraint," paper presented by J. W. Y. at the Boston Conference on Gravitation and Quantization, 2 November 1972 (to be published); N. O'Murchadha, thesis, January, 1973.

[26]D. R. Brill and S. Deser, Ann. Phys. (N.Y.) **50,** 548 (1968); and N. O'Murchadha and J. W. York, "Positive Energy in General Relativity" (to be published).

[27]H. Yamabe, Osaka Math. J. 12, 21 (1960). That Yamabe's original proof is incomplete was noticed by N. S. Trudinger, Annal. Scuola di Pisa, Ser. 3, 22, 265 (1968), who gave a partial proof of the theorem, and by T. Aubin, C.R. Acad. Sci. A **266,** 69 (1968), who gave a new proof of Yamabe's theorem. This theorem is quite interesting, but is not assumed in the present considerations. There is no analogous theorem for nonclosed manifolds.

[28]R. Gowdy, "Vacuum Spacetimes with Two-Parameter Spacelike Isometries: Topologies and Boundary Conditions", to be published (1972).

[29]Cf. J. A. Wheeler's discussion of D. R. Brill's treatment of the time-symmetric initial-value problem in *Relativity, Groups and Topology,* edited by C. DeWitt and B. DeWitt (Gordon and Breach, New York, 1964).

[30]Y. Choquet-Bruhat discussed mathematical aspects of the case $\tau=0$ in C.R. Acad. Sci. A **274,** 682 (1972).

[31]J. W. York, "General Decompositions of Symmetric Tensors" (to be published).

[32]N. O'Murchadha and J. W. York, "Coupled Gravitational-Electromagnetic Initial-Value Problem" (to be published).

[33]In this reference, matter tensor "currents" were not included, but they can be included without difficulty, as pointed out in Ref. 11. See the discussion of (57) below.

[34]See Ref. 30.

[35]The transformation (58) follows from a dimensional analysis or from consideration of electromagnetism as a source for gravity (Ref. 32).

[36]J. W. York, "Conformal Thin-Sandwich Theorem" (to be published).

[37]If O is an operator, $O^{-1}(0)$ denotes the kernel of the operator O.

[38]An analogous duality exists whenever one decomposes a symmetric tensor of any valence r in terms of a transverse part (or TT part) and a longitudinal part, where the longitudinal part is defined in terms of a symmetric tensor of valence $(r-1)$. For example, in decomposing a symmetric tensor T^{abc}, the elliptic operator defining the "tensor potential" P^{ab} [analogous to W^a in (10)] has as its kernel only "Killing tensors" C^{ab}, satisfying $\nabla^{(a}C^{bc)}=0$, if any exist on the given (M,g).

[39]It is probably well known that the functional derivative with respect to the metric of a conformally invariant functional is TT. For example, one may obtain the "contracted Bianchi identities" in this way for a sourceless Maxwell field coupled to the Einstein gravitational field.

Classification of space-times in general relativity

Ronald J. Adler and Charles Sheffield

The American University, Washington, D.C. 20016
(Received 21 July 1972)

The relationship between the Petrov classification of Riemannian space-times and the Debever-Penrose principal null directions is developed in explicit algebraic terms.

1. INTRODUCTION

The Petrov classification of Riemannian space-time in terms of its algebraic properties at a point offers an invariant (coordinate independent) description of aspects of space-time.[1] The relationship between the Petrov classification and the Debever-Penrose principal null directions at a point of space-time was first given in full by Penrose in his classic paper on the use of spinors in general relativity.[2] Subsequent articles and papers[3,4] have followed Penrose's treatment, which is essentially geometrical and does not display an explicit algebraic relationship between the two classification systems.

It is our purpose in this paper to develop the relationship between the Petrov classification and the Debever-Penrose directions in explicit algebraic terms. We show that the Petrov approach and Penrose's spinor approach each leads to a fundamental 3×3 matrix, and that these two matrices are related to each other by a similarity transformation. Further, we give simple explicit forms for the curvature eigenspinors in the Penrose formulation and use these to classify the space-times.

For convenience and simplicity, the analysis is carried out for the case of a zero matter tensor. The nonvacuum case is very similar, but the Weyl tensor must then be used instead of the Riemann tensor.

2. THE PETROV CLASSIFICATION

Consider the Riemann tensor $R_{\mu\nu\rho\lambda}$, which is antisymmetric in μ, ν and in ρ, λ, and is symmetric in the pairs $\mu\nu, \rho\lambda$.

We map the indices μ, ν, ρ, λ onto indices $\bar{A}, \bar{B}$ by the homomorphism

$$\begin{bmatrix} \mu\nu: & 23 & 31 & 12 & 10 & 20 & 30 \\ \bar{A}: & 1 & 2 & 3 & 4 & 5 & 6 \end{bmatrix}.$$

In terms of $\bar{A}$ and $\bar{B}$, $R_{\bar{A}\bar{B}}$ is a second order symmetric tensor with indices running from 1 to 6, and it can therefore be written as a 6×6 matrix of the form

$$R_{\bar{A}\bar{B}} = \begin{bmatrix} M & N \\ N^T & Q \end{bmatrix}, \tag{1}$$

where M and Q are real 3×3 symmetric matrices and N is a real 3×3 traceless matrix. We assume the use of real coordinates and a space-time signature $(+,-,-,-)$. The form of (1) follows directly from the symmetries of the Riemann tensor, and $R_{\bar{A}\bar{B}}$ is isomorphic to the tensor $R_{\mu\nu\rho\lambda}$.

Raising the first index on $R_{\bar{A}\bar{B}}$ and employing a local Lorentz coordinate frame leads to the form

$$R^{\bar{A}}{}_{\bar{B}} = \begin{bmatrix} M & N \\ -N^T & -Q \end{bmatrix}. \tag{2}$$

For the vacuum case, the field equations lead to the additional relations

$$Q = -M, \quad N = N^T, \quad \mathrm{Tr}(M) = 0. \tag{3}$$

In this case we therefore have the form

$$R^{\bar{A}}{}_{\bar{B}} = \begin{bmatrix} M & N \\ -N & M \end{bmatrix}, \tag{4}$$

where M and N are both real symmetric trace-free 3×3 matrices.

We now introduce the dual of the Riemann tensor

$$*R^{\mu\nu}{}_{\rho\lambda} = \tfrac{1}{2}\sqrt{-g}\, R^{\mu\nu}{}_{\gamma\delta}\epsilon^{\gamma\delta\alpha\beta} g_{\alpha\rho} g_{\beta\lambda}, \tag{5}$$

where $\epsilon^{\gamma\delta\alpha\beta}$ is the usual unit tensor density. The general notation follows that of Ref. 5. We observe that mapping μ, ν, ρ, λ to $\bar{A}, \bar{B}$ as before and applying the field equations yields the form

$$*R^{\bar{A}}{}_{\bar{B}} = \begin{bmatrix} N & -M \\ M & N \end{bmatrix}. \tag{6}$$

This result may be derived easily by element-by-element evaluation in the local Lorentz frame.

Thus, defining a self-dual tensor $R^{(+)\bar{A}}{}_{\bar{B}}$, we have

$$R^{(+)\bar{A}}{}_{\bar{B}} = R^{\bar{A}}{}_{\bar{B}} + i*R^{\bar{A}}{}_{\bar{B}} = \begin{bmatrix} M + iN & -i(M + iN) \\ i(M + iN) & M + iN \end{bmatrix}. \tag{7}$$

This can be written in the direct product form

$$R^{(+)\bar{A}}{}_{\bar{B}} = \begin{bmatrix} P & -iP \\ iP & P \end{bmatrix} = P \otimes J, \tag{8}$$

where $P = M + iN$ and $J = \begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix}$. P is a complex 3×3 traceless symmetric matrix and has ten algebraically independent components.

The eigenvectors of $R^{(+)\bar{A}}{}_{\bar{B}}$ are the direct products of the eigenvectors of P and J, and the eigenvalues are the algebraic products of those of P and J. For J, we have eigenvalues 0 and 2 and corresponding eigenvectors $(1, -i)$ and $(1, i)$. Thus, $R^{(+)\bar{A}}{}_{\bar{B}}$ has at least 3 zero eigenvalues. For the Petrov classification, it is necessary to consider only the eigenvalues and eigenvectors of $R^{(+)\bar{A}}{}_{\bar{B}}$ constructed using the nonzero eigenvalue of J. These eigenvalues of $R^{(+)\bar{A}}{}_{\bar{B}}$ may or may not be zero.

The Petrov classification corresponds to the following simple set of statements about the eigenvectors and eigenvalues of the matrix P:

(a) If P has 3 distinct eigenvectors and 3 distinct eigenvalues, space-time is Petrov Type I.

(b) If P has 3 distinct eigenvectors and 2 eigenvalues equal, space-time is Petrov Type I-D.

(c) If P has 2 distinct eigenvectors and 2 distinct eigenvalues, space–time is Petrov Type II.

(d) If P has 2 distinct eigenvectors but has equal (and hence zero, since P is traceless) eigenvalues, space–time is Petrov Type II-N.

(e) If P has only one eigenvector, space–time is Petrov Type III.

Equivalently, we can state the Petrov classification in terms of the elementary divisors of P:

(a) If P has linear elementary divisors and 3 distinct eigenvalues, the Petrov classification at the point of space–time concerned is *Petrov Type I*.

(b) If P has linear elementary divisors and has 2 equal eigenvalues, space–time is *Petrov Type I-D*.

(c) If P has just one linear elementary divisor and has 2 distinct eigenvalues, it is *Petrov Type II*.

(d) If P has just one linear elementary divisor and 2 equal eigenvalues, it is *Petrov Type II-N*.

(e) If P has no linear elementary divisors, it is *Petrov Type III*.

Finally, the Petrov classification is sometimes stated in terms of the Segre characteristic of the matrix[6] P:

Type	Segre characteristic
I	(1, 1, 1)
I-D	((1, 1), 1)
II	(2, 1)
II-N	((2, 1))
III	(3)

The Petrov classification provides a meaningful way of categorizing space–times because it is independent of the choice of coordinates used in setting up the Riemann tensor. For, consider the coordinate transformation $x \to \bar{x}$, so

$$\bar{R}^{\alpha\beta}{}_{\gamma\delta} = \frac{\partial \bar{x}^\alpha}{\partial x^\rho}\frac{\partial \bar{x}^\beta}{\partial x^\lambda}\frac{\partial x^\mu}{\partial \bar{x}^\gamma}\frac{\partial x^\nu}{\partial \bar{x}^\delta} R^{\rho\lambda}{}_{\mu\nu}. \tag{9}$$

With the index mappings $\binom{\alpha,\beta}{\rho,\lambda}\to\binom{A}{L}$; $\binom{\gamma,\delta}{\mu,\nu}\to\binom{C}{M}$, the coordinate transformation (9) can be written as a matrix transformation of R:

$$\bar{R}^A{}_C = T^A{}_L \bar{T}^M{}_C R^L{}_M, \tag{10}$$

where

$$T^A{}_L = \frac{\partial \bar{x}^\alpha}{\partial x^\rho}\frac{\partial \bar{x}^\beta}{\partial x^\lambda} \quad \text{and} \quad \bar{T}^M{}_C = \frac{\partial x^\mu}{\partial \bar{x}^\gamma}\frac{\partial x^\nu}{\partial \bar{x}^\delta}. \tag{11}$$

We can write this as a matrix product:

$$\bar{R} = TR\bar{T} \tag{12}$$

But, $T\bar{T} = T^A{}_L \bar{T}^L{}_C = \delta^\alpha{}_\gamma \delta^\beta{}_\delta$. Thus $T = T^{-1}$, and T is a similarity transformation on R.

The eigenvalues and the number of linear elementary divisors of a matrix are unchanged under a similarity transformation. Thus, *the Petrov classification is invariant under coordinate transformations*.

The 3×3 complex matrix P is the fundamental matrix of the Petrov classification. We will show that P is related by a similarity transformation to another 3×3 complex matrix arising from a spinor approach, and thus either matrix can be used as the basis for the Petrov classification.

3. THE SPINOR APPROACH

The classification of space–times in terms of the Debever–Penrose principal null directions is most readily approached using a spinor treatment. We will follow the notation used in Refs. 2 and 3, where the proofs may be found of many of the necessary spinor results that we will quote and use here without proof.

Tensor and spinor quantities are transformed to each other via the relationships

$$\begin{aligned} T^{A\dot{B}}{}_{C\dot{D}} &= \sigma_\lambda{}^{A\dot{B}}\, T^\lambda{}_\nu \sigma^\nu{}_{C\dot{D}} \\ T^\lambda{}_\nu &= \sigma^\lambda{}_{A\dot{B}} T^{A\dot{B}}{}_{C\dot{D}} \sigma_\nu{}^{C\dot{D}} \end{aligned} \tag{13}$$

where the quantities $\sigma_\lambda{}^{A\dot{B}}$ satisfy the equation

$$\sigma_\mu{}^{A}{}_{\dot{C}}\sigma_\nu{}^{B\dot{C}} + \sigma_\nu{}^{A}{}_{\dot{C}}\sigma_\mu{}^{B\dot{C}} = g_{\mu\nu}\epsilon^{AB}. \tag{14}$$

The σ's relate spinor space to tensor space, and ϵ^{AB} is a skew-symmetric metric spinor for the two-dimensional complex spinor space, enabling spinor indices to be raised and lowered. Spinor indices take on the values 1 and 2.

Equations (13) extend easily to any number of tensor indices. In a tangent flat space with signature $(+,-,-,-)$ a suitable set of σ's consists of multiples of the Pauli spin matrices, plus the 2×2 identity matrix.

To any second rank tensor $F_{\mu\nu}$ there corresponds a four-index spinor $F_{A\dot{B}C\dot{D}}$. If $F_{\mu\nu}$ is skew, $F_{A\dot{B}C\dot{D}}$ can be shown to have the unique spinor decomposition

$$F_{A\dot{B}C\dot{D}} = (\phi_{AC}\epsilon_{\dot{B}\dot{D}} + \epsilon_{AC}\psi_{\dot{B}\dot{D}}), \tag{15}$$

where ϕ_{AC} and $\psi_{\dot{B}\dot{D}}$ are symmetric in their spinor indices. Further, if $F_{\mu\nu}$ is real it may be shown that $\psi_{\dot{B}\dot{D}} = \bar{\phi}_{\dot{B}\dot{D}}$. Note that (15) differs from Penrose's definitions of ϕ and ψ by a factor of 2.

If we apply (15) to the Riemann tensor $R_{\mu\nu\rho\lambda}$, which is skew in both pairs of indices $\mu\nu$ and $\rho\lambda$, the corresponding 8-index spinor may be shown to have the form

$$\begin{aligned} R_{\mu\nu\rho\lambda} \leftrightarrow R_{A\dot{E}B\dot{F}C\dot{G}D\dot{H}} = (&\chi_{ABCD}\epsilon_{\dot{E}\dot{F}}\epsilon_{\dot{G}\dot{H}} + \epsilon_{CD}\phi_{AB\dot{G}\dot{H}}\epsilon_{\dot{E}\dot{F}} \\ &+ \epsilon_{AB}\bar{\phi}_{\dot{E}\dot{F}CD}\epsilon_{\dot{G}\dot{H}} + \epsilon_{AB}\epsilon_{CD}\bar{\chi}_{\dot{E}\dot{F}\dot{G}\dot{H}}). \end{aligned} \tag{16}$$

The symmetry of $R_{\mu\nu\rho\lambda}$ in the pairs $\mu\nu$ and $\rho\lambda$ leads to the symmetries on χ:

$$\chi_{ABCD} = \chi_{BACD} = \chi_{ABDC} = \chi_{CDAB}. \tag{17}$$

In the vacuum case, applying the field equations leads to the additional conditions

$$\phi_{AB\dot{C}\dot{D}} = 0, \qquad \chi_{ABCD} = \chi_{ADCB}. \tag{18}$$

Thus, χ_{ABCD} is *totally symmetric* in all indices, and Eq. (16) reduces to the form

$$R_{\mu\nu\rho\lambda} \leftrightarrow R_{A\dot{E}B\dot{F}C\dot{G}D\dot{H}} = (\chi_{ABCD}\epsilon_{\dot{E}\dot{F}}\epsilon_{\dot{G}\dot{H}} + \epsilon_{AB}\epsilon_{CD}\bar{\chi}_{\dot{E}\dot{F}\dot{G}\dot{H}}). \tag{19}$$

The dual tensor ${}^*R_{\mu\nu\rho\lambda}$ can be shown to have the spinor equivalent form

$${}^*R_{\mu\nu\rho\lambda} \leftrightarrow {}^*R_{A\dot{E}B\dot{F}C\dot{G}D\dot{H}} = i(-\chi_{ABCD}\epsilon_{\dot{E}\dot{F}}\epsilon_{\dot{G}\dot{H}} + \epsilon_{AB}\epsilon_{CD}\bar{\chi}_{\dot{E}\dot{F}\dot{G}\dot{H}}). \tag{20}$$

Thus the tensor $R^{(+)}{}_{\mu\nu\rho\lambda}$ has the spinor equivalent form

$$R^{(+)}{}_{\mu\nu\rho\lambda} \leftrightarrow R^{(+)}{}_{A\dot{E}B\dot{F}C\dot{G}D\dot{H}} = 2\chi_{ABCD}\epsilon_{\dot{E}\dot{F}}\epsilon_{\dot{G}\dot{H}}. \tag{21}$$

Using (13), we have explicitly

$$R^{(+)}{}_{\mu\nu\rho\lambda} = \sigma_{\mu}{}^{A\dot{E}}\sigma_{\nu}{}^{B\dot{F}}\sigma_{\rho}{}^{C\dot{G}}\sigma_{\lambda}{}^{D\dot{H}}(2\chi_{ABCD}\epsilon_{\dot{E}\dot{F}}\epsilon_{\dot{G}\dot{H}}). \tag{22}$$

or, raising the μ and ν indices,

$$R^{(+)\mu\nu}{}_{\rho\lambda} = \sigma^{\mu}{}^{A\dot{E}}\sigma^{\nu B\dot{F}}\sigma_{\rho}{}^{C\dot{G}}\sigma_{\lambda}{}^{D\dot{H}}(2\chi_{ABCD}\epsilon_{\dot{E}\dot{F}}\epsilon_{\dot{G}\dot{H}}). \tag{23}$$

Since $R^{(+)\mu\nu}{}_{\rho\lambda}$ is antisymmetric in μ, ν and in ρ, λ, we write (23) in the explicitly antisymmetrized form

$$R^{(+)\mu\nu}{}_{\rho\lambda} = \Sigma^{\mu\nu AB}\chi_{ABCD}\Sigma_{\rho\lambda}{}^{CD} \equiv \Sigma^{\mu\nu}{}_{AB}\chi^{AB}{}_{CD}\Sigma_{\rho\lambda}{}^{CD}, \tag{24}$$

where in (24) we define $\Sigma^{\mu\nu}{}_{AB}$ and $\Sigma_{\rho\lambda}{}^{CD}$ to be

$$\begin{aligned}\Sigma^{\mu\nu}_{AB} &= 1/\sqrt{2}\,(\sigma^{\mu}{}_{A}{}^{\dot{E}}\sigma^{\nu}{}_{B}{}^{\dot{F}} - \sigma^{\nu}{}_{A}{}^{\dot{E}}\sigma^{\mu}{}_{B}{}^{\dot{F}})\epsilon_{\dot{E}\dot{F}},\\ \Sigma^{CD}_{\rho\lambda} &= 1/\sqrt{2}\,(\sigma_{\rho}{}^{C\dot{G}}\sigma_{\lambda}{}^{D\dot{H}} - \sigma_{\lambda}{}^{C\dot{G}}\sigma_{\rho}{}^{D\dot{H}})\epsilon_{\dot{G}\dot{H}}.\end{aligned} \tag{25}$$

Now we again introduce the tensor index mapping $\mu, \nu \to \bar{A}, \rho, \lambda \to \bar{B}$, as in Eq. (1). Bars distinguish the matrix and spinor indices, and we have from (24)

$$R^{(+)\bar{A}}{}_{\bar{B}} = \Sigma^{\bar{A}}{}_{AB}\chi^{AB}{}_{CD}\Sigma^{CD}{}_{\bar{B}} \tag{26}$$

Now, from Eq. (8) we note that the upper left 3×3 submatrix of $R^{(+)\bar{A}}{}_{\bar{B}}$ is the matrix P. Thus, if we restrict the range of the indices $\bar{A}$ and $\bar{B}$ to 1, 2 and 3, Eq. (26) gives

$$P^{\bar{A}}{}_{\bar{B}} = \Sigma^{\bar{A}}{}_{AB}\chi^{AB}{}_{CD}\Sigma^{CD}{}_{\bar{B}}, \tag{27}$$

where in (27), $\bar{A}$ and $\bar{B}$ take on the values 1, 2 and 3, and A, B, C and D take on the values 1 and 2.

Equation (27) is an equation relating a 3×3 array, P, to a 4×4 array, χ. In general, the four-index spinor will have four eigenspinors, whereas the matrix P can have at most three eigenvectors. Thus it would seem at first sight that χ is not directly relatable to the Petrov classification given earlier and based on the properties of P.

However, since χ is symmetric in all spinor indices, *any* antisymmetric two-index spinor, z^{CD}, is an eigenspinor of χ, with zero eigenvalue, since

$$\chi_{ABCD}z^{CD} \equiv 0. \tag{28}$$

In addition, it can be shown that any antisymmetric two-index spinor can be written as a scalar multiple of the skew metric spinor ϵ^{AB}. Thus we need consider only the symmetric eigenspinors y^{AB} of χ that are at most three in number and satisfy

$$\chi_{ABCD}y^{CD} = \lambda \cdot y_{AB}. \tag{29}$$

Then, since y and χ are both symmetric in their spinor indices, the 4×4 system for the eigenspinors of χ can be written as a 3×3 system:

$$X_{ij}Y^{j} = \lambda \cdot Y_{i}, \tag{30}$$

where (29) and (30) are related by

$$Y_1 = y_{11}, \quad Y_2 = \sqrt{2}y_{12} = \sqrt{2}y_{21}, \quad Y_3 = y_{22},$$

and

$$\begin{aligned}&X_{11} = \chi_{1111}, \quad X_{12} = X_{21} = \sqrt{2}\,\chi_{1112},\\ &X_{13} = X_{31} = \chi_{1122}, \quad X_{22} = 2\chi_{1122},\\ &X_{23} = X_{32} = \sqrt{2}\,\chi_{1222}, \quad X_{33} = \chi_{2222}.\end{aligned} \tag{31}$$

Using (31) in (29) yields (30) by direct substitution. Thus, the eigenvectors of the 3×3 matrix X are simply related to and in one-to-one correspondence with the symmetric eigenspinors of the four-index spinor, χ. Note that X is a symmetric matrix, as is P. Equations (29) and (30) may be related by the index mapping

$$\begin{bmatrix} AB & 11 & (12) & 22 \\ \downarrow & & & \\ i & 1 & 2 & 3 \end{bmatrix}, \tag{32}$$

and we may then write (27) in the form

$$P^{\bar{A}}{}_{\bar{B}} = \Sigma^{\bar{A}}{}_{i}X^{i}{}_{j}\Sigma^{j}{}_{\bar{B}}, \tag{33}$$

where now all indices run from 1 to 3 and both P and X are traceless symmetric complex 3×3 matrices. (Note: X is symmetric only if all indices are up or all are down).

We now show that Eq. (33) is actually a similarity transformation between P and X. Thus, either P or X may be used in performing the Petrov classification in terms of their eigenvalues and eigenvectors. Equally, because of the relation between X and χ, the Petrov classification can be done using the symmetric eigenspinors and eigenvalues of χ.

The condition for (33) to be a similarity transformation is

$$\Sigma^{\bar{A}}{}_{i}\Sigma^{i}{}_{\bar{B}} = \delta^{\bar{A}}{}_{\bar{B}}. \tag{34}$$

This is most easily checked by again using the local Lorentz frame, in which the σ matrices entering the definition of Σ can be chosen as Pauli spin matrices.

Note that, since $\bar{A}$ and $\bar{B}$ range only over the values 1, 2, and 3, the index 0 is not used in the set μ, ν, ρ, λ. Thus we need only σ_1, σ_2, and σ_3, for which we take the representations

$$\sigma_1 = 1/\sqrt{2}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \quad \sigma_2 = 1/\sqrt{2}\begin{pmatrix}0 & -i\\ i & 0\end{pmatrix}, \sigma_3 = 1/\sqrt{2}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \tag{35}$$

Using the definitions (25), we have

$$\Sigma^{\bar{A}}{}_{i}\Sigma^{i}{}_{\bar{B}} = \tfrac{1}{2}(\sigma^{\mu}{}_{A\dot{F}}\sigma^{\nu}{}_{B}{}^{\dot{F}} - \sigma^{\nu}{}_{A\dot{F}}\sigma^{\mu}{}_{B}{}^{\dot{F}})(\sigma_{\rho}{}^{A}{}_{\dot{H}}\sigma_{\lambda}{}^{B\dot{H}} - \sigma_{\lambda}{}^{A}{}_{\dot{H}}\sigma_{\rho}{}^{B\dot{H}}). \tag{36}$$

Consider one term of this product. Using standard spinor properties, we have

$$\begin{aligned}\sigma^{\mu}{}_{A\dot{F}}\sigma^{\nu}{}_{B}{}^{\dot{F}}\sigma_{\rho}{}^{A}{}_{\dot{H}}\sigma_{\lambda}{}^{B\dot{H}} &= \sigma^{\mu P\dot{Q}}\epsilon_{PA}\epsilon_{\dot{Q}\dot{F}}\sigma^{\nu R\dot{F}}\epsilon_{RB}\sigma_{\rho}{}^{AS}\epsilon_{S\dot{H}}\sigma_{\lambda}{}^{B\dot{H}}\\ &= \mathrm{Tr}(\sigma_{\rho}\epsilon\bar{\sigma}_{\lambda}\epsilon\sigma^{\nu}\epsilon\bar{\sigma}^{\mu}\epsilon).\end{aligned} \tag{37}$$

In (37) we have used $\sigma^{\mu P\dot{Q}} = \bar{\sigma}^{\mu}{}_{\dot{Q}P}$, and we note that since the trace is taken in spinor space, tensor indices may be up or down. To evaluate the trace, we use the fact that with the representations of the σ-matrices used in (35), $\epsilon\bar{\sigma}_{\lambda}\epsilon = \sigma_{\lambda}$.

Thus,

$$\mathrm{Tr}(\sigma_{\rho}\epsilon\bar{\sigma}_{\lambda}\epsilon\sigma^{\nu}\epsilon\bar{\sigma}^{\mu}\epsilon) = \mathrm{Tr}(\sigma_{\rho}\sigma_{\lambda}\sigma^{\nu}\sigma^{\mu}).$$

Now using the relation, well-known for the Pauli spin matrices,

$$\sigma_\mu\sigma_\nu = \tfrac{1}{2}\sigma_{\mu\nu} + i/\sqrt{2}\,\epsilon_{\mu\nu\lambda}\sigma_\lambda \tag{38}$$

we find

$$\mathrm{Tr}(\sigma_\rho\sigma_\lambda\sigma^\nu\sigma^\mu) = \tfrac{1}{2}(\delta_{\rho\lambda}\delta^{\mu\nu} - \delta_\rho{}^\nu\delta_\lambda{}^\mu + \delta_\rho{}^\mu\delta_\lambda{}^\nu). \tag{39}$$

Performing the same calculation for each term of (36) and combining the results, we find

$$\Sigma^{\bar{A}}{}_i\Sigma^i{}_{\bar{B}} = (\delta_\lambda{}^\nu\delta_\rho{}^\mu - \delta_\lambda{}^\mu\delta_\rho{}^\nu) = \delta^{\bar{A}}{}_{\bar{B}}. \tag{40}$$

Note that since we require that $\delta^{\bar{A}}{}_{\bar{B}} = -1$ when $\bar{A} = (\mu,\nu)$ and $\bar{B} = (\nu,\mu)$, Eq. (40) is the appropriate form for $\delta^{\bar{A}}{}_{\bar{B}}$. Equation (40) confirms that the relation (27) is indeed a similarity transformation.

4. THE EIGENSPINORS OF χ AND THE DEBEVER-PENROSE DIRECTIONS

We have shown that the Petrov classification may be described in terms of the properties of a complex traceless symmetric matrix P, and that this classification is invariant under a similarity transformation on P. Further, we have shown that P is related by a similarity transformation to a matrix X, and that the eigenvalues and eigenvectors of X are in one-to-one correspondence with the symmetric eigenspinors and eigenvalues of the wholly symmetric four-index spinor χ. It now remains only to determine the symmetric eigenspinors and eigenvalues of χ and relate these to the Debever–Penrose directions, in order to complete the relationship of the Petrov classification to the Debever–Penrose description of space-time.

First, since χ_{ABCD} is symmetric in all indices, it follows that there exists a unique decomposition of χ into a symmetrized product of one-index spinors (see Refs. 2 and 3), thus

$$\chi_{ABCD} = k_{(A}m_B r_C s_{D)}. \tag{41}$$

It can also be shown that each single index spinor k_A defines a null vector in Riemannian space, thus the spinor χ in general defines a set of up to 4 such null vectors. To relate these null vectors, which are the Debever–Penrose principal null directions, to the Petrov classification, we need to obtain the symmetric eigenspinors of χ_{ABCD} in terms of the single index spinor factors of χ.

To find these eigenspinors, let us consider the product $\chi_{ABCD}k^{(C}m^{D)}$. This can be written

$$\begin{aligned}\chi_{ABCD}k^{(C}m^{D)} = [&k_{(A}m_{B)}r_{(C}s_{D)} + k_{(A}r_{B)}m_{(C}s_{D)}\\ &+ k_{(A}s_{B)}m_{(C}r_{D)} + r_{(A}m_{B)}k_{(C}s_{D)}\\ &+ s_{(A}m_{B)}k_{(C}r_{D)} + r_{(A}s_{B)}k_{(C}m_{D)}]k^{(C}m^{D)}.\end{aligned} \tag{42}$$

Noting that $k_As^A = k_1s_2 - k_2s_1 = -s_Ak^A$, etc., we write k_As^A as $k\wedge s$ and contract on C and D in (42) to give

$$\begin{aligned}\chi_{ABCD}k^{(C}m^{D)} = {}&k_{(A}m_{B)}(r\wedge ks\wedge m + r\wedge ms\wedge k)\\ &+ k_{(A}r_{B)}m\wedge ks\wedge m + k_{(A}s_{B)}m\wedge kr\wedge m + r_{(A}m_{B)}m\wedge kk\wedge s\\ &+ s_{(A}m_{B)}m\wedge kk\wedge r + r_{(A}s_{B)}m\wedge kk\wedge m.\end{aligned} \tag{43}$$

To simplify (43), we need the following spinor identity:

$$\sum_P P(k_{(A}m_B)r_{[C}s_{D]}) = 0. \tag{44}$$

This indicates that the sum over all cyclic permutations of the 1-index spinors k, m, r and s, symmetrized and antisymmetrized as indicated, is identically zero.

It is most easily proved by symmetry arguments. In particular, taking $C = 1$ and $D = 2$, (44) gives

$$\sum_P P(k_{(A}m_{B)}r\wedge s) = 0. \tag{45}$$

To make use of this result, we regroup terms in (43) and write

$$\begin{aligned}\chi_{ABCD}k^{(C}m^{D)} = {}&k_{(A}m_{B)}(r\wedge ks\wedge m + r\wedge ms\wedge k)\\ &+ r_{(A}s_{B)}m\wedge kk\wedge m + m\wedge k(k_{(A}r_{B)}s\wedge m + s_{(A}m_{B)}k\wedge r)\\ &+ m\wedge k(k_{(A}s_{B)}r\wedge m + r_{(A}m_{B)}k\wedge s).\end{aligned} \tag{46}$$

Now applying (45) with the symbols ordered as k, r, s, m to the third term of (46) and applying it with the symbols ordered as k, s, r, m to the fourth term of (46), we find

$$\begin{aligned}\chi_{ABCD}k^{(C}m^{D)} = {}&k_{(A}m_{B)}(r\wedge ks\wedge m + r\wedge ms\wedge k)\\ &+ 3r_{(A}s_{B)}m\wedge kk\wedge m.\end{aligned} \tag{47}$$

Similarly, we find

$$\begin{aligned}\chi_{ABCD}r^{(C}s^{D)} = {}&r_{(A}s_{B)}(r\wedge ks\wedge m + r\wedge ms\wedge k)\\ &+ 3k_{(A}m_{B)}s\wedge rr\wedge s.\end{aligned} \tag{48}$$

Taking linear combinations of (47) and (48) we at once obtain the eigenspinors, as

$$y^{\pm}{}_{AB} = k_{(A}m_{B)}r\wedge s \pm r_{(A}s_{B)}k\wedge m \tag{49}$$

with eigenvalues

$$\lambda^{\pm} = 2(r\wedge ks\wedge m + r\wedge ms\wedge k) \mp 6k\wedge mr\wedge s. \tag{50}$$

Since we could have equally well have used $k^{(C}r^{D)}$ or $k^{(C}s^{D)}$ in equation (42), we find as the full set of symmetric eigenspinors

$$\begin{aligned}y_1{}^{\pm} &= k_{(A}m_{B)}r\wedge s \pm r_{(A}s_{B)}k\wedge m,\\ y_2{}^{\pm} &= k_{(A}r_{B)}m\wedge s \pm m_{(A}s_{B)}k\wedge r,\\ y_3{}^{\pm} &= k_{(A}s_{B)}r\wedge m \pm r_{(A}m_{B)}k\wedge s,\end{aligned} \tag{51}$$

with corresponding eigenvalues:

$$\begin{aligned}\lambda_1{}^{\mp} &= 2(r\wedge ks\wedge m + r\wedge ms\wedge k) \mp 6k\wedge mr\wedge s,\\ \lambda_2{}^{\mp} &= 2(m\wedge ks\wedge r + m\wedge rs\wedge k) \mp 6k\wedge rm\wedge s,\\ \lambda_3{}^{\mp} &= 2(r\wedge km\wedge s + r\wedge sm\wedge k) \mp 6k\wedge sr\wedge m.\end{aligned} \tag{52}$$

The form of (51) suggests that there are six symmetric eigenspinors, but using (45) again reveals at once that $y_1{}^+ = y_3{}^+$, $y_1{}^- = y_2{}^-$, and $y_2{}^+ = -y_3{}^-$. We thus have at most three symmetric eigenspinors, which we choose as $y_1 = y_1{}^+$, $y_2 = y_2{}^-$ and $y_3 = y_3{}^-$, with corresponding eigenvalues $\lambda_1{}^-$, $\lambda_2{}^+$ and $\lambda_3{}^+$.

If now k, m, r and s are all distinct one-index spinors, then we have three independent symmetric eigenspinors of χ, and four distinct Debever-Penrose principal null directions.

However, since k, m, r, and s can coincide in various ways it is necessary to examine the behavior of the symmetric eigenspinors (51) and their eigenvalues (52) in such cases. The possible distinct situations that can arise are the following:

(a) *k, m, r and s all distinct.* We have 4 independent

Debever–Penrose directions. The symmetric eigenspinors of χ and their corresponding eigenvalues are all distinct, thus P has 3 eigenvalues and eigenvectors and the space is therefore *Petrov Type I*.

(b) *If* $k = r \neq m \neq s$, 2 Debever–Penrose directions coincide. In this case, $k \wedge r = 0$, so the symmetric eigenspinors are

$$y_1 = k_{(A}m_{B)}r \wedge s + r_{(A}s_{B)}k \wedge m,$$
$$y_3 = -y_2 = k_{(A}k_{B)}m \wedge s \tag{53}$$

and we have $\lambda_2 = \lambda_3 \neq \lambda_1$.

Thus we have 2 eigenvalues and 2 symmetric eigenspinors, hence P has 2 distinct eigenvalues and eigenvectors and the space is *Petrov Type II*.

(c) *If* $k = r = m \neq s$, 3 Debever–Penrose directions coincide. In this case we find $y_1 = y_2 = y_3$, and $\lambda_1 = \lambda_2 = \lambda_3 = 0$. We have one symmetric eigenspinor, P has only one eigenvector, and the space is *Petrov Type III*.

(d) *If* $k = r \neq m = s$, the Debever-Penrose directions coincide in pairs. In this case, some care must be taken in using (51), since the forms of the eigenspinors become degenerate. It is then easier to rederive the eigenspinors in this case directly from (47), and one finds the 3 symmetric eigenspinors

$$y_1 = k_{(A}m_{B)}, \quad y_2 = k_{(A}k_{B)}, \quad y_3 = m_{(A}m_{B)}$$

and

$$\lambda_2 = \lambda_3 \neq \lambda_1.$$

Thus χ has 3 distinct symmetric eigenspinors, but only 2 distinct eigenvalues. Then P has 3 eigenvectors and 2 eigenvalues, and the space is therefore *Petrov Type I-D*.

(e) *If* $k = r = m = s$, all 3 Debever–Penrose directions coincide. In this case, we have the symmetric eigenspinors $y_1 = k_{(A}k_{B)}$ and $y_2 = k_{(A}j_{B)}$, where j_B is any 1-index spinor independent of k_B. The eigenvalues are zero. Thus P has 2 eigenvectors and zero eigenvalues, and the space is *Petrov Type II-N*.

A final case must be added for completeness. If $\chi = 0$, space is flat, anything is an eigenspinor and there are no preferred directions.

*Supported in part by the National Science Foundation, GP-16565.

[1]A. Z. Petrov, Sci. Notices Kazan State Univ. **114**, 55 (1954). Also, see A. Z. Petrov, *Einstein Spaces* (Pergamon, New York, 1964).
[2]R. Penrose, Ann. Phys. (N.Y.) **10**, 171 (1960).
[3]F. A. E. Pirani, "Introduction to Gravitational Radiation Theory," in *Lectures in General Relativity: Brandeis Summer Institute in Theoretical Physics* (Prentice-Hall, New York, 1964), Vol. I.
[4]A. Schild, "Lectures on General Relativity Theory," in *Lectures in Applied Mathematics*, edited by J. Ehlers (Amer. Math. Soc., Providence, R.I., 1967), Vol. 8.
[5]R. Adler, M. Bazin, and M. Schiffer, *Introduction to General Relativity* (McGraw-Hill, New York, 1965).
[6]R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary Matrices* (Cambridge U.P., Cambridge, 1952).

Exact nonplanar solutions of the classical relativistic three-body problem*

De-Ching Chern† and Peter Havas

Department of Physics, Temple University, Philadelphia, Pennsylvania 19122
(Received 5 June 1972)

It was shown recently by Havas that for a Newtonian system of three particles there exists a class of exact solutions which is nonplanar, such that the particles revolve with constant angular velocity in circular orbits in parallel planes. Such motions are possible for central forces which are arbitrary functions of the mutual separations, provided that the forces are repulsive between one pair of particles and attractive between the two other pairs (which for Coulomb forces corresponds to two like charges and one unlike one). In this paper, it is established that such motions are possible also for relativistic three-body systems with time-symmetric electromagnetic or mesic interactions; they reduce to the Newtonian solutions in the appropriate limit. Beyond this, there also exist exact relativistic solutions with the same orbits which have no nonrelativistic counterpart. These include solutions with all electric charges of the same sign, as well as solutions with all three particles on the same side of the axis of rotation; degenerate cases include planar and nonplanar two-particle solutions and a one-particle circular solution.

I. INTRODUCTION

In Newtonian dynamics, the problem of two bodies interacting through central forces depending on the mutual separation alone can be reduced to a one-body problem, which can always be integrated. The problem of $n > 2$ bodies, on the other hand, cannot be solved in general except for the case of the oscillator potential. Apart from this, the only interaction which has been studied extensively is the Newtonian gravitational one. For this case, some exact solutions of the three-body problem were found by Euler and Lagrange two centuries ago[1]; apart from generalization to $n > 3$, no further exact solutions are known. The known solutions[2] are all homographic, i.e., the configuration remains similar to itself for all time. Viewed from an inertial system in which the center of mass is at rest, they are either homothetic (i.e., the configuration is dilating without rotation), or planar (i.e., all n bodies move within the same fixed plane).

The general problem of homographic motions of a Newtonian system of particles interacting through central forces which are arbitrary functions of the mutual separations has been solved recently by Havas[3]; the special case of Coulomb interactions had been studied earlier by Müntz.[4] For any type of interactions for which the forces between pairs of particles are not all attractive or all repulsive, it is possible to have homographic motions which are not planar. In particular, it is possible to have nonplanar motions which are flat, i.e., the configurations is such that for every t there exists a plane $\Pi(t)$ containing all n bodies, and stationary. Then Π is rotating with constant angular velocity about an axis contained within it. For the special case of three particles, and forces which are repulsive for like "charges" and attractive for unlike ones, it was shown in HII that the configuration must be such that, viewed along the axis of rotation, the particle situated between the other two must have different sign of charge and be placed on the opposite side of the axis from the two like charges.

In relativistic dynamics, the two-body problem cannot be reduced to a one-body problem; the only known exact solutions are the planar, stationary solutions of Smith[5] and of Schild[6] for time-symmetric electrodynamic interactions, as well as similar solutions for time-symmetric gravitational interactions[5,7] (using equations of motion arising from a Lorentz-invariant approximation method in general relativity). In all these solutions, the particles travel in concentric circles and are situated on a diameter on opposite sides of the center.

As discussed in HI, another class of exact relativistic solutions can be obtained for a type of interaction suggested by Havas and Plebański.[8] In particular, these solutions exist for all n-body problems for which the Newtonian orbit of each body is circular, and the motion is flat and nonplanar, and thus for the three-body motions considered in HII, provided that all speeds are less than c.

Apart from these solutions, no exact relativistic solutions are known for $n > 2$. In this paper, we are concerned with another relativistic generalization of the three-body solutions with Coulomb or more general forces considered in HII, replacing these interactions by time-symmetric electromagnetic or mesic interactions. The equations of motion for point particles interacting through mesic fields were first derived by Bhabha and Harish-Chandra,[9] using the method devised by Dirac[10] for the electromagnetic case. The interactions they considered are the retarded ones. The corresponding field-theoretical equations of motion for the time-symmetric case, as well as a consistent procedure for obtaining the corresponding (nonequivalent) equations of motion of action-at-a-distance theory, were considered by Havas.[11] The equations of motion derived from the field-theoretical point of view contain self-action terms which are in the form of an integral over the previous motion in the retarded case and over the entire motion in the symmetric case. However, such an action of a particle on itself has no meaning from the point of view of action at a distance; it can be discarded without inconsistency, as discussed in Ref. 11.

In Sec. II we discuss the formal solution for general time-symmetric interactions of a type encompassing the special interactions mentioned above for the three-particle problem; some of the considerations are extended to n-particle nonplanar, but flat, solutions. In Sec. III the results are applied to these special interactions; the existence of exact solutions is established for special values of the parameters characterizing the three-particle system, and several limiting cases are studied. Apart from solutions which reduce to the nonrelativistic solutions of HII in the appropriate limit, we also obtain several types of solutions which have no nonrelativistic counterpart.

II. THE GENERAL SCHEME OF THE SOLUTION

A. Notation

We shall use a Minkowski space with coordinates x^μ, where $\mu = 0, 1, 2, 3$, and the zeroth component represents time. The metric tensor $\eta_{\mu\nu}$ is taken as

$$\eta_{00} = 1, \quad \eta_{11} = \eta_{22} = \eta_{33} = -c^{-2}, \quad \eta_{\mu\nu} = 0 \quad \text{if } \mu \neq \nu. \tag{1}$$

Summation over repeated Greek indices is understood. The coordinates of a particle j are denoted by z_j^μ and its world line is given by $z_j^\mu(t_j)$.

We consider three particles 1, 2 and 3 in a plane containing the axis of rotation, which is chosen as the 3-axis. At $t = t_1 = t_2 = t_3 = 0$ all the particles are in the 2-3 plane; then their world lines are

$$z_j^0 = t_j, \quad z_j^1 = -r_j \sin\omega t_j, \quad z_j^2 = r_j \cos\omega t_j, \quad z_j^3 = z_j, \quad j = 1, 2, 3, \tag{2}$$

where ω is the angular velocity, which will be taken as positive. To allow a simple description of the motion, we have taken r_j as the coordinate z_j^2 at $t = 0$, and allow it to be either positive or negative; because the relativistic case allows a greater variety of possible configurations than the nonrelativistic one, this is more convenient than the conventions and coordinates used in HII.

Because only two-body, time-symmetric interactions are considered, and the motion of the particles is steady and possesses cylindrical symmetry, we may choose the "field point" to be in the 2-3 plane and its time to be zero and to serve as a reference point. Thus the 1-component is tangential to the orbit of each particle and the 2-component is radial. Then the four-distance between particles i and j is given by

$$\begin{aligned} &s_{ij}^\mu \equiv z_i^\mu(0) - z_j^\mu(-t_{ij}): \\ &s_{ij}^0 = t_{ij}, \quad s_{ij}^1 = -r_j \sin\omega t_{ij}, \\ &s_{ij}^2 = r_i - r_j \cos\omega t_{ij}, \quad s_{ij}^3 = z_i - z_j, \\ &t_{ij} \equiv t_i - t_j, \quad t_i = 0. \end{aligned} \tag{3}$$

When two particle indices are attached to a single quantity, the first one always refers to the "field point" and the second one to the "source point." When self-action terms are involved, we have $i = j$ and thus $t_{ii} = t_i - t_i'$, with $t_i = 0$.

Occasionally we shall also use the proper times τ_i as parameters, with

$$d\tau_i = (\eta_{\mu\nu} dz_i^\mu dz_i^\nu)^{1/2}. \tag{4}$$

B. Geometric parameters

A natural choice of the unit of length for our problem is c/ω. We define six dimensionless quantities

$$\beta_i \equiv r_i\omega/c, \quad \alpha_{ij} \equiv (z_i - z_j)\omega/c, \tag{5}$$

where the β's are the usual v/c and can be positive or negative. The α's are antisymmetric in i and j, with

$$\alpha_{12} + \alpha_{23} + \alpha_{31} = 0. \tag{6}$$

Hence, there are only five independent quantities, which will be called the geometric parameters.

A natural choice of the unit of time is $1/\omega$. Therefore, we define

$$\varphi_{ij} \equiv \omega t_{ij}. \tag{7}$$

Transcribed into these parameters, Eq. (3) becomes

$$\begin{aligned} &s_{ij}^0 = \frac{\varphi_{ij}}{\omega}, \quad s_{ij}^1 = -\frac{c}{\omega}\beta_j \sin\varphi_{ij}, \\ &s_{ij}^2 = \frac{c}{\omega}(\beta_i - \beta_j \cos\varphi_{ij}), \quad s_{ij}^3 = \frac{c}{\omega}\alpha_{ij}. \end{aligned} \tag{8}$$

C. Equations of motion

Since only two-body forces are considered, the equations of motion for a three-particle problem are generally of the form

$$m_i \frac{d^2 z_i^\mu}{d\tau_i^2} = F_{ij}^\mu + F_{ik}^\mu + F_{ii}^\mu, \quad i \neq j \neq k \neq i, \tag{9}$$

where m_i is the mass of particle i. F_{ij}^μ is the force acting on particle i due to particle j; F_{ii}^μ represents the self-action of particle i (which is absent in an action-at-a-distance theory).

In our problem, all the terms in Eq. (9) must be independent of time, and be functions of the geometric parameters. Furthermore, F_{ij}^μ must depend bilinearly on the charges of i and j, or

$$\begin{aligned} &F_{ij}^0 = e_i e_j \frac{\omega^2}{c^3} f(\beta_i, \beta_j, \alpha_{ij}), \\ &F_{ij}^m = e_i e_j \frac{\omega^2}{c^2} f^m(\beta_i, \beta_j, \alpha_{ij}), \quad m = 1, 2, 3, \end{aligned} \tag{10}$$

where the f's stand for some dimensionless functions to be calculated for specific interactions. The e's characterize the charges of the particles and $e^2/(\text{length})^2$ is assumed to have the dimension of force. For conciseness we shall use e regardless of the nature of the interaction.

Our purpose is to find under what conditions Eq. (9) is satisfied by the world lines defined by Eq. (2).

D. Formal solution

With a few assumptions, the equations of motion can be simplified and a formal solution obtained without adopting any specific interaction.

We assume that the force terms can be derived from a Fokker-type interaction Lagrangian[11-13]

$$J_{ij} = \iint_{-\infty}^{\infty} d\tau_i d\tau_j \Lambda_{ij}, \quad i \neq j \tag{11}$$

where

$$\begin{aligned} &\Lambda_{ij} = \Lambda_{ji} = \Lambda_{ij}(s_{ij}, \omega_{ij}, v_{ij}), \\ &s_{ij} \equiv (\eta_{\mu\nu} s_{ij}^\mu s_{ij}^\nu)^{1/2}, \\ &\omega_{ij} \equiv \eta_{\mu\nu} v_i^\mu v_j^\nu, \quad v_i^\mu \equiv \frac{dz_i^\mu}{d\tau_i}, \\ &v_{ij} \equiv v_i v_j, \quad v_i \equiv (\eta_{\mu\nu} v_i^\mu v_i^\nu)^{1/2}. \end{aligned} \tag{12}$$

We shall prove that, for our problem,

(i) $F_{ij}^0 = 0, \quad F_{ij}^1 = 0,$

(ii) $F_{ij}^3 \gamma_i^{-1} = -F_{ji}^3 \gamma_j^{-1}, \quad \gamma_i \equiv (1 - \beta_i^2)^{-1/2},$

(iii) $F_{ii}^0 = F_{ii}^1 = F_{ii}^3 = 0.$

After the standard variational procedure, we obtain from Eq. (11)

$$F_{ij}^\mu = \int f_{ij}^\mu d\tau_j, \tag{13}$$

where

$$f^{\mu}_{ij} = 2s^{\mu}_{ij}\frac{\partial\Lambda_{ij}}{\partial s^2_{ij}} - \frac{d}{d\tau_i}\left(\frac{\partial\Lambda_{ij}}{\partial\omega_{ij}}v^{\mu}_j + \frac{\partial\Lambda_{ij}}{\partial v_{ij}}v^{\mu}_i\right). \tag{14}$$

The integrand f^{μ}_{ij} of Eq. (13) may be interpreted as the force acting on i due to the particle j in the interval between τ_j and $\tau_j + d\tau_j$.

For our problem, $d\tau_j = -\gamma_j^{-1}dt_{ij}$ and all the quantities in Eq. (14) can be expressed as functions of t_{ij}. Clearly, s_{ij}, ω_{ij}, and v_{ij} are even functions of t_{ij}. Therefore Λ_{ij} must also be even in t_{ij}. It is also easily seen that

$$\begin{aligned} f^0_{ij}(t_{ij}) &= -f^0_{ij}(-t_{ij}), \quad f^1_{ij}(t_{ij}) = -f^1_{ij}(-t_{ij}), \\ f^2_{ij}(t_{ij}) &= +f^2_{ij}(-t_{ij}), \quad f^3_{ij}(t_{ij}) = +f^3_{ij}(-t_{ij}). \end{aligned} \tag{15}$$

From this and Eq. (13) we obtain

$$F^0_{ij} = 0, \quad F^1_{ij} = 0, \tag{16}$$

establishing (i). If self-action terms are present [which can not be included in (11)], they must have the same properties under time reversal as the interactions and thus

$$F^0_{ii} = 0, \quad F^1_{ii} = 0. \tag{17}$$

Eqs. (16) and (17) reflect the fact that radiation reaction is excluded by the use of time-symmetric interactions. If (17) would not hold, we would have

$$\begin{aligned} F^0_{ii} &= e_i^2\frac{\omega^2}{c^3}f(\beta_i) \neq 0, \\ F^1_{ii} &= e_i^2\frac{\omega^2}{c^2}f^1(\beta_i) \neq 0. \end{aligned} \tag{18}$$

Let particle i be constrained to move according to Eq. (2), and the other particles be chargeless. Then the 0- and 1- components of the equations of motion become

$$0 = F^0_{ii} + F^0_c, \quad 0 = F^1_{ii} + F^1_c, \tag{19}$$

where F^{μ}_c is the constant constraining force. Since radiation is excluded, conservation of energy and of angular momentum is violated unless $F^0_c = F^1_c = 0$. Hence, Eq. (17) follows from conservation requirements. A similar argument also shows that $F^3_{ii} = 0$, establishing (iii). Hence the only possible nonvanishing component of the self-action force in the field-theoretical case is F^2_{ii}.

To prove (ii), we note that $v^3_i = v^3_j = 0$ and $s^3_{ij} = z_i - z_j = \text{const}$ in our problem. Therefore, Eq. (13) reduces to

$$F^3_{ij} = 2(z_i - z_j)\int d\tau_j\frac{\partial\Lambda_{ij}}{\partial s_{ij}}, \tag{20}$$

or, using $d\tau_j = \gamma_j^{-1}dt'$,

$$F^3_{ij} = 2(z_i - z_j)\gamma_j^{-1}\int dt'\frac{\partial\Lambda_{ij}}{\partial s_{ij}}. \tag{21}$$

As the integrand is symmetric in i and j, we find

$$F^3_{ij}\gamma_i^{-1} = -F^3_{ji}\gamma_j^{-1}, \tag{22}$$

establishing (ii). Clearly, these terms represent the impulse transfered from one particle to another; thus, Eq. (22) represents a generalized Newton's third law.

Lagrangians of the above type are used for the electromagnetic[12] and scalar or vector meson interactions.[11] Therefore, the results hold for all these cases.[14]

Equation (21), together with Eqs. (10) and (5), indicates that F^3_{ij} must be of the form

$$F^3_{ij} = e_ie_j\frac{\omega^2}{c^2}\alpha_{ij}\gamma_i\mathfrak{F}_{ij}(\beta_i, \beta_j, \alpha_{ij}), \tag{23}$$

where $\mathfrak{F}_{ij}$ is symmetric in i and j. Therefore

$$F^3_{ij} = 0 \quad \text{if } \alpha_{ij} = 0, \tag{24}$$

i.e., if the motion is planar.

Following the pattern of Eq. (23), we also write

$$F^2_{ij} = e_ie_j\frac{\omega^2}{c^2}\gamma_i\beta_i\mu_{ij}(\beta_i, \beta_j, \alpha_{ij}); \tag{25}$$

but here the function μ_{ij} is not necessarily symmetric in i and j, and $\beta_i = 0$ does not necessarily imply $F^2_{ij} = 0$.

The only nonvanishing component of the four-acceleration for the "field point" at $t = 0$ is the radial one, $d^2z_i^2/d\tau_i^2 = -c\omega\gamma_i^2\beta_i$. Applying this and the above results to Eq. (9), we find that the 0- and 1-components are identically zero. The other two components are

$$-m_ic\omega\gamma_i^2\beta_i = (\omega/c)^2e_i\beta_i\gamma_i\sum_j e_j\mu_{ij}, \tag{26}$$

$$0 = (\omega/c)^2e_i\gamma_i(e_j\alpha_{ij}\mathfrak{F}_{ij} + e_k\alpha_{ik}\mathfrak{F}_{ik}). \tag{27}$$

Thus the ratios of the e's are determined by Eq. (27). Introducing ϵ_k by

$$\alpha_{ij}\mathfrak{F}_{ij} = \sum_k \pi_{ijk}\epsilon_k, \tag{28}$$

where π_{ijk} is the permutation symbol (antisymmetric in all indices, with $\pi_{123} = 1$), we have

$$e_1 : e_2 : e_3 = \epsilon_1 : \epsilon_2 : \epsilon_3. \tag{29}$$

From this and Eq. (26) we obtain

$$-\frac{m_ic^3}{\omega e_i^2} = \gamma_i^{-1}\sum_j\frac{\epsilon_j}{\epsilon_i}\mu_{ij}, \tag{30}$$

provided that none of the ϵ's vanishes.

Equations (29) and (30) contain five independent equations, whose right hand sides are functions of the geometric parameters. Therefore, these equations indicate the relation of the five geometric parameters to the five independent particle parameters, which can be chosen as the ratios

$$e_1 : e_2 : e_3; \quad \frac{m_i}{e_i^2\omega}, \quad i = 1, 2, 3. \tag{31}$$

Once the ϵ's and μ's are given, a set of geometric parameters determines a set of particle parameters through Eqs. (29) and (30). In this sense, the geometric parameters can be considered as the independent variables, while the particle parameters are the dependent variables.

These ten parameters constitute a solution. However, not all the particle parameters are acceptable physically; the mass, and therefore the $m_i/(e_i^2\omega)$, must be positive. This in turn imposes three "positive mass conditions" on the geometric parameters. From Eq. (30) we obtain

$$\sum_j\frac{\epsilon_j}{\epsilon_i}\mu_{ij} < 0. \tag{32}$$

If none of the ϵ's vanishes, we are assured of the existence of a physical solution whenever these three inequalities are satisfied.

Equation (32) can be put in a different form. Defining

$$\mu_i = \gamma_i^{-1}\left(\frac{\mu_{ij}}{\epsilon_k} + \frac{\mu_{ik}}{\epsilon_j} + \frac{\epsilon_i \mu_{ii}}{\epsilon_j \epsilon_k}\right), \tag{33}$$

we get, multiplying Eq. (30) by $\epsilon_i^2/(\epsilon_1\epsilon_2\epsilon_3)$ and using Eq. (29),

$$m_1 : m_2 : m_3 = \mu_1 : \mu_2 : \mu_3. \tag{34}$$

The positive mass conditions may now be written as

$$\operatorname{sign}\mu_i = -\operatorname{sign}(\epsilon_1\epsilon_2\epsilon_3), \quad i = 1, 2, 3. \tag{35}$$

The procedure described above provides us with solutions for the particle parameters (31) in terms of the geometric parameters (5). The usual procedure in dynamics is to search for solutions for a system with given masses and charges; for the particular type of solutions considered here and in HII, it is to be expected that solutions exist only for a limited range of these quantities. However, the determination of this range is a difficult problem which has not even been fully solved for the nonrelativistic case treated in HII. For the relativistic case treated here we have not attempted to find a systematic procedure for determining this range; however, it will be shown later that solutions may exist for values of the charges which do not allow a nonrelativistic solution.

In the nonrelativistic case, it was shown in HII that, for Coulomb interactions and the configurations described in the Introduction, solutions exist for arbitrary choice of the geometric parameters, or, in the terminology used above, that the positive mass conditions are always satisfied. In the mesic case, no such simple result holds, as discussed in Sec. IIIB.

E. Special cases

We now consider those general cases where some of the ϵ's vanish. If one or two of them vanish, this implies by Eq. (29) that the corresponding charges also vanish. Since the uncharged particles do not interact with the other particles and thus cannot influence their motions, we then are no longer dealing with a three-particle system.

If only one particle has a nonvanishing charge, say e_3, Eq. (27) is identically satisfied. From Eq. (26) we obtain

$$-\frac{m_3 c^3}{e_3^2\omega} = \gamma_3^{-1}\mu_{33} < 0 \tag{36}$$

as the positive mass condition. Therefore, a one-particle solution, i.e., a particle traveling around a circle under the influence of its own field, is possible if the self-action term μ_{ii} is negative. This type of solution does not exist in a theory of action at a distance.

If one of the charges vanishes, say $e_1 = \epsilon_1 = 0$, this implies

$$\alpha_{23} = 0 \quad \text{or} \quad \mathfrak{F}_{23} = 0. \tag{37}$$

If $\alpha_{23} = 0$, the solution reduces to the planar two-particle solutions of Smith[5] and of Schild.[6] If $\mathfrak{F}_{23} = 0$ and $\alpha_{23} \neq 0$, on the other hand, we obtain a nonplanar two-particle solution.

For both the planar and nonplanar cases, Eq. (26) reduces to

$$-\frac{m_2 c^3}{e_2^2\omega} = \gamma_2^{-1}\left[\frac{e_3}{e_2}\ \mu_{23} + \mu_{22}\right] < 0 \tag{38}$$

and a similar equation with 2 and 3 interchanged, which are the positive mass conditions in the case of field theory. In the case of the theory of action at a distance, for which $\mu_{ii} = 0$, it can be shown by a proper choice of the sign of $e_2 e_3$ that the only condition is

$$\frac{m_2}{m_3} = \frac{\gamma_3}{\gamma_2}\frac{\mu_{23}}{\mu_{32}} > 0. \tag{39}$$

Only three of the five equations of motion have been used, because particle 1 is considered absent. However, if we consider it as a test particle instead, the other two equations provide a solution to the restricted three-body problem. We have

$$\frac{e_2}{e_3} = \frac{\epsilon_2}{\epsilon_3}, \tag{40}$$

$$\frac{m_1}{e_1} = \frac{m_2}{e_2}\frac{\gamma_2}{\gamma_1}\frac{\mu_{12} + \mu_{13}\epsilon_3/\epsilon_2}{\mu_{23}\epsilon_3/\epsilon_2 + \mu_{22}}. \tag{41}$$

In case all of the ϵ's vanish, Eq. (29) does not imply that any of the e's should be zero, but ceases to specify the ratios of the charges. Then the ϵ's in Eqs. (30) and (32) should be replaced by e's. In this sense, the ratios of the e's become independent variables. But the geometric parameters have to satisfy the condition that all the ϵ's vanish. In particular, all the α's may be zero, which corresponds to a planar solution.

F. Generalization to n particles

Instead of three particles, we now consider a system of n particles each of which has a world line of the form (2), and corresponding sets of $2n-1$ geometric and $2n-1$ particle parameters. Equation (26) still holds, with the summation running from 1 to n. Equation (27), on the other hand, contains $n-1$ terms instead of 2, and thus relation (28) no longer holds. The equations corresponding to Eq. (27) may be written in matrix form as

$$0 = \begin{pmatrix} 0 & \epsilon_{12} & \epsilon_{13} & \cdots & \epsilon_{1n} \\ \epsilon_{21} & 0 & \epsilon_{23} & \cdots & \\ \epsilon_{31} & \epsilon_{32} & 0 & \cdots & \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \epsilon_{n1} & \cdot & & \cdots & 0 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{pmatrix}, \tag{42}$$

where $\epsilon_{ij} = -\epsilon_{ji} = \alpha_{ij}\mathfrak{F}_{ij}$. A nontrivial solution exists only if

$$\det\epsilon_{ij} = 0, \tag{43}$$

i.e., if the rank r of the determinant is smaller than n. As proved by Frobenius,[15] for an antisymmetric determinant r is always even and there exists at least one nonsingular $r \times r$ minor which is antisymmetric around the original diagonal. Hence, without loss of generality, we may rewrite Eq. (42) as

$$\begin{pmatrix} 0 & \epsilon_{12} & \epsilon_{13} & \cdots & \epsilon_{1r} \\ \epsilon_{21} & 0 & \epsilon_{23} & \cdots & \\ \epsilon_{31} & \epsilon_{32} & 0 & \cdots & \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \epsilon_{r1} & \cdot & & \cdots & 0 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \cdot \\ \cdot \\ \cdot \\ e_r \end{pmatrix}$$

$$= - \begin{pmatrix} \epsilon_{1\,r+1} & \epsilon_{1\,r+2} & \cdots & \epsilon_{1n} \\ \epsilon_{2\,r+1} & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \epsilon_{r\,r+1} & \cdot & \cdots & \epsilon_{rn} \end{pmatrix} \begin{pmatrix} e_{r+1} \\ e_{r+2} \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{pmatrix}, \tag{44}$$

where the square matrix on the left hand side is nonsingular. This equation may be solved by Cramer's rule. We obtain

$$e_j = \Delta_j/\Delta, \quad j \le r, \tag{45}$$

where Δ is the determinant of the square matrix and Δ_j is the same determinant with the jth column replaced by the right-hand side of Eq. (44). Generally, the determinant of an antisymmetric matrix of even order can be expressed as a perfect square of a polynomial of its elements. Hence Δ is always positive.

At least one member of the last column matrix of Eq. (44) must be nonzero to have a nontrivial solution. Let $e_n \neq 0$ and define

$$\epsilon_j = \frac{\Delta_j}{e_n \Delta^{1/2}}, \quad 1 \le j \le r,$$
$$\epsilon_j = \frac{e_j \Delta^{1/2}}{e_n}, \quad r < j \le n. \tag{46}$$

(If $r = 0$, the first line is absent and in the second line Δ is replaced by 1.) Then we have

$$\epsilon_1 : \epsilon_2 : \epsilon_3 : \cdots = e_1 : e_2 : e_3 : \cdots . \tag{47}$$

Eqs. (30) and (32) retain their form except that the summations run from 1 to n. Then we have n positive mass conditions.

For $r < n - 1$, the ϵ_j are functions of the ratios of $n - r$ e's as well as of the geometric parameters. For $n = 3$ and $r = 0$, we have the case discussed at the end of Sec. IIE.

If n is an odd number, Eq. (43) is always satisfied and r may equal $n - 1$. In this case, all ϵ_j are functions of the geometric parameters only. For $n = 3$ and $r = 2$, Eq. (46) becomes Eq. (28).

If n is an even number, Eq. (43) becomes a condition imposed on the geometric parameters, and $r \le n - 2$. For $n = 2$, Eq. (43) becomes Eq. (37), and $r = 0$.

The solutions considered here are flat. For the nonrelativistic case, n-particle solutions which are not flat have been given in HII. Their relativistic generalization goes beyond the scope of this paper; however, the method developed here can be adapted to the nonflat case.

III. SPECIFIC INTERACTIONS

A. Exact form of force terms

In the electromagnetic case, the force terms are derived from the time-symmetric Liénard–Wiechert potentials. The equations of motion for time-symmetric neutral scalar and vector mesic interactions, derived from the point of view of both field theory and from that of the theory of action at a distance, are given in Ref. 11; for charge-symmetric mesic interactions they are given in Ref. 16. In the mesic case, there is a new parameter χ ($=mc/\hbar$ in a quantized theory, where m is the rest mass of the meson). The calculations presented here refer to the neutral case; however, the results also represent special solutions for the charge-symmetric case.[16]

In the following, the double subscripts ij are omitted for brevity where no confusion may arise; the letters E, S, V after the number of the equations indicate the cases of electromagnetic, scalar mesic, and vector mesic interactions, respectively. With the help of Eqs. (23) and (25), we obtain[17]

$$\mathcal{F}_{ij} = -\frac{1}{K}\frac{\partial}{\partial\varphi}\frac{P}{K}, \tag{48E}$$

$$\mathcal{F}_{ij} = \frac{1}{\gamma_i\gamma_j}\left(\frac{1}{K}\frac{\partial}{\partial\varphi}\frac{1}{K} + \frac{1}{\lambda^2}\frac{1}{2K} - \frac{1}{\lambda^4}\Psi_2[1]\right), \tag{48S}$$

$$\mathcal{F}_{ij} = -\frac{1}{K}\frac{\partial}{\partial\varphi}\frac{P}{K} - \frac{1}{\lambda^2}\frac{P}{2K} + \frac{1}{\lambda^4}\Psi_2[P], \tag{48V}$$

$$\mu_{ij} = -\frac{1}{K}\frac{\partial}{\partial\varphi}\frac{Q}{K}, \tag{49E}$$

$$\mu_{ij} = \frac{\gamma_i}{\gamma_j}\left(-\frac{1}{K} + \frac{1}{\lambda^2}\Psi_1[1]\right) + \frac{1}{\gamma_i\gamma_j}\left(\frac{1}{K}\frac{\partial}{\partial\varphi}\frac{R}{K} + \frac{1}{\lambda^2}\frac{R}{2K} - \frac{1}{\lambda^4}\Psi_2[R]\right), \tag{49S}$$

$$\mu_{ij} = -\frac{1}{K}\frac{\partial}{\partial\varphi}\frac{Q}{K} - \frac{1}{\lambda^2}\frac{Q}{2K} + \frac{1}{\lambda^4}\Psi_2[Q], \tag{49V}$$

$$\mu_{ii} = 0, \tag{50E}$$

$$\mu_{ii} = \frac{1}{\lambda^2}\Psi_1[1] - \frac{1}{\gamma_i^2}\frac{1}{\lambda^4}\Psi_2[R], \tag{50S}$$

$$\mu_{ii} = \frac{1}{\lambda^4}\Psi_2[Q], \tag{50V}$$

all quantities not under the integral operator Ψ are evaluated at $\varphi = \theta$, where $\theta \equiv \theta_{ij}$ is defined by the positive root of

$$\theta^2 = \alpha_{ij}^2 + \beta_i^2 + \beta_j^2 - 2\beta_i\beta_j\cos\theta, \tag{51}$$

and

$$K \equiv K_{ij}(\varphi) \equiv \varphi - \beta_i\beta_j\sin\varphi,$$
$$P \equiv P_{ij}(\varphi) \equiv 1 - \beta_i\beta_j\cos\varphi, \tag{52}$$
$$Q \equiv Q_{ij}(\varphi) \equiv 1 - \frac{\beta_j}{\beta_i}(\varphi\sin\varphi + \cos\varphi) - \beta_j(\beta_i\cos\varphi - \beta_j),$$
$$R \equiv R_{ij}(\varphi) \equiv 1 - \frac{\beta_j}{\beta_i}\cos\varphi,$$
$$\lambda \equiv \frac{\omega}{c\chi}.$$

The integral operator Ψ_n is defined as

$$\Psi_n[X] \equiv \Psi_n[X_{ij}] = \int_\theta^\infty \frac{J_n(c\chi s)}{(c\chi s)^n} X(\varphi)\, d\varphi, \tag{53}$$

where J_n is the Bessel function. The lower limit of integration θ is given by Eq. (51); when applied to the self-action terms Eqs. (50S, V), it reduces to $\theta_{ii} = 0$.

Through integrations by parts, Eqs. (48S, V) and (49S, V) become

$$\mathcal{F}_{ij} = -\frac{1}{\gamma_i\gamma_j}\Psi_0\left[\frac{\partial}{\partial\varphi}\left(\frac{1}{K}\frac{\partial}{\partial\varphi}\frac{1}{K}\right)\right], \tag{54S}$$

$$\mathcal{F}_{ij} = \Psi_0\left[\frac{\partial}{\partial\varphi}\left(\frac{1}{K}\frac{\partial}{\partial\varphi}\frac{P}{K}\right)\right], \tag{54V}$$

$$\mu_{ij} = \frac{\gamma_i}{\gamma_j}\Psi_0\left[\frac{\partial}{\partial\varphi}\frac{1}{K}\right] - \frac{1}{\gamma_i\gamma_j}\Psi_0\left[\frac{\partial}{\partial\varphi}\left(\frac{1}{K}\frac{\partial}{\partial\varphi}\frac{R}{K}\right)\right], \tag{55S}$$

$$\mu_{ij} = \Psi_0\left[\frac{\partial}{\partial\varphi}\left(\frac{1}{K}\frac{\partial}{\partial\varphi}\frac{Q}{K}\right)\right]. \tag{55V}$$

These equations are too complicated to be analyzed in general, and thus we only consider some limiting cases in the following. The values of the particle parameters for the general case can be obtained through numerical methods. As an example, a numerical and graphical analysis for the case of electromagnetic interaction with $\beta_1 = -\beta_2 = \beta_3$ and $\alpha_{12} = \alpha_{23} = \frac{1}{2}\alpha_{13} < 3\pi$ is given in Appendix A of Ref. 17. It is found that there are forbidden regions of these parameters where the positive mass conditions are not satisfied; these are indicated in a diagram. The relative signs of the charges for the allowed regions and the planar and nonplanar two-particle solutions are also discussed; in particular, it is found that in the nonplanar case for certain values of the geometric parameters all charges can be of the same sign.

Although we now restrict ourselves to limiting cases, it should be noted that they are used only to show the existence and location of some solutions, but that these solutions, of the form (2), are still exact.

B. Nonrelativistic limit

In this limit we take all the geometric parameters to be small; in the mesic case, λ is also assumed to be small. The results are as expected; the electromagnetic and mesic interactions reduce to the static Coulomb and Yukawa interactions, respectively. We have

$$\mathcal{F}_{ij} \cong -\frac{1}{\theta^3}, \tag{56E}$$

$$\mathcal{F}_{ij} \cong \frac{\exp(-\theta/\lambda)}{\theta^3}\left(1+\frac{\theta}{\lambda}\right), \tag{56S}$$

$$\mathcal{F}_{ij} \cong -\frac{\exp(-\theta/\lambda)}{\theta^3}\left(1+\frac{\theta}{\lambda}\right), \tag{56V}$$

$$\mu_{ij} \cong -\frac{1}{\theta^3}\left(1-\frac{\beta_j}{\beta_i}\right), \tag{57E}$$

$$\mu_{ij} \cong \frac{\exp(-\theta/\lambda)}{\theta^3}\left(1+\frac{\theta}{\lambda}\right)\left(1-\frac{\beta_j}{\beta_i}\right), \tag{57S}$$

$$\mu_{ij} \cong -\frac{\exp(-\theta/\lambda)}{\theta^3}\left(1+\frac{\theta}{\lambda}\right)\left(1-\frac{\beta_j}{\beta_i}\right). \tag{57V}$$

The self-action terms are too small to distinguish the field theory from the theory of action at a distance in this limit, as will be shown in Sec. IIIE.

By Eq. (33), we have

$$\mu_i \cong \frac{1-\beta_j/\beta_i}{\alpha_{ij}} - \frac{1-\beta_k/\beta_i}{\alpha_{ik}}, \qquad \pi_{ijk} = +1, \tag{58}$$

for all three cases. Employing Eq. (35), we easily obtain the following results:

(i) There is no nonplanar solution in this limit for the scalar meson case. This is due to the fact that the scalar (in contrast to the vector) mesic interaction is attractive for like charges. Thus, this interaction is analogous to the gravitational rather than the electromagnetic one.

(ii) For the electromagnetic and vector meson cases, a nonplanar solution is always possible. But the signs of the charge and radius of the middle particle must be different from those of the outer two, in agreement with the results of HII.

C. Close range limit

In this limit we take all the geometric parameters to be small, but let λ remain finite. For the electromagnetic case, this is identical to the nonrelativistic limit. For the mesic cases, the results are similar to the nonrelativistic limit except that the factor $\exp(-\theta/\lambda) \times (1+\theta/\lambda)$ reduces to 1 in Eqs. (56 S, V) and (57 S, V). Eq. (58) remains valid. The solutions exist under the same conditions as in B.

However, if the higher order terms are considered, we find nonanalyticities at $\lambda = 1/N$ (N integer). The characteristics of these nonanalyticities are similar to those of the self-action terms which are discussed in more detail in Sec. IIIE. They do not show up in the nonrelativistic limit where λ was taken to be infinitesimal. The apparent inconsistency of the disappearance of the nonanalyticities as $\lambda \to 0$ is explained by property (iii) of the power series discussed in Sec. IIIE.

D. Asymptotic limit

In this limit we take the α's to be large. In the electromagnetic case the results are relatively simple. We have

$$\mathcal{F}_{ij} \cong -\beta_i\beta_j\frac{\sin\theta}{\theta^2}, \tag{59E}$$

$$\mu_{ij} \cong \frac{\beta_j}{\beta_i}\frac{\cos\theta}{\theta}, \tag{60E}$$

assuming that $|\sin\theta|$ and $|\cos\theta|$ are large in comparison with $|1/\alpha|$.

For the mesic interactions, we shall first consider the case of action at a distance. The self-action terms will be discussed later. By carrying out the integrations on Eqs. (54 S, V) and (55 S, V) through an expansion of the operands in this limit, we again find nonanalyticities at $1/\lambda = N$. For the region $\lambda > 1$ we have

$$\mathcal{F}_{ij} \cong \frac{\beta_i\beta_j}{\gamma_i\gamma_j}\frac{\cos[(1-\lambda^{-2})^{1/2}\theta]}{\theta^3}\left(1-\frac{1}{\lambda^2}\right), \tag{61S}$$

$$\mathcal{F}_{ij} \cong -\beta_i\beta_j\frac{\sin[(1-\lambda^{-2})^{1/2}\theta]}{\theta^2}\left(1-\frac{1}{\lambda^2}\right)^{1/2}, \tag{61V}$$

$$\mu_{ij} \cong \frac{\gamma_i}{\gamma_j}\frac{\beta_j}{\beta_i}(1-2\beta_i^2)\frac{\sin[(1-\lambda^{-2})^{1/2}\theta]}{\theta^2}\left(1-\frac{1}{\lambda^2}\right)^{1/2}, \tag{62S}$$

$$\mu_{ij} \cong \frac{\beta_j}{\beta_i} \frac{\cos[(1-\lambda^{-2})^{1/2}\theta]}{\theta}, \tag{62V}$$

assuming that $|\sin[(1-\lambda^{-2})^{1/2}\theta]|$, $|\cos[(1-\lambda^{-2})^{1/2}\theta]|$, and, in the scalar case, $|1-2\beta^2|$ are large in comparison with $|1/\alpha|$. For the region $1 > \lambda > \frac{1}{2}$ we have

$$\mathfrak{F}_{ij} \cong \frac{2\beta_i\beta_j}{\gamma_i\gamma_j} \frac{\sin[(4-\lambda^{-2})^{1/2}\theta]}{\theta^4} \left(1-\frac{1}{4\lambda^2}\right)^{3/2}, \tag{63S}$$

$$\mathfrak{F}_{ij} \cong 2\beta_i^2\beta_j^2 \frac{\cos[(4-\lambda^{-2})^{1/2}\theta]}{\theta^3} \left(1-\frac{1}{4\lambda^2}\right), \tag{63V}$$

$$\mu_{ij} \cong \frac{\gamma_i}{\gamma_j} \beta_j^2(3\beta_i^2-2) \frac{\cos[(4-\lambda^{-2})^{1/2}\theta]}{\theta^3} \left(1-\frac{1}{4\lambda^2}\right), \tag{64S}$$

$$\mu_{ij} \cong 2\beta_j^2 \frac{\sin[(4-\lambda^{-2})^{1/2}\theta]}{\theta^2} \left(1-\frac{1}{4\lambda^2}\right)^{1/2}, \tag{64V}$$

assuming that $|\sin[(4-\lambda^{-2})^{1/2}\theta]|$, $|\cos[(4-\lambda^{-2})^{1/2}\theta]|$, and, in the scalar case, $|2-3\beta^2|$ are large in comparison with $|1/\alpha|$.

In the region $\lambda > 1$, the leading terms of the expansions yield, after integration, the results indicated by Eqs. (61 S, V) and (62 S, V). But in the region $1 > \lambda > \frac{1}{2}$, these terms fall off exponentially with $\theta(\theta \cong \alpha$ in this limit). We obtain the results (63 S, V) and (64 S, V) from the second order terms of the expansions; for smaller λ we have to use higher-order terms.

It is easy to show that the positive mass conditions can be satisfied and that solutions exist in the limit considered. The signs of the β's do not affect the existence of the solutions, and thus can be chosen arbitrarily, including the case that they are all positive, i.e., that the three particles are on the same side of the axis of rotation. By proper choice of the values of the θ's, the signs of the e's can also be chosen arbitrarily.

The nonplanar two-particle solutions can also easily be found from the above results for all cases.

E. Self-action terms

The self-action terms (50 S, V) can be evaluated as triple series. The characteristics of these power series in $(\beta/\lambda)^2$ can be summarized as follows:

(i) The series are ascending power series, with coefficients depending on λ.

(ii) The series converge absolutely for all values of β and λ; they converge rapidly when the order of $(\beta/\lambda)^2$ becomes larger than the magnitude of β/λ.

(iii) As a function of λ, the series have nonanalyticities at $1/\lambda = N$ (N integer). For larger N, the order of the nonanalyticities as well as the order of the terms which contain the nonanalyticities also become larger. More specifically, a general term of the series contains the factor $\mathrm{Re}(1-N^2\lambda^2)^{M+1/2}(\beta/\lambda)^M$, where $M \geq N^{-1}$; thus in the nonrelativistic limit, where β/λ is constant, the nonanalyticities disappear as $\lambda \to 0$.

Up to order β^2, we have

$$\mu_{ii} \cong \frac{1}{\lambda}\left\{1+\tfrac{1}{3}\left(\frac{\beta}{\lambda}\right)^2[1-(1-\lambda^2)^{3/2}]\right\}-\frac{1-\beta^2}{\lambda^3}$$
$$\times\left\{\tfrac{1}{3}[1-(1-\lambda^2)^{3/2}]+\tfrac{1}{15}\left(\frac{\beta}{\lambda}\right)^2[\tfrac{3}{2}-2(1-\lambda^2)^{5/2}\right.$$
$$\left.+\tfrac{1}{2}(1-4\lambda^2)^{5/2}]\right\}, \tag{65S}$$

$$\mu_{ii} \cong \frac{1+\beta^2}{\lambda^3}\left\{\tfrac{1}{3}[1-(1-\lambda^2)^{3/2}]+\tfrac{1}{15}\left(\frac{\beta}{\lambda}\right)^2\right.$$
$$\times[\tfrac{3}{2}-2(1-\lambda^2)^{5/2}+\tfrac{1}{2}(1-4\lambda^2)^{5/2}]\Big\}$$
$$-\frac{1}{\lambda}\left\{(1-\lambda^2)^{1/2}+\tfrac{1}{3}\left(\frac{\beta}{\lambda}\right)^2[(1-\lambda^2)^{3/2}\right.$$
$$\left.-\lambda^2(1-4\lambda^2)^{3/2}]\right\}. \tag{65V}$$

The imaginary terms, if any, are omitted here and in the following.

Obviously the series are suitable expansions either for large λ or for small β. If $\lambda \gg 1$, we have

$$\mu_{ii} \cong \frac{1}{\lambda}, \tag{66S}$$

$$\mu_{ii} \cong \frac{1}{3\lambda^3}(1+\beta^2). \tag{66V}$$

Hence the self-action terms are small in this limit. A solution for the case of field theory can always be found in the neighborhood of the solutions for the case of action at a distance.

But if λ is not large, no conclusion can be reached in general. However, two interesting results may be obtained by assuming β^2 to be small:

(i) Close range solution: By setting $\beta = 0$ in Eqs. (65 S, V), we obtain

$$\mu_{ii} \cong \frac{1}{3\lambda^2}[3\lambda^2-1+(1-\lambda^2)^{3/2}], \tag{67S}$$

$$\mu_{ii} \cong \frac{1}{3\lambda^3}[1-(1+2\lambda^2)(1-\lambda^2)^{1/2}], \tag{67V}$$

which are constant in this limit. Hence the existence of the solution is not affected by the existence of the self-action terms. For the region

$$\lambda^2 < \tfrac{1}{2}3^{1/2} \cong 0.866, \tag{68V}$$

μ_{ii}, given by Eq. (67V), becomes negative; therefore we have one-particle solutions in this region.

(ii) Nonrelativistic limit: In this case both β and λ approach zero while the ratio β/λ remains finite. One would at first expect dependence on $(\beta/\lambda)^2$ in this limit from Eqs. (65 S, V); but it turns out that there is no such dependence in the leading terms, and

$$\mu_{ii} \cong \frac{1}{2\lambda}, \tag{69S}$$

$$\mu_{ii} \cong -\frac{1}{2\lambda}. \tag{69V}$$

Thus, the mesic self-action terms have no effect other than to change the values of the masses by amounts

$$\omega e_i^2\mu_{ii}c^{-3}\gamma^{-1} \cong \pm\tfrac{1}{2}\chi e_i^2c^{-2} = \pm\tfrac{1}{2}m\frac{e_i^2}{\hbar c}$$

(where m is the meson mass) in the nonrelativistic limit, in agreement with Bhabha's[18] and Harish-Chandra's[19] results. The interaction terms are two orders higher in c than the self-action terms. Hence, except in the one-particle solution of the vector case, the self-action terms are too small to affect the existence of the solutions, as mentioned in Sec. IIIB.

IV. DISCUSSION

In Secs. II and III, we have established the existence of exact solutions for three particles with half-retarded,

half-advanced interactions. For the case of electromagnetic and mesic interactions, they were shown in Sec. IIIB to reduce to the nonrelativistic solutions found in HII in the appropriate limit. But beyond this, we also found a number of types of solutions which have no nonrelativistic counterpart. These include solutions with all electric charges of the same sign (Sec. IIID) and solutions with all three particles on the same side of the axis of rotation (Sec. IIID), as well as planar and nonplanar two-particle solutions (Secs. IIE, IIID), and a one-particle circular solution (Secs. IIE, IIIE).

The solutions with all charges of the same sign show that a given relativistic two-particle interaction can be either repulsive or attractive, depending on the character of the entire motion. This particular feature of relativistic interactions has been noted earlier for the case of motion of a particle in an external field.[20]

The three-, two-, and one-particle solutions with all particles on the same side of the axis of rotation violate our preconceived (Newtonian) notions of conservation of momentum. However, in relativistic dynamics the conservation laws either have to include field as well as particle variables, or involve integrals over the entire motion.[21,13] Our results must be consistent with the conservation laws, since the equations of motion are;[10,9,22] calculations verifying this explicitly will be published elsewhere.

The nonplanar two-particle solutions can be used to construct examples of more than one exact solution with the same initial positions and velocities, as will be shown elsewhere.[23]

For the case of mesic interactions, we found that in the relativistic region the solutions contain nonanalyticities at $1/\lambda$ = integer (Secs. IIIC, IIID, IIIE).

One question remaining to be answered is that of the stability of the solutions, which has not been resolved even for the nonrelativistic solutions obtained in HIII. For the relativistic Kepler problem (i.e., for the planar solutions of Smith[5] and of Schild[6]) it has been investigated by Staruszkiewicz[24] and by Andersen and von Baeyer,[25] who arrived at opposite conclusions. We intend to return to this question elsewhere.

Some of the highly relativistic solutions obtained occur in the region where quantum mechanical effects cannot be ignored for the known particles and interactions. The left-hand side of Eq. (30) may be written as

$$\frac{mc^3}{e^2\omega} = \frac{\hbar c}{e^2}\,\frac{c/\omega}{\hbar/mc}.$$

The first function on the right hand side characterizes the strength of the interaction. If the particles are taken to be electrons, it is the reciprocal of the fine structure constant and thus ~137. For the strongly interacting nucleons, it is of order 1. The denominator of the second function is the Compton wavelength of the particle. The numerator is subject to our choice. If the separations between particles are large (as in the solutions discussed in Sec. IIID), the right-hand sides of the three equations represented by Eq. (30) must be of order of magnitude less than 1. In order to have this type of solution, we must have c/ω smaller than the Compton wavelength; obviously, quantum mechanical effects cannot be ignored then.

Similarly, since the value of λ can be written as $\hbar\omega/mc^2$ (where m is the meson mass), we must have $\lambda \ll 1$ in the classical limit, which precludes detection of the nonanalyticities in that limit.

The one-particle solutions, however, are not excluded by such considerations. On the other hand, their existence requires the interactions to be time-symmetric rather than retarded. It also requires that field theory be valid rather than the theory of action at a distance.[26] Furthermore, to be experimentally detectable, the motions must be stable.

The use of time-symmetric interactions was necessary to allow the existence of stationary solutions. As the prime motivation of this investigation was a study of typically relativistic features of solutions of the equations of motion of interacting particles[27] rather than to obtain results directly comparable to experiment, the question of retarded vs time-symmetric interactions is not relevant here. It should be noted, however, that for problems involving many particles the use of time-symmetric interactions does not preclude a description of radiation.[28,11] Furthermore, relativistic quantum mechanical stationary states are expected to correspond to such classical states with time-symmetric interactions.

---

*Research supported in part by the National Science Foundation.

†Present address: Department of Physics, National Central University, Chung-Li, Taiwan.

[1]L. Euler, Nova Comm. Petrop. **11**, 144 (1767); J. L. Lagrange, *Oeuvres* (Gauthier-Villars, Paris, 1873), Vol. VI, p. 229. These solutions are described in A. Wintner, *The Analytical Foundations of Celestial Mechanics* (Princeton U.P., Princeton, N.J., 1947), Chap. 5, and E. T. Whittaker, *Analytical Dynamics of Particles* (Cambridge U.P., Cambridge, 1964), 4th ed., Sec. 162.

[2]For detailed reviews see A. Wintner, Ref. 1, and Y. Hagihara, *Celestial Mechanics* (MIT, Cambridge, Mass., 1970), Vol. 1, Chap. 3.

[3]P. Havas, in Celestial Mechanics (scheduled for Vol. 7, No. 3, this paper contains the classification of the possible motions and will be referred to as HI); *ibid.* (to be submitted shortly; this paper contains a discussion of possible configurations, and will be referred to as HII).

[4]Ch. H. Müntz, Math. Z. 15, 169 (1922); the classification given is not quite complete, as discussed in HI.

[5]S. F. Smith, thesis (Lehigh University, 1960) (unpublished), Appendix A.

[6]A. Schild, Phys. Rev. **131**, 2762 (1963).

[7]S. F. Smith and P. Havas, Phys. Rev. **138**, B 493 (1965).

[8]P. Havas and J. Plebánski, Bull. Am. Phys. Soc. **5**, 433 (1960); see also P. Havas, in *Statistical Mechanics of Equilibrium and Non-Equilibrium*, edited by J. Meixner (North-Holland, Amsterdam, 1965), p. 1.

[9]H. J. Bhabha and Harish-Chandra, Proc. R. Soc. A **183**, 134 (1944) and Proc. R. Soc. A **185**, 250 (1946); Harish-Chandra, Proc. R. Soc. A **185**, 269 (1946).

[10]P. A. M. Dirac, Proc. R. Soc. A **167**, 148 (1938).

[11]P. Havas, Phys. Rev. **87**, 309 (1952).

[12]A. D. Fokker, Z. Phys. **58**, 309 (1929); J. A. Wheeler and R. P. Feynman, Rev. Mod. Phys. 21, 425 (1949).

[13]P. Havas, in *Problems in the Foundations of Physics*, edited by M. Bunge (Springer-Verlag, Berlin, 1971), p. 31. The form (11) is less general than the interactions considered in that paper, but more general than those of Refs. 11 and 12.

[14]However, they do not hold for all interactions considered in Ref. 13.

[15]G. Frobenius, Crelle's J. 132, 230 (1876); see also T. Muir, *Theory of Determinants* (MacMillan, New York, 1920), Vol. 3, p. 275.

[16]P. Havas, Phys. Rev. **91**, 997 (1953).

[17]For the details of the calculations, see D.-C. Chern, thesis (Temple University, 1971) (unpublished). There are a few errors in the formulas and discussion, and Appendix A has been revised; a corrected copy can be obtained from the authors on request.

[18]H. J. Bhabha, Proc. R. Soc. A **172**, 384 (1939).

[19]Harish-Chandra, Proc. Indian Acad. Sci. **21**, 135 (1945).

[20]J. Werle, Acta Phys. Pol. **14**, 233 (1955).

[21]J. W. Dettman and A. Schild, Phys. Rev. **95**, 1057 (1954).

[22]P. Havas, in *Recent Developments in General Relativity* (Pergamon, New York, 1962), p. 259.

[23]D.-C. Chern and P. Havas (to be published).
[24]A. Staruszkiewicz, Acta Phys. Pol. **33,** 1007 (1968).
[25]C. M. Andersen and H. C. von Baeyer, Phys. Rev. D **5,** 802 (1972).
[26]For other possible experimental distinctions between the two theories compare C. R. Mehl and P. Havas, Phys. Rev. **91,** 393 (1953); P. Havas, Phys. Rev. **93,** 882 (1954); J. E. Chatelain and P. Havas, Phys. Rev. **129,** 1459 (1963); A. D. Craft and P. Havas, Phys. Rev. **154,** 1460 (1967); A. Rosenblum and P. Havas, Phys. Rev. D **6,** 1522 (1972).
[27]For such features for particles in an external field see P. Havas, Acta Phys. Austriaca **3,** 342 (1949).
[28]J. A. Wheeler and R. P. Feynman, Rev. Mod. Phys. **17,** 157 (1945); see also P. Havas, Phys. Rev. **74,** 456 (1948) and Phys. Rev. **86,** 974 (1952).

General scaling method for electromagnetic fields with application to a matching problem

Tse Chin Mo and C. H. Papas

Electrical Engineering Department, California Institute of Technology, Pasadena, California 91109

Carl E. Baum

Air Force Weapons Laboratory, Kirtland AFB, Albuquerque, New Mexico 87117

(Received 24 October 1972)

A scaling method that reduces an electromagnetic problem described by a complicated geometry in a complicated medium to one described by a simple Cartesian geometry in a simple medium is explored and developed. This method creates and identifies an equivalent class of problems and their solutions from the Cartesian simple medium problem. We illustrate the usefulness of the method by applying it to the design of a reflectionless, distortionless, loaded matching section connecting a cylindrical and a conical coaxial waveguide, with the TEM fields being explicitly found everywhere.

I. INTRODUCTION AND SUMMARY

The transform methods in mechanics and fluid dynamics, which can carry a problem and its solution into a class of equivalent problems and solutions has been of interest for many years.[1] However, the application in electromagnetic (EM) theory of a scaling method of such similar nature has not received extensive attention and only a few works have recently been devoted to it.[2] The purpose of the present work is to investigate and develop for EM theory such a similarity or scaling transform.

In Sec. II we first develop the general theory of scaling, carrying an EM problem P into an equivalent P′, with the accompanying transformations for media, geometry, and fields. The advantage of such a procedure is, hopefully, to make the complexities of the geometry and the medium "cancel" each other in such a way that the resulting fields are simple and have known solutions. Also, some special cases of interest for 3-geometry[3] are discussed with results listed in Sec. II. Section III is devoted to the solving of a matching problem between a coaxial cylindrical and a conical waveguide.

II. THEORY OF ELECTROMAGNETIC SCALING

A. General theory

Basically, scaling is possible because the covariant divergences of the antisymmetric EM field tensors can be rewritten as ordinary divergences by including the metric determinant.[4] Let us consider a coordinate system $\{x^0 \equiv t, x^1, x^2, x^3\}$ with invariant length ds^2 and metric coefficients $g_{\mu\nu}$ (see Ref. 5):

$$ds^2 = g_{\mu\nu}\,dx^\mu dx^\nu, \qquad \mu, \nu = 0, 1, 2, 3, \tag{1}$$

where the summation convention, as in the whole text, is employed for repeated indices unless otherwise stated. For observers $\{0\}$ fixed in this frame $\{x^\mu\}$ with their spatial locations $x^i \equiv$ fixed, the Maxwell equations are[6]

$$\begin{cases} \left[\sqrt{-g}\,e^j_{(i)}\dfrac{B^{(i)}}{\sqrt{g_{00}}}\right]_{,j} + \left[\sqrt{-g}\,\eta^{ijk}e^0_{(i)}e^l_{(j)}E_{(k)}\right]_{,l} = 0 & (2) \\ \left[\sqrt{-g}\,\eta^{ikl}e^j_{(i)}e^\nu_{(k)}E_{(l)}\right]_{,\nu} = \left[\sqrt{-g}\,e^j_{(i)}\dfrac{B^{(i)}}{\sqrt{g_{00}}}\right]_{,0}, & (3) \end{cases}$$

$$\begin{cases} \left[\sqrt{-g}\,e^j_{(l)}\dfrac{D^{(l)}}{\sqrt{g_{00}}}\right]_{,j} - \left[\sqrt{-g}\,\eta^{ilk}e^0_{(i)}e^j_{(l)}H_{(k)}\right]_{,j} = \sqrt{-g}\left[\dfrac{q}{\sqrt{g_{00}}} + J^{(i)}e^0_{(i)}\right] & (4) \\ -\left[\sqrt{-g}\,\eta^{ikl}e^j_{(i)}e^\nu_{(k)}H_{(l)}\right]_{,\nu} = \left[\sqrt{-g}\,e^j_{(i)}\dfrac{D^{(i)}}{\sqrt{g_{00}}}\right]_{,0} + \sqrt{-g}J^{(i)}e^j_{(i)}. & (5) \end{cases}$$

Here we have used the following notation: $g \equiv \det(g_{\mu\nu})$; $\eta^{ijk} \equiv 0$ if (ijk) are not all different, ± 1 if (ijk) is an even or an odd order of (123); the Latin indices i, j, a, b, etc. equal 1, 2, 3; $e^\nu_{(\mu)}$ are the ν-contravariant components of the local comoving unit-tetrad vectors[7] $\mathbf{e}_{(\mu)}$ of $\{0\}$; $(\mathbf{D}, \mathbf{H})$, $(\mathbf{E}, \mathbf{B})$ are the usual EM field as seen by $\{0\}$; $D^{(l)}$, etc. are the physical vector components on $\mathbf{e}_{(l)}$ for $\{0\}$; $(q, \mathbf{J})$ are the physical charge-current source as seen by $\{0\}$; and $(\,)_{,\lambda} \equiv \partial/\partial x^\lambda(\,)$. Also we assume the medium to be linear with constitutive relations

$$D^{(i)} = \epsilon^{(i)}_{(j)}E^{(j)} + \alpha^{(i)}_{(j)}B^{(j)}, \tag{6}$$

$$H^{(i)} = \beta^{(i)}_{(j)}E^{(j)} + K^{(i)}_{(j)}B^{(j)}, \tag{7}$$

$$J^{(i)} = \sigma^{(i)}_{(j)}E^{(j)}. \tag{8}$$

If there are conducting boundaries, they are described by $F(\mathbf{x}) = 0$. The above, with appropriate boundary conditions, defines an EM problem P.

Now, a scaling can transform P into an EM problem P′ with simple Cartesian geometry and correspondingly scaled medium properties, sources, and boundary conditions as the following. If we define the scaled "fictitious" EM fields $(\mathbf{e}, \mathbf{b})$, $(\mathbf{h}, \mathbf{d})$ and their Cartesian components e^i, b^i, h^i, and d^i by

$$e^j \equiv \tfrac{1}{2}\sqrt{-g}\,\eta^{jab}\eta^{ikl}e^{(i)a}e^{(k)b}E^{(l)} \tag{9}$$

$$b^j \equiv \sqrt{-g}\,e^{(l)j}\left[\frac{B^{(l)}}{\sqrt{g_{00}}} + \eta^{ilk}e^{(i)0}E^{(k)}\right], \tag{10}$$

$$d^j \equiv -\sqrt{-g}\,e^{(l)j}\left[\frac{D^{(l)}}{\sqrt{g_{00}}} - \eta^{ilk}e^{(i)0}H^{(k)}\right] \tag{11}$$

$$h^j \equiv \tfrac{1}{2}\sqrt{-g}\,\eta^{jab}\eta^{ikl}e^{(i)a}e^{(k)b}H^{(l)} \tag{12}$$

in the coordinate frame $\{x^\mu\} \equiv \{t, x^1, x^2, x^3\}$, which is *now taken* to be Cartesian with

$$dS^2 = dt^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2, \tag{13}$$

then the Maxwell equations (2) to (5) become respectively

$$\begin{cases} \nabla \cdot \mathbf{b} = 0 & (14) \\ \nabla \times \mathbf{e} = -\dfrac{\partial \mathbf{b}}{\partial t}, & (15) \end{cases}$$

$$\begin{cases} \nabla \cdot \mathbf{d} = \rho & (16) \\ \nabla \times \mathbf{h} = \mathbf{j} + \dfrac{\partial}{\partial t}\mathbf{d}, & (17) \end{cases}$$

as in the simple Cartesian sense. The scaled medium, corresponding to (6) to (8), is still linear and has as constitutive relations

$$\begin{aligned} d^j &= \xi^{js} e^s + A^{jb} b^b \\ &\equiv -e^{(l)j}\hat{e}^{(n)b}\left\{\hat{e}^{(m)a}\eta^{abs}\left[\frac{1}{2}\eta^{kmn}\left(\frac{\epsilon^{(l)}_{(k)}}{\sqrt{g_{00}}} - \eta^{ilp}e^{(i)0}\beta^{(p)}_{(k)}\right)\right.\right. \\ &\quad \left.\left. - \sqrt{g_{00}}e^{(m)0}\left(\frac{\alpha^{(l)}_{(m)}}{\sqrt{g_{00}}} - \eta^{ilp}e^{(i)0}K^{(p)}_{(n)}\right)\right] e^s \right. \\ &\quad \left. - \sqrt{g_{00}}\left[\frac{\alpha^{(l)}_{(n)}}{\sqrt{g_{00}}} - \eta^{ilp}e^{(i)0}K^{(p)}_{(n)}\right] b^b \right\}, \end{aligned} \tag{18}$$

$$\begin{aligned} h^j &= B^{js}e^s + \lambda^{jt}b^t \\ &\equiv \frac{1}{2}\eta^{jab}\eta^{ikl}e^{(i)a}e^{(k)b}\hat{e}^{(p)t}\left[\hat{e}^{(m)r}\eta^{rts}\left(\frac{\beta^{(l)}_{(n)}}{2}\eta^{nmp}\right.\right. \\ &\quad \left.\left. - \sqrt{g_{00}}K^{(l)}_{(p)}e^{(m)0}\right) e^s - \sqrt{g_{00}}K^{(l)}_{(p)}b^t\right], \end{aligned} \tag{19}$$

$$j^k = \Sigma^{kt}e^t \equiv -\tfrac{1}{2}e^{(i)k}\sigma^{(i)}_{(j)}\eta^{jmn}\hat{e}^{(m)a}\hat{e}^{(n)b}\eta^{abt}e^t, \tag{20}$$

where $e^{(i)k} \equiv -e^k_{(i)}$ and $\hat{e}^{(m)a} \equiv [e^{(i)j}]^{-1}$ exist since $\det[e^{(i)j}] \neq 0$. The source is scaled by[8]

$$\begin{cases} \rho = \sqrt{-g}\left(\dfrac{q}{\sqrt{g_{00}}} + J^{(i)}e^0_{(i)}\right) \\ j^k = \sqrt{-g}J^{(i)}e^k_{(i)}. \end{cases} \tag{21}$$

The corresponding boundary conditions are given through (9)–(12) and (18)–(19) to regulate the boundary behavior of the "fictitious" fields at the *same mathematical* boundaries with $\{x^\mu\}$ *interpreted as Cartesian coordinates*. For example, conducting boundaries are still described by the surfaces $F(\mathbf{x}) = 0$, on which e satisfies $N^i\hat{e}^{(l)j}\eta^{kjm}e^m\hat{e}^{(l)k} = 0$ where N^i is normal to $F(\mathbf{x}) = 0$.

Apparently, the reduction of P to P′ with greatly simplified geometry and differential equations is achieved at the expense of the much complicated medium properties. However, we must realize first that the "fictitious" fields and the "fictitious" problem P′ are the equivalent of and as real as the "real" problem P, and they can play a reverse role at our disposal.[9] Thus we can require the apparently complicated medium properties (18)–(20) to be simple enough so that we know the solution of the EM problem P′. Then, through the inverse scaling P′ → P, we obtain the whole class of P and their solutions with each of them corresponding to a particular choice of $g_{\mu\nu}$.

The inverse scaling P′→ P has, corresponding to (9)–(12),

$$\begin{cases} E^{(l)} = (1/2\sqrt{-g})\eta^{lmn}\hat{e}^{(m)a}\hat{e}^{(n)b}\eta^{abs}e^s & (22) \\ B^{(l)} = (-\sqrt{g_{00}}/\sqrt{-g})\{\hat{e}^{(l)j}b^j + e^{(m)0}\hat{e}^{(m)a}\hat{e}^{(l)b}\eta^{abs}e^s\}, & (23) \end{cases}$$

$$\begin{cases} D^{(l)} = (-\sqrt{g_{00}}/\sqrt{-g})\{\hat{e}^{(l)j}d^j - e^{(m)0}\hat{e}^{(m)a}\hat{e}^{(l)b}\eta^{abs}h^s\} & (24) \\ H^{(l)} = (1/2\sqrt{-g})\eta^{lmn}\hat{e}^{(m)a}\hat{e}^{(n)b}\eta^{abs}h^s, & (25) \end{cases}$$

and corresponding to (18) to (20)

$$\begin{aligned} D^{(l)} &= -\sqrt{g_{00}}\hat{e}^{(l)b}e^{(p)k}\{[\tfrac{1}{2}(\xi^{bi} \\ &\quad - e^{(m)0}\hat{e}^{(m)a}\eta^{abc}B^{ci})\eta^{ikd}\eta^{pns}e^{(n)d} \\ &\quad - (A^{bk} - e^{(m)0}\hat{e}^{(m)a}\eta^{abc}\lambda^{ck})\eta^{ips}e^{(i)0}]E^{(s)} \\ &\quad - (1/\sqrt{g_{00}})(A^{bk} - e^{(m)0}\hat{e}^{(m)a}\eta^{abc}\lambda^{ck})B^{(p)}\}, \end{aligned} \tag{26}$$

$$\begin{aligned} H^{(l)} &= \tfrac{1}{2}\eta^{lmn}\hat{e}^{(m)a}\hat{e}^{(n)b}\eta^{abc}e^{(i)s}\{\eta^{inp}[\frac{B^{cj}}{2}\eta^{jsd}e^{(n)d} \\ &\quad + \lambda^{cs}e^{(n)0}]E^{(p)} - \lambda^{cs}[B^{(i)}/\sqrt{g_{00}}]\}, \end{aligned} \tag{27}$$

$$J^{(l)} = -\tfrac{1}{2}\hat{e}^{(l)k}\Sigma^{kj}\eta^{jab}\eta^{imn}e^{(i)a}e^{(m)b}E^{(n)}. \tag{28}$$

Also, the inverse of (21) is

$$\begin{cases} q = (\sqrt{g_{00}}/\sqrt{-g})(\rho - \hat{e}^{(i)k}e^{(i)0j^k}) \\ J^{(i)} = (-1/\sqrt{-g})\hat{e}^{(i)k}j^k. \end{cases} \tag{29}$$

B. Special geometries and media

For the P → P′ scaling, the mixed constitutive terms A^{ij} and B^{ij} in (18) and (19) are caused partly by the medium's intrinsic constitutive mixture $\alpha^{(i)}_{(j)}$ and $\beta^{(i)}_{(j)}$ in (6) and (7), and partly by the nontime-orthogonality term $e^{(i)0}$ of the frame $\{x^\mu\}$ with $g_{0i} \neq 0$. If both $\alpha^{(i)}_{(j)} = \beta^{(i)}_{(j)} = 0$ for the $\{x^\mu\}$ of problem P, then P′ has the simplified version of (18) and (19),

$$d^j = -e^{(l)}\hat{e}^{(m)a}\hat{e}^{(n)b}\eta^{abs}(\eta^{kmn}/2\sqrt{g_{00}})\epsilon^{(l)}_{(k)}e^s, \tag{30}$$

$$h^j = (-\sqrt{g_{00}}/2)\eta^{jab}\eta^{ikl}e^{(i)a}e^{(k)b}\hat{e}^{(p)t}K^{(l)}_{(p)}b^t. \tag{31}$$

If, in addition to the above, the original frame $\{x^\mu\}$ has a diagonal metric, i.e., $g_{\mu\nu} \equiv 0$ for $\mu \neq \nu$, we have[10]

$$e^\nu_{(\mu)} = \delta^\nu_\mu/\sqrt{|g_{\mu\mu}|} \equiv \delta^\nu_\mu/g_\mu. \tag{32}$$

Then (30), (31), (20) of the scaled medium further reduce to

$$\begin{aligned} \begin{pmatrix} d^1 \\ d^2 \\ d^3 \end{pmatrix} &= \frac{1}{\sqrt{g_{00}}}\begin{pmatrix} g_2g_3 & & \bigcirc \\ & g_3g_1 & \\ \bigcirc & & g_1g_2 \end{pmatrix}\begin{pmatrix} \epsilon^{(1)}_{(1)} & \epsilon^{(1)}_{(2)} & \epsilon^{(1)}_{(3)} \\ \epsilon^{(2)}_{(1)} & \epsilon^{(2)}_{(2)} & \epsilon^{(2)}_{(3)} \\ \epsilon^{(3)}_{(1)} & \epsilon^{(3)}_{(2)} & \epsilon^{(3)}_{(3)} \end{pmatrix} \\ &\quad \times \begin{pmatrix} 1/g_1 & & \bigcirc \\ & 1/g_2 & \\ \bigcirc & & 1/g_3 \end{pmatrix}\begin{pmatrix} e^1 \\ e^2 \\ e^3 \end{pmatrix}, \end{aligned} \tag{33}$$

$$b^i \equiv \eta^{ij}h^j = [\text{in (33), replace } e^i \text{ by } h^i \text{ and } \epsilon^{(i)}_{(j)} \text{ by } \mu^{(i)}_{(j)}], \tag{34}$$

$$(1/\sqrt{g_{00}})j^i = [\text{in (33), replace } \epsilon^{(i)}_{(j)} \text{ by } \sigma^{(i)}_{(j)}], \tag{35}$$

where $\mu \equiv K^{-1}$ such that $B^{(i)} = \mu^{(i)}_{(j)} H^{(j)}$. For this case, (9)–(12) which link fields of P′ and P reduce simply to (no summation here)

$$\begin{cases} e^j = g_0 g_j E^{(j)} & (36) \\ b^j = g_1 g_2 g_3 [B^{(j)}/g_j], & (37) \\ d^j = (g_1 g_2 g_3/g_j) D^{(j)} & (38) \\ h^j = g_0 g_j H^{(j)}. & (39) \end{cases}$$

The inverse of (33)–(39) for P′→ P is obvious.

C. Remarks on medium restricted scalings in Euclidean 3-space

Physically limiting ourselves to certain class of media puts a restriction to the scaling. Preparing for a particular application we will treat in Sec. III, let us examine such limits in the following in a Euclidean 3-space and choosing *orthogonal coordinates* with $g_0 \equiv 1$.

If we require both P and P′ to have isotropic media, i.e., $\epsilon^{(i)}_{(j)} = -\epsilon\delta^{ij}$ and $\xi^{ij} = -\xi\delta^{ij}$, etc., then (33)–(35) imply

$$g_1 = g_2 = g_3 \tag{40}$$

and

$$\xi/\epsilon = \eta/\mu = \Sigma/\sigma = g_1. \tag{41}$$

In Euclidean 3-space, there exist only two[11] such coordinate frames, namely: Cartesian with $dS^2 = dt^2 - dx^2 - dy^2 - dz^2$, and inverse sphere with $dS^2 = dt^2 - a^4(dx'^2 + dy'^2 + dz'^2)/(x'^2 + y'^2 + z'^2)^2$.

If we require both P and P′ to have uniaxial media, i.e.,

$$\epsilon^{(i)}_{(j)} = \begin{pmatrix} \epsilon & & \\ & \epsilon & \\ & & \epsilon_3 \end{pmatrix}, \quad \xi^{ij} = \begin{pmatrix} \xi & & \\ & \xi & \\ & & \xi_3 \end{pmatrix}, \text{ etc., (42)}$$

then (33)–(35) imply

$$g_1 = g_2, \tag{43}$$

$$\xi/\epsilon = \eta/\mu = \Sigma/\sigma = g_3, \tag{44}$$

$$\xi_3/\epsilon_3 = \eta_3/\mu_3 = \Sigma_3/\sigma_3 = g_1^2/g_3. \tag{45}$$

For this case, the wave "impedances" $(\mu/\epsilon)^{1/2}$ and $(\mu_3/\epsilon_3)^{1/2}$ are unchanged in the scaling. This property will be used in Sec. III. Also, a simple calculation gives the results that we can obtain from any orthogonal coordinates (v^1, v^2, v^3) with $d\sigma^2 = f_1^2(dv^1)^2 + f_2^2(dv^2)^2 + f_3^2(dv^3)^2$ a coordinate system (x^1, x^2, x^3) having

$$\begin{cases} d\sigma^2 = f_2^2[(dx^1)^2 + (dx^2)^2] + \{f_3^2(dx^3)^2/[F'(x^3)]^2\} & (46) \\ x^1 \equiv \int (f_1/f_2) dv^1 + \text{const}, \quad x^2 \equiv v^2, \quad x^3 \equiv F(v^3), & (47) \end{cases}$$

if[12]

$$f_1/f_2 = \text{function of } v^1 \text{ only}. \tag{48}$$

If we require only that both P and P′ have diagonal media, then no restriction on g_i is imposed. As a trivial example of this and to illustrate the P′→ P procedure, consider P′ as a parallel plate waveguide with plate at $x = a$ and $x = b$, filled with uniform η, ξ-simple medium, and having a TEM z-propagating wave $E^{(x)} = (\eta/\xi)^{1/2} H^{(y)} = \exp(i\omega\sqrt{\xi\eta}\, z)$. Take the scaling $(x, y, z) \to (\theta, \phi, r)$ as spherical coordinate, then (33)–(39) right away gives a legitimate P and its solution. The P is a conical waveguide with cones at $\theta = a/c_1$ and $\theta = b/c_1$ and filled with diagonal simple medium

$$\frac{\epsilon^{(i)}_{(j)}}{\xi} = \frac{\mu^{(i)}_{(j)}}{\eta} = \begin{pmatrix} c_2/(c_1 \sin\theta) & & \\ & (c_1 \sin\theta)/c_2 & \\ & & c_1 c_2/(r^2 \sin\theta) \end{pmatrix} \tag{49}$$

and has a TEM propagation with $rE^{(\Theta)}/c_1 = (\eta/\xi)^{1/2}$ $(r \sin\theta/c_2) H^{(\Phi)} = \exp(i\omega\sqrt{\xi\eta}\, r)$. Here c_1 and c_2 are arbitrary length constants.

III. PERFECT μ,ϵ-LOADED MATCH BETWEEN CYLINDRICAL AND CONICAL COAXIAL WAVEGUIDES

To illustrate the use of the scaling method, we here apply it to examine and obtain a reflectionless and distortionless perfect match between a cylindrical (region I) and a conical (region III) coaxial waveguide. The given situation P in region I (see Fig. 1), which has perfectly conducting boundaries at $\rho = A$ and $\rho = B$[13] and is filled with ϵ, μ-simple uniform medium, is a TEM incident wave with $E^{(\rho)} = H^{(\Phi)}(\mu/\epsilon)^{1/2} = [\exp(i\omega\sqrt{\epsilon\mu}\, z - i\omega t)]/\rho$. The solution of the problem may not be unique, and it is the simplest solution we want to find. Of course, from here on we use Euclidean 3-space and choose $g_{00} \equiv 1$, $g_{0i} \equiv 0$.

A. Selection and design of the scaling match

For mathematical simplicity we examine only orthogonal coordinates, and because of the nature of the problem we choose rotational ones in 3-space. Also, since only TEM wave is of interest, we designate the x^3-direction as the propagation direction and can try the "transverse isotropy" (42)–(45) that leaves ϵ_3, μ_3 free as the simplest general possibility.

Now the problem is to find a common P′ that underlies

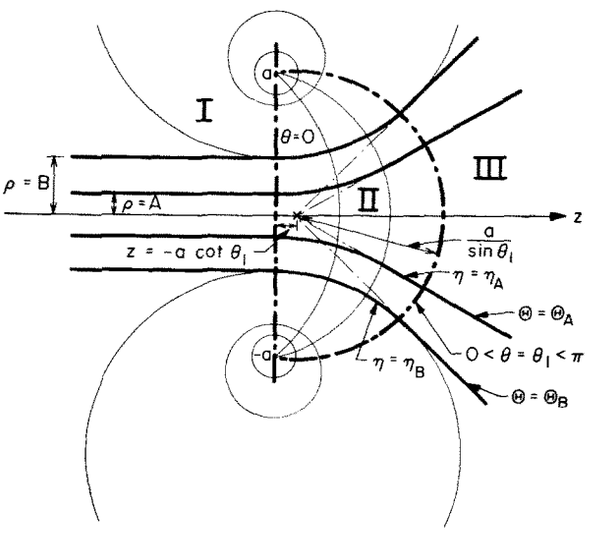


FIG. 1. Loaded perfect match between a cylindrical and a conical coaxial waveguide.

the whole range of regions I, III, and II which is the matching section to be found. Then a $P' \to P$ scaling will, hopefully, give back the desired result by choosing different scalings for different regions. In region I the P' as implied by the given P in region I is fixed. Through the use of (36)–(39) and (42)–(48) the P' is therefore specified by a TEM wave

$$e^1 = \exp(i\omega\sqrt{\mu\epsilon}\, x^3)/c_1, \tag{50}$$

$$h^2 = (\epsilon/\mu)^{1/2} \exp(i\omega\sqrt{\mu\epsilon}\, x^3)/c_1 \tag{51}$$

in a Cartesian coordinate (x^1, x^2, x^3) and in a Cartesian parallel plate waveguide which has boundaries at $x^1 = c_1 \ln(A/\rho_0)$ and $x^1 = c_1 \ln(B/\rho_0)$ and is filled with the medium

$$\frac{\xi^{ij}}{\epsilon} = \frac{\eta^{ij}}{\mu} = \frac{\Sigma^{ij}}{\sigma} = \begin{pmatrix} 1 & & \bigcirc \\ & 1 & \\ \bigcirc & & (\rho_0^2/c_1^2)\exp(2x^1/c_1) \end{pmatrix}. \tag{52}$$

Here in I the $(x^1, x^2, x^3) \equiv (c_1 \ln(\rho/\rho_0), c_1\phi, z)$ is obtained by using the cylindrical (ρ, ϕ, z) as the (v^1, v^2, v^3) to furnish the scaling, where c_1, ρ_0 are some length constants to be restricted later.

To choose a $P' \to P$ for region II from the above P', we first realize from the geometry of I and III that we need a coordinate system for (v^1, v^2, v^3) whose constant coordinate-surfaces can carry a plane into spheres convex w.r.t. the plane. The toroidal coordinates (η, ϕ, θ) provide just that.[14] As to region III, of course, we use the spherical coordinates (Θ, ϕ, r) as the (v^1, v^2, v^3).

Now, by using such (v^i) to convert the (x^i) of P' and by making sure that the x^i are continuous at junctions, simple straightforward calculations from (42) to (48) and the inverse of (36) to (39) give the required perfect matching. This match is shown in Fig. 1 and has the properties listed in Table I.

Notice that in the above table $F(\theta)$, $G(r)$ in general are dimensionless arbitrary smooth functions that satisfy $F(0) = 0$, $G(a/\sin\theta_1) = \tan\theta_1/2$, and the choices as shown are the results of requiring $\epsilon^{(1)}_{(1)} \geq \epsilon$ in II and $\epsilon^{(1)}_{(1)} = \epsilon$ in III. Also ρ_0 and a $(> B > A > 0)$ are arbitrary length constants. Arrows at the top of the table denote boundaries that divide the regions.

B. Remarks and discussion

Firstly, since only TEM wave exists, the match can be realized by employing an isotropic medium using its transverse isotropy. Physically, it is obvious that there should be no reflection caused by μ, ϵ discontinuities and changes since η/ξ and therefore μ/ϵ is an invariant constant throughout the scaling. Also obviously, there should be no distortion because the smaller $(\mu\epsilon)$ near the inner matching conductor $\eta = \eta_A$ bends the plane phase front from the cylindrical I into a spherical phase front to match the spherical III.

TABLE I.

Regions / Quantities	I	$z = 0 \to \leftarrow \theta = 0$ II	$\theta = \theta_1 \to \leftarrow r = a/\sin\theta_1$ $\theta < \theta_1 < \pi$ III
x^1	$a \ln(\rho/\rho_0)$	$a \ln[th(\eta/2)] + b$	$a \ln\left(\frac{\tan(\Theta/2)}{\tan(\theta_1/2)}\right) + b$
x^2	$a\phi$	$a\phi$	$a\phi$
x^3	z	$[\equiv aF(\theta)] = a\tan(\theta/2)$ and $F(0) = 0$	$[\equiv aG(r)] = r - a\cot\theta_1$ and $G(a/\sin\theta_1) = \tan(\theta_1/2)$
$g_1 = g_2$	(ρ/a)	$\sinh\eta/(\cosh\eta + \cos\theta)$	$r(\sin\Theta)/a$
g_3	1	$(\equiv 1/(\cosh\eta + \cos\theta)F'(\theta))$ $= (1 + \cos\theta)/(\cosh\eta + \cos\theta)$	$[\equiv 1/aG'(r)]$ 1
Boundaries			
$x^1 = a\ln(A/\rho)$ to $x^1 = a\ln(B/\rho_0)$	$\rho = A$ to $\rho = B$ $0 < A < B < a$	$\eta = 2\tanh^{-1}A/a$ to $\eta = 2\tanh^{-1}B/a$	$\Theta = 2\tan^{-1}[(A/a)\tan(\theta_1/2)$ to $\Theta = 2\tan^{-1}[(B/a)\tan(\theta_1/2)$
Media			
$\epsilon^{(i)}_{(j)}/\epsilon = \mu^{(i)}_{(j)}/\mu = \sigma^{(i)}_{(j)}/\sigma$ $\left[\epsilon^{(1)}_{(1)} = \epsilon^{(2)}_{(2)} = \xi/g_3\right.$ and $\left.\epsilon^{(3)}_{(3)} = \xi_3(g_3/g_1^2)\right]$	δ^{ij}	$\begin{pmatrix} \frac{(\cosh\eta + \cos\theta)}{(1 + \cos\theta)}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & \bigcirc \\ \bigcirc & \frac{\sinh\eta(1 + \cosh\eta)}{(1 + \cosh\eta)^2} \end{pmatrix}$	δ^{ij}
	[now $\sigma^{(i)}_{(j)} \equiv 0$ for all regions, since $\sigma = 0$]		
Fields			
$\mathbf{E}\, e^{i\omega t}$	$\mathbf{e}_{(\rho)}(e^{i\omega\sqrt{\mu\epsilon}z}/\rho)$	$\mathbf{e}_{(\eta)}\frac{(\cosh\eta + \cos\theta)}{a\sinh\eta}$ $\times e^{i\omega\sqrt{\mu\epsilon}\,a\tan(\theta/2)}$	$\mathbf{e}_{(\Theta)}(1/r\sin\theta)$ $\times e^{i\omega\sqrt{\mu\epsilon}(r - a\cot\theta_1)}$
$\mathbf{H}\, e^{i\omega t}$	$\mathbf{e}_{(\phi)}\sqrt{\epsilon/\mu}$ $\times \frac{e^{i\omega\sqrt{\mu\epsilon}z}}{\rho}$	$\mathbf{e}_{(\phi)}\sqrt{\frac{\epsilon}{\mu}}\frac{(\cosh\eta + \cos\theta)}{a\sinh\eta}$ $\times e^{i\omega\sqrt{\mu\epsilon}\,a\tan(\theta/2)}$	$\mathbf{e}_{(\phi)}(\epsilon/\mu)^{1/2}$ $\times \frac{e^{i\omega\sqrt{\mu\epsilon}(r - a\cot\theta_1)}}{r\sin\theta}$

The match so obtained is by no means unique. It is merely the simplest one. In fact, any rotational coordinates that can match smoothly with (ρ, ϕ, z) at left and (Θ, ϕ, r) at right can be used to provide (x^1, x^2, x^3) in II for $P' \to P$.

Concerning the realizability of the loading the required taper of μ is difficult. Since for normal incidence there is no reflection if and only if the impedance $(\epsilon/\mu)^{1/2}$ is constant and yet we need $(\mu\epsilon)$ to vary to furnish a transition, it is *impossible* to achieve a perfect match by an *orthogonal* scaling with only a varied ϵ and a fixed μ. But oblique incidence immediately suggests itself a Brewster angle tapering which may provide a perfect match through a nonorthogonal scaling with a fixed μ and varying ϵ. This and other possibilities of fixed-μ match are currently being investigated.

ACKNOWLEDGMENT

We acknowledge the support from the Boeing Co., and the partial support from the U.S. Air Force Office of Scientific Research under grant No. AFOSR-70-1935 during the investigation of this work.

---

[1]For general principle see, e.g., L. D. Landau and E. M. Lifshitz, *Fluid Dynamics* (Pergamon, New York, 1959), Secs. 19, 53, 118, 119; also, *Mechanics* (Pergamon, New York, 1960), Sec. 10; for an earlier example of such transforms, see E. J. Routh, Proc. Lond. Math. Soc. **12,** 73 (1881).

[2]For the theory of conformal mapped waveguides see, e.g., F. E. Borgnis and C. H. Papas, "Electromagnetic Waveguides and Resonators," in *Handbuch der Physik* (Springer-Verlag, Berlin, 1958), 16, 358; for some examples see F. J. Tischer, Proc. IEEE 51, 1050 (1963); 53, 168 (1965); also J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill, New York, 1941), p. 217; also P. Krasnooshkin, J. Phys. USSR **10,** 434 (1946); for an approach using invariance groups in differential forms, see B. K. Harrison and F. B. Estabrook, J. Math. Phys. **12,** 653 (1971); for a frequency scaling of reflection, see J. H. Davis and J. R. Cogdell, IEEE Trans. Antennas Propag. **19,** 58 (1971); for a scaling for reducing constantly moving uniform simple media, see R. J. Pogorzelski, IEEE Trans. Antennas Propag. **19,** 455 (1971).

[3]For most recent treatments on scaling see C. E. Baum, EMP Sensor and Simulation Notes DASA **32,** 1800 (1967), and Ph.D. thesis, Caltech Antenna Lab. Report **47,**California Institute of Technology (1968). For an example of tapering the dielectric to suit propagation, see P. L. Uslenghi, IEEE Trans. Antennas Propag. **17,** 644 (1969).

[4]See any text on relativistic electrodynamics, e.g., V. Fock, *The Theory of Space, Time and Gravitation* (Pergamon, New York, 1964), Sec. 24.

[5]Notice that the signature $(+---)$ is used therefore for special relativity flat space-time $dS^2 = dt^2 - dx^2 - dy^2 - dz^2$. Also geometrized MKS unit with $\mu_0 = \epsilon_0 = 1$ for vacuum is used for simple formalisms. For retrieval to MKS see, e.g., attached table in T. C. Mo, Radio Sci. **6,** 673 (1971). For any applications in special relativistic EM theory, only insert appropriate powers of c $(3 \times 10^8$ meters/sec) in the final answer to fix dimensions right.

[6]T. C. Mo, J. Math. Phys. **11,** 2589 (1970), Sec. 4. Also for 3-vectors, $D^{(l)} \equiv -D_{(l)}$, etc.

[7]See Ref. 6, Sec. 2; and any textbook with tensor calculus, e.g., J. L. Synge, *General Relativity* (Interscience, New York, 1960), Sec. 3.

[8]Notice that the $(\rho, \mathbf{j})/\sqrt{g_{00}}$ of this scaled current-density is the contravariant 4-vector component J^μ in the frame $\{x^\mu\}$ of (1).

[9]For both P and P' to be really physical, care must be taken into account to make x^i have dimensions of length such that $g_{\mu\nu}$ are dimensionless pure numbers.

[10]Here we use g_μ in the diagonal $dS^2 = g_0^2 dt^2 - (g_i)^2 (dx^i)^2$, instead of the conventional notation $dS^2 = h_0^2 - (h_i)^2 (dx^i)^2$ just to avoid confusion with the scaled magnetic field **h**.

[11]Ref. 3, Antenna Report **47,** p. 76.

[12]Notice that for geometry (46), $x^3 = \text{const}$ is either a sphere or a plane; see Ref. 11, p. 83; and L. P. Eisenhart, *A Treatise on the Differential Geometry of Curves and Surfaces* (Dover, New York, 1960), p. 449.

[13]The ρ in cylindrical $(\rho\phi z)$ coordinates from here on should not cause confusion with the previous charge density ρ.

[14]P. Moon and D. E. Spencer, *Field Theory Handbook* (Springer-Verlag, Berlin, 1961), p. 112.

Extension of a compact Lorentz manifold*

J. G. Miller† **and M. D. Kruskal**

Program in Applied Mathematics, Princeton University, Princeton, New Jersey 08540
(Received 18 July 1972)

For a certain geodesically incomplete, compact Lorentz manifold T, we construct an analytic non-Hausdorff extension in which no geodesic bifurcates. The extension is geodesically incomplete, but is a maximal analytic Lorentz manifold in the sense that any further analytic extension has bifurcating geodesics. We also obtain a maximal analytic Hausdorff extension of the universal covering space of T. The latter extension is geodesically complete.

The positive definite metric tensor of a connected Hausdorff Riemannian manifold induces a metric topology on the manifold equivalent to the manifold topology.[1] Since a Hausdorff Riemannian manifold is a metric space, Cauchy sequences can be defined. The Hopf–Rinow theorem[2] on connected Hausdorff Riemannian manifolds states that Cauchy completeness is equivalent to geodesic completeness. An immediate consequence of this theorem is that a compact, connected Hausdorff Riemannian manifold is geodesically complete.

This last statement is not true for pseudo-Riemannian manifolds. Misner[3] has given a metric with Lorentz signature defined on the (compact) 2-torus T that is geodesically incomplete. A compact, connected Hausdorff manifold has no Hausdorff extension, i.e., it cannot be properly imbedded (with codimension zero) in another connected Hausdorff manifold. Therefore, any extension of Misner's example must be non-Hausdorff.

Given a compact Riemannian or pseudo-Riemannian manifold, it is trivial to construct non-Hausdorff extensions that have bifurcating geodesics (e.g., simply duplicate any point.) The Hopf–Rinow theorem implies that any non-Hausdorff extension of a compact, connected Hausdorff Riemannian manifold has at least one bifurcating geodesic. We will show that compact Hausdorff pseudo-Riemannian manifolds may have non-Hausdorff extensions in which no geodesic bifurcates. This will be accomplished by constructing an analytic non-Hausdorff extension of Misner's example. The extension is geodesically incomplete, but is a maximal analytic Lorentz manifold in the sense that any further analytic extension has bifurcating geodesics. We also obtain a maximal analytic Hausdorff extension of the universal covering space of T. The latter extension is geodesically complete.

Consider the metric[3]

$$ds^2 = \cos x\,(dy^2 - dx^2) + 2\sin x\,dx dy \tag{1}$$

defined on the 2-torus T obtained from the (x, y) plane by imposing the identification[4]

$$(x, y) = (x + 4m\pi, y + n) \quad \text{for all integers } m, n. \tag{2}$$

The incomplete null geodesics[3] $y = 2\ln\lambda$, $x = (k + \frac{1}{2})\pi$ where λ is an affine parameter and k is an integer, are Killing horizons[5,6] with respect to the Killing vector field $X \equiv \partial_y \equiv \partial/\partial y$. The usual procedure for obtaining an extension of the two-dimensional Lorentz manifold is to introduce null coordinates. The transformation

$$\xi = -\sin(\tfrac{1}{2}x - \tfrac{1}{4}\pi)e^{-y/2}, \qquad \eta = \cos(\tfrac{1}{2}x - \tfrac{1}{4}\pi)e^{-y/2} \tag{3}$$

takes metric (1) into

$$ds^2 = 8d\xi d\eta/(\xi^2 + \eta^2). \tag{4}$$

Metric (4) is analytic on the punctured plane $M = \{(\xi, \eta) \in R^2 \,|\, \xi^2 + \eta^2 \neq 0\}$. The Killing vector field in (ξ, η) coordinates is

$$X = -\tfrac{1}{2}(\xi\partial_\xi + \eta\partial_\eta), \tag{5}$$

and the one-parameter group of isometries $\{U_t\}$ generated by X is given by

$$U_t(\xi, \eta) = (\xi e^{-t/2}, \eta e^{-t/2}). \tag{6}$$

Let $G = \{U_n\}$, a discrete subgroup of $\{U_t\}$. Then T is isometric to the quotient space M/G.

In the (ξ, η) coordinate patch, the geodesics of metric (4) that approach (0, 0) are complete, but there are geodesics that run off to infinity with finite affine length. Consider the transformation $u = \tan^{-1}\xi$, $v = \tan^{-1}\eta$, which maps M onto to punctured square $|u| < \pi/2$, $|v| < \pi/2$, $u^2 + v^2 \neq 0$. In (u, v) coordinates (4) and (5) become

$$ds^2 = \frac{8dudv}{\sin^2 u\cos^2 v + \cos^2 u\sin^2 v} \tag{7}$$

and

$$X = -\tfrac{1}{2}(\sin u\cos u\,\partial_u + \sin v\cos v\,\partial_v). \tag{8}$$

Metric (7) and the Killing vector field (8) extend analytically to $\bar{M} = R^2 - \{(m\pi/2, n\pi/2) \,|\, m + n \text{ even}\}$. It can easily be seen that $\bar{M}$ is geodesically complete. The set of zeros of X is $N = \{(m\pi/2, n\pi/2) \,|\, m + n \text{ odd}\}$. The Killing vector field X is a timelike rotation near its zeros, and the null orbits of $\{U_t\}$ are null geodesics and are branches of bifurcate Killing horizons[6] (see Fig. 1). Let $\bar{G}$ be the analytic continuation of G to $\bar{M}$. Then $\bar{M}/\bar{G}$ is not a T_1 topological space (because any zero of X is approached by an infinite sequence of points which will all be identified under $\bar{G}$) and hence not a manifold, but $(\bar{M} - N)/\bar{G}$ is a non-Hausdorff Lorentz manifold in which no geodesic bifurcates.[7] Since $\bar{M}$ is geodesically complete, we conclude that $(\bar{M} - N)/\bar{G}$ is a maximal analytic extension of T with no bifurcating geodesics.

From (3) we see that the universal covering space of M

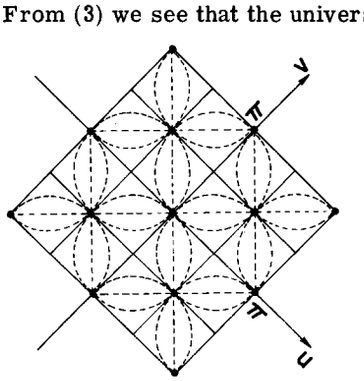


FIG. 1. The manifold $\bar{M}$. The solid lines are bifurcate Killing horizons. Each one of the four branches of a bifurcate Killing horizon is a null orbit of $\{U_t\}$, and the four branches meet at a fixed point of U_t (zero of X). The dashed lines are spacelike and timelike orbits of $\{U_t\}$.

is isometric to metric (1) defined on the (x, y) plane. Thus, the universal covering space of $\bar{M}$ is an analytic Hausdorff extension of the universal covering space of T. Since $\bar{M}$ is geodesically complete, its universal covering space is complete and, hence, is a maximal Hausdorff manifold.

---

*Work supported in part by the National Science Foundation under Grant No. GP 30799 X.

†Present address: Department of Mathematics, University of California, Los Angeles, California 90024.

[1]N. Hicks, *Notes on Differential Geometry* (Van Nostrand, Princeton, N. J., 1965), p. 70.

[2]H. Hopf and W. Rinow, Comment. Math. Helv. **3,** 209 (1931).

[3]C. Misner, J. Math. Phys. **4,** 924 (1963). See also M. Fierz and R. Jost, Helv. Phys. Acta **38,** 137 (1965); R. Geroch, J. Math. Phys. **9,** 450 (1968); D. Feinblum, in *Relativity and Gravitation*, edited by C. Kuper and A. Peres (Gordon and Breach, New York, 1971), p. 155.

[4]Imposing the identification $(x, y) = (x + ma, y + nb)$ for all intergers m and n, a being an integral multiple of 2π and b an arbitrary constant, on metric (1) defines an analytic Lorentz manifold. The various identifications lead to qualitatively similar but not isometric results. We specialize here to the case $a = 4\pi$, $b = 1$ only for convenience. Misner originally chose $a = 2\pi$, $b = 2\pi$.

[5]B. Carter, J. Math. Phys. **10,** 70 (1969).

[6]R. H. Boyer, Proc. R. Soc. A **311,** 245 (1969).

[7]P. Hajicek, J. Math. Phys. **12,** 157 (1971).

Global analysis of the Kerr–Taub–NUT metric*

J. G. Miller†

Program in Applied Mathematics, Princeton University, Princeton, New Jersey 08540
(Received 18 July 1972)

The Kerr–Taub–NUT metric is a local analytic solution of the vacuum Einstein–Maxwell equations. When the metric is expressed in Schwarzschild-like coordinates, two types of coordinate singularity are present. One occurs at certain values of the "radial" coordinate where g_{rr} becomes infinite and corresponds to bifurcate Killing horizons; the other occurs at $\theta=0,\pi$, where the determinant of the components of the metric vanishes. It is shown that for nonzero NUT parameter the fixed points of the bifurcate Killing horizons and the degeneracies at $\theta=0,\pi$ cannot all be covered in one manifold. A maximal analytic manifold is constructed which covers the degeneracies at $\theta=0,\pi$. It is non-Hausdorff but contains maximal Hausdorff subspaces, topologically $S^3\times R$, which reduce to Taub–NUT space for vanishing Kerr parameter. Kerr–Taub space can be interpreted as a closed, inhomogeneous electromagnetic-gravitational wave undergoing gravitational collapse. Another maximal analytic manifold is constructed which covers the fixed points of the bifurcate Killing horizons and the degeneracy at $\theta=0$. It is suggested that this manifold represents the superposition of the Kerr geometry and a massless source of angular momentum at $\theta=\pi$ characterized by the NUT parameter.

I. INTRODUCTION

Many new exact solutions of the vacuum Einstein and Einstein–Maxwell equations have appeared in the literature in recent years. Most of the solutions (four-dimensional metrics with Lorentz signature) have been found by local methods and have coordinate singularities and incomplete geodesics. If the curvature invariants are finite at the coordinate singularities, it may be possible to cover the singularities in an extension. Maximal extensions are needed in order to study the global properties of a metric. The purpose of this paper is to study the global properties of the combined Kerr–Taub–NUT metric first given by Demiański and Newman[1] by an algebraic trick and later derived by Carter,[2] Kinnersley,[3] Kramer and Neugebauer,[4] Robinson *et al.*,[5] and Talbot.[6] Fortunately, the information needed for the global analysis can be obtained from the symmetries of the metric. We are also guided by previous work on the Kerr and Taub–NUT metrics.[7-15]

When a metric has symmetries, it is possible to impose various identifications on the metric. A metric with a given identification (or no identification at all) defines a manifold, and we can consider the extensions of that manifold. A connected Lorentz manifold $\bar{M}$ is an extension of a connected Lorentz manifold M, if a proper submanifold (codimension zero) of $\bar{M}$ is isometric to M. In this paper only analytic extensions will be considered. Non-Hausdorff extensions of Lorentz manifolds having no bifurcating geodesics may arise in a natural way, and we do not wish to exclude them. Non-Hausdorff Lorentz manifolds with bifurcating geodesics will not be considered because they are too pathological for any physical interpretation. A necessary and sufficient condition for a non-Hausdorff manifold to have no bifurcating geodesics is given by Hajicek.[13] We will say that a connected Lorentz manifold with no bifurcating geodesics is maximal, if every further extension has bifurcating geodesics.

Although the Kerr–Taub–NUT metric can be expressed in Schwarzschild-like coordinates, for nonzero NUT parameter the singularities at $\theta=0,\pi$ are not the usual degeneracies of spherical coordinates on the 2-sphere.[1] This problem was first encountered in the Taub–NUT metric.[10] There are several identifications to impose on the Taub–NUT metric that have physical interpretations. Misner[10] showed that the singularities at $\theta=0,\pi$ in the Taub–NUT metric are the degeneracies of spherical coordinates on the 3-sphere provided a certain identification is imposed on the metric. The t coordinate essentially becomes an Euler angle coordinate on S^3 and thus is periodic. With this identification, the fixed points of the bifurcate Killing horizons cannot be regular points of the manifold[14,15] and the maximal analytic extensions are non-Hausdorff.[13] The Taub region with this identification is interpreted as a cosmological model.[11] Seeking to avoid a timelike periodic coordinate in the NUT region, Bonnor[12] imposed only part of the identification considered by Misner. With Bonnor's identification, the singularity at $\theta=0$ and the fixed points of the bifurcate Killing horizons can be covered in an extension but not the singularity at $\theta=\pi$.[14] Bonner interpreted the $\theta=\pi$ singularity as a massless source of angular momentum.

In Sec. II we present the Kerr–Taub–NUT metric in Schwarzschild-like coordinates and discuss its local properties. In Sec. IIIA we consider Misner's identification and show that the singularities at $\theta=0,\pi$ can be covered in an extension. A maximal analytic extension of the Kerr–Taub–NUT metric with this identification is constructed in Sec. IIIB. A maximal analytic extension of the Kerr–Taub–NUT metric with Bonnor's identification is constructed in Sec. IIIC. The fixed points of the bifurcate Killing horizons and the singularity at $\theta=0$ are covered in the latter extension. Carter[2] has shown that the Hamilton–Jacobi equation for the geodesics in the Kerr–Taub–NUT metric separates in certain coordinate systems. We use this fact in Sec. IV in the discussion of the geodesics and show that the extensions constructed in Secs. IIIB and IIIC are maximal analytic manifolds. In Sec. V we summarize the results.

II. KERR-TAUB-NUT METRIC AND LOCAL PROPERTIES

The Kerr–Taub–NUT metric, represented here in Schwarzschild-like coordinates,[16]

$$ds^2=\Sigma(\Delta^{-1}dr^2+d\theta^2)+\Sigma^{-1}\sin^2\theta(adt-\rho d\phi)^2-\Sigma^{-1}\Delta(dt-Ad\phi)^2, \quad (1)$$

$$\Sigma\equiv r^2+(l+a\cos\theta)^2, \quad (2)$$

$$\Delta\equiv r^2-2mr-l^2+a^2+e^2, \quad (3)$$

$$A\equiv a\sin^2\theta-2l\cos\theta, \quad (4)$$

$$\rho \equiv r^2 + l^2 + a^2 = \Sigma + aA, \tag{5}$$

where m, l, a, and e are, respectively, the Schwarzschild, NUT, Kerr, and electromagnetic parameters, is a local analytic solution of the vacuum Einstein–Maxwell equations with electromagnetic field tensor

$$F = -2e\Sigma^{-2}[(\Sigma dr - rd\rho)\wedge(dt - Ad\phi) + rdA \wedge (adt - \rho d\phi)]. \tag{6}$$

When e is zero, the electromagnetic field tensor vanishes and the metric satisfies the vacuum Einstein equations. When a is zero, the metric reduces to the Taub–NUT metric,[17] and when l is zero it reduces to the Kerr metric.[18]

Two types of singularity may be present in a metric. One occurs where the components g_{ij} of the metric are singular; the other occurs where the determinant of the matrix (g_{ij}) vanishes. In metric (1), the first type occurs where Σ or Δ vanishes; the second occurs at $\theta = 0, \pi$. A curvature invariant becomes infinite where Σ vanishes,[1] but Σ is always positive for $a^2 < l^2$. We will consider only the case in which the zeros of $\Delta[r = r_\pm \equiv m \pm (m^2 + l^2 - a^2 - e^2)^{1/2}]$ are real and distinct. In this case, the singularities correspond to bifurcate Killing horizons.[19,20] Metric (1) is nonsingular (except for the possible curvature singularity) in three disjoint regions which are distinguished by the ranges of the "radial" coordinate. The regions $r_+ < r < \infty$ and $-\infty < r < r_-$ will each be called Kerr–NUT regions and labeled I and III, respectively, and the region $r_- < r < r_+$ will be called a Kerr–Taub region and labeled II. In each of these regions $0 < \theta < \pi, -\infty < \phi < \infty, -\infty < t < \infty$.

The Weyl tensor is of type D in the Petrov–Pirani classification,[21] and the two double principal null vector fields on region II, for example, are given by

$$N_\pm \equiv \rho\Delta^{-1}\partial_t \pm \partial_r + a\Delta^{-1}\partial_\phi, \tag{7}$$

where $\partial_t \equiv \partial/\partial t$, etc. The vector fields have been normalized so that $\nabla_{N_\pm} N_\pm = 0$. (All equations in this paper with plus-or-minus signs should be read as two equations, one with the upper sign throughout and one with the lower sign.) The integral curves of these vector fields give two shear-free null geodesic congruences,[22] which have nonzero rotation unless both a and l vanish. Clearly, ∂_t and ∂_ϕ are Killing vector fields, and when a is zero

$$\cos\phi\partial_\theta - \cot\theta \sin\phi\, \partial_\phi + 2l \csc\theta \sin\phi\, \partial_t,$$
$$\sin\phi\partial_\theta + \cot\theta \cos\phi\, \partial_\phi - 2l \csc\theta \cos\phi\, \partial_t, \tag{8}$$

are also Killing vector fields.[11]

III. GLOBAL PROPERTIES

A. Kerr-Taub and Kerr-NUT spaces

For nonzero NUT parameter, the singularities at $\theta = 0, \pi$ are not the usual degeneracies of spherical coordinates on the 2-sphere. Demiański and Newman[1] showed that the singularities are the degeneracies of spherical coordinates on the 3-sphere provided a certain identification is imposed on the metric, the same identification that Misner[10] imposed on the Taub–NUT metric. That identification on the Schwarzschild-like coordinates of metric (1) is

$$(\phi, t) = (\phi + (n + m)2\pi, t + (n - m)4l\pi) \quad \text{for all integers } n, m. \tag{9}$$

Let M_I, M_{II}, and M_{III} be the connected Lorentz manifolds obtained by imposing identification (9) on metric (1) in the regions I, II, and III, respectively. We now show how to obtain extensions of M_I, M_{II}, and M_{III} that are analytic on $S^3 \times R$. The set of quaternions $Q = \{q = w + ix + jy + kz \,|\, w^2 + x^2 + y^2 + z^2 \neq 0\}$ is a Lie group, diffeomorphic to $S^3 \times R$, and the left-invariant vector fields are analytic on $S^3 \times R$. The dual 1-forms of the left-invariant vector fields are also analytic, and metric (1) with identification (9) can be expressed in the global basis given by these 1-forms.

Left- and right-invariant vector fields in (w, x, y, z) coordinates can be computed from the definition of quaternion multiplication[23] and are given, respectively, by

$$\begin{aligned} X_0 &\equiv \tfrac{1}{2}(w\partial_w + x\partial_x + y\partial_y + z\partial_z),\\ X_1 &\equiv \tfrac{1}{2}(-x\partial_w + w\partial_x + z\partial_y - y\partial_z),\\ X_2 &\equiv \tfrac{1}{2}(-y\partial_w - z\partial_x + w\partial_y + x\partial_z),\\ X_3 &\equiv \tfrac{1}{2}(-z\partial_w + y\partial_x - x\partial_y + w\partial_z), \end{aligned} \tag{10}$$

and

$$\begin{aligned} Y_0 &\equiv \tfrac{1}{2}(w\partial_w + x\partial_x + y\partial_y + z\partial_z),\\ Y_1 &\equiv \tfrac{1}{2}(-x\partial_w + w\partial_x - z\partial_y + y\partial_z),\\ Y_2 &\equiv \tfrac{1}{2}(-y\partial_w + z\partial_x + w\partial_y - x\partial_z),\\ Y_3 &\equiv \tfrac{1}{2}(-z\partial_w - y\partial_x + x\partial_y + w\partial_z). \end{aligned} \tag{11}$$

They satisfy the commutation relations

$$\begin{aligned} [X_i, X_j] &= C_{ij}^k X_k,\\ [Y_i, Y_j] &= -C_{ij}^k Y_k,\\ [X_i, Y_j] &= 0, \end{aligned} \tag{12}$$

where $C_{ij}^k = \epsilon_{ijk}$ for $1 \le i, j, k \le 3$ and otherwise 0, ϵ_{ijk} being the totally antisymmetric symbol with $\epsilon_{123} = +1$. The ϵ_{ijk} are the structure constants of the rotation group $SO(3)$. $SO(3)$ is covered twice by its universal covering group $SU(2)$, and $SU(2)$ is isomorphic to the group of quaternions of unit norm.[10] The X_i and Y_i for $1 \le i \le 3$ are tangent to S^3. The dual 1-forms of the left-invariant vector fields are

$$\begin{aligned} \omega^0 &\equiv 2|q|^{-2}(wdw + xdx + ydy + zdz),\\ \omega^1 &\equiv 2|q|^{-2}(-xdw + wdx + zdy - ydz),\\ \omega^2 &\equiv 2|q|^{-2}(-ydw - zdx + wdy + xdz),\\ \omega^3 &\equiv 2|q|^{-2}(-zdw + ydx - xdy + wdz), \end{aligned} \tag{13}$$

where $|q|^2 \equiv w^2 + x^2 + y^2 + z^2$, and they satisfy

$$d\omega^k = -\tfrac{1}{2}C_{ij}^k \omega^i \wedge \omega^j. \tag{14}$$

The X_i and ω^i are invariant under left translations L_p: $q \to pq$ and the Y_i are invariant under right translations R_p: $q \to qp$.

Euler angle coordinates[24] (θ, ϕ, ψ) on S^3 may be defined by $q = w + ix + jy + kz = e^{r/2}e^{k\phi/2}e^{i\theta/2}e^{k\psi/2}$ or

$$\begin{aligned} w &= e^{r/2} \cos\theta/2 \cos(\phi + \psi)/2,\\ x &= e^{r/2} \sin\theta/2 \cos(\phi - \psi)/2,\\ y &= e^{r/2} \sin\theta/2 \sin(\phi - \psi)/2,\\ z &= e^{r/2} \cos\theta/2 \sin(\phi + \psi)/2, \end{aligned} \tag{15}$$

where $-\infty < r < \infty, 0 < \theta < \pi$, and

$$(\phi,\psi) = (\phi + (n+m)2\pi, \psi + (n-m)2\pi). \tag{16}$$

There are degeneracies at $\theta = 0, \pi$, where the Jacobian of the transformation vanishes. In (r,θ,ϕ,ψ) coordinates the left- and right-invariant vector fields are respectively,

$$\begin{aligned} X_0 &= \partial_r, \\ X_1 &= \cos\psi\,\partial_\theta + \csc\theta\,\sin\psi\,\partial_\phi - \cot\theta\,\sin\psi\,\partial_\psi, \\ X_2 &= -\sin\psi\,\partial_\theta + \csc\theta\,\cos\psi\,\partial_\phi - \cot\theta\,\cos\psi\,\partial_\psi, \\ X_3 &= \partial_\psi, \end{aligned} \tag{17}$$

and

$$\begin{aligned} Y_0 &= \partial_r, \\ Y_1 &= \cos\phi\,\partial_\theta - \cot\theta\,\sin\phi\,\partial_\phi + \csc\theta\,\sin\phi\,\partial_\psi, \\ Y_2 &= \sin\phi\,\partial_\theta + \cot\theta\,\cos\phi\,\partial_\phi - \csc\theta\,\cos\phi\,\partial_\psi, \\ Y_3 &= \partial_\phi; \end{aligned} \tag{18}$$

and the dual 1-forms of the left-invariant vector fields are

$$\begin{aligned} \omega^0 &= dr, \\ \omega^1 &= \cos\psi\,d\theta + \sin\theta\,\sin\psi\,d\phi, \\ \omega^2 &= -\sin\psi\,d\theta + \sin\theta\,\cos\psi\,d\phi, \\ \omega^3 &= d\psi + \cos\theta\,d\phi. \end{aligned} \tag{19}$$

The transformation $\psi = (2l)^{-1}t$ for $l \neq 0$ takes metric (1) into

$$\begin{aligned} ds^2 &= \Sigma\Delta^{-1}(\omega^0)^2 + \Sigma(\omega^1)^2 + \Sigma(\omega^2)^2 - 4l^2\Sigma^{-1}\Gamma(\omega^3)^2 \\ &\quad + a^2(2 - \Sigma^{-1}\Gamma)(\sin\theta\,\sin\psi\omega^1 + \sin\theta\cos\psi\omega^2)^2 \\ &\quad - 4la(1 - \Sigma^{-1}\Gamma)(\sin\theta\,\sin\psi\omega^1 + \sin\theta\,\cos\psi\omega^2)\omega^3, \end{aligned} \tag{20}$$

where

$$\Gamma \equiv \Delta - a^2\sin^2\theta, \tag{21}$$

and identification (9) into identification (16). For $a^2 < l^2$, this metric in (w,x,y,z) coordinates is analytic and has Lorentz signature on

$$\bar{M}_{\rm I} = \{(w,x,y,z) \in R^4 \mid e^{r_+} < w^2 + x^2 + y^2 + z^2 < \infty\},$$

and

$$\bar{M}_{\rm II} = \{(w,x,y,z) \in R^4 \mid e^{r_-} < w^2 + x^2 + y^2 + z^2 < e^{r_+}\},$$
$$\bar{M}_{\rm III} = \{(w,x,y,z) \in R^4 \mid 0 < w^2 + x^2 + y^2 + z^2 < e^{r_-}\},$$

Metric (20) defined on $\bar{M}_{\rm II}$ will be called Kerr–Taub space. It reduces to Taub space[17] when a vanishes, in which case Y_1 and Y_2 become Killing vector fields in addition to X_3 and Y_3. The Killing vector fields $X_3 - Y_3 = y\partial_x - x\partial_y$ and $X_3 + Y_3 = w\partial_z - z\partial_w$ vanish at $\theta = 0 (x = y = 0)$ and $\theta = \pi (w = z = 0)$, respectively, and are spacelike rotations.[20] They cannot be represented as a linear combination (with constant coefficients) of coordinate vector fields where they vanish. The degeneracies of Euler angle coordinates at $\theta = 0, \pi$ on S^3 are like polar coordinate singularities, where $\delta = (\psi - \phi)/2$ is an azimuthal angle coordinate near $\theta = 0$ and $\sigma = (\psi + \phi)/2$ is an azimuthal angle coordinate near $\theta = \pi$. In (r,θ,δ,σ) coordinates[25] $X_3 - Y_3 = \partial_\delta$ and $X_3 + Y_3 = \partial_\sigma$. It can be seen from transformation (15) that the submanifolds defined by $\theta = 0 (x = y = 0)$ and $\theta = \pi (w = z = 0)$ are topologically cylinders, $S^1 \times R$. These cylinders are totally geodesic.[20]

The "radial" coordinate is timelike in Kerr–Taub space, and the hypersurfaces of constant r are spacelike 3-spheres. The volume of these hypersurfaces goes to zero as r approaches $r_\pm$. Kerr–Taub space can be interpreted as a closed, inhomogeneous electromagnetic-gravitational wave undergoing gravitational collapse (see Brill[17] and Gowdy[26]). The timelike congruence given by the vector field ∂_r has nonzero rotation unless a vanishes, in which case the congruence is orthogonal to the homogeneous hypersurfaces of Taub space. Thus Kerr–Taub space has an intrinsic rotation which vanishes when the Kerr parameter is zero.

Metric (20) defined on $\bar{M}_{\rm I}$ or $\bar{M}_{\rm III}$ will be called a Kerr–NUT space. In the Kerr–NUT spaces the "radial" coordinate is spacelike, and the hypersurfaces of constant r are timelike 3-spheres and have closed timelike curves. We show the relationship of $\bar{M}_{\rm I}$ and $\bar{M}_{\rm III}$ to $\bar{M}_{\rm II}$ by giving a maximal analytic extension of Kerr–Taub space.

B. Kerr-Taub-NUT space

Two distinct Hausdorff analytic extensions of Kerr–Taub space can be obtained simply by straightening out the null vector fields

$$N_\pm = \rho(2l\Delta)^{-1}\partial_\psi \pm \partial_r + a\Delta^{-1}\partial_\phi. \tag{22}$$

The extensions are isometric to each other and representative of what we will call Kerr–Taub–NUT space, but they cannot be adopted simultaneously without abandoning the Hausdorff property. The transformation[27]

$$\begin{aligned} \psi_\pm &= \psi \mp \int \rho(2l\Delta)^{-1}dr, \\ \phi_\pm &= \phi \mp \int a\Delta^{-1}dr, \\ r &= r,\ \theta = \theta, \end{aligned} \tag{23}$$

for $r_- < r < r_+$ takes metric (20) into the metric

$$\begin{aligned} ds^2 &= \Sigma d\theta^2 + \Sigma^{-1}(\rho^2\sin^2\theta - \Delta A^2)d\phi_\pm^2 - 4l^2\Sigma^{-1}\Gamma d\psi_\pm^2 \\ &\quad - 4l\Sigma^{-1}(a\rho\sin^2\theta - \Delta A)d\phi_\pm d\psi_\pm \\ &\quad \mp 4l d\psi_\pm dr \pm 2A d\phi_\pm dr \\ &= \Sigma(\omega_\pm^1)^2 + \Sigma(\omega_\pm^2)^2 - 4l^2\Sigma^{-1}\Gamma(\omega_\pm^3)^2 \\ &\quad + a^2(2 - \Sigma^{-1}\Gamma)(\sin\theta\,\sin\psi_\pm\omega_\pm^1 + \sin\theta\,\cos\psi_\pm\omega_\pm^2)^2 \\ &\quad - 4la(1 - \Sigma^{-1}\Gamma)(\sin\theta\,\sin\psi_\pm\omega_\pm^1 + \sin\theta\,\cos\psi_\pm\omega_\pm^2)\omega_\pm^3 \\ &\quad \mp 4l\omega_\pm^3\omega_\pm^0 \pm 2a(\sin\theta\,\sin\psi_\pm\omega_\pm^1 + \sin\theta\,\cos\psi_\pm\omega_\pm^2)\omega_\pm^0, \end{aligned} \tag{24}$$

where $\omega_\pm^i$ are defined by placing the subscript $\pm$ on ω^i, ϕ,ψ in (19). Identification (16) becomes

$$(\phi_\pm,\psi_\pm) = (\phi_\pm + (n+m)2\pi, \psi_\pm + (n-m)2\pi). \tag{25}$$

We define vector fields X_{i_+}, Y_{i_+} in (r,θ,ϕ_+,ψ_+) coordinates and vector fields X_{i_-}, Y_{i_-} in (r,θ,ϕ_-,ψ_-) coordinates by placing the subscript $\pm$ on X_i, Y_i, ϕ, ψ in (17) and (18). Note that on Kerr–Taub space $X_{3_\pm} = X_3$, $Y_{3_\pm} = Y_3$, $\omega_\pm^0 = \omega^0$, and $N_\pm = \pm X_{0_\pm} = \pm Y_{0_\pm}$. In (r,θ,ϕ_+,ψ_+) coordinates on Kerr–Taub space

$$N_+ = \partial_r, \quad N_- = \rho(l\Delta)^{-1}\partial_{\psi_+} - \partial_r + 2a\Delta^{-1}\partial_{\phi_+}, \tag{26}$$

and in (r,θ,ϕ_-,ψ_-) coordinates

$$N_+ = \rho(l\Delta)^{-1}\partial_{\psi_-} + \partial_r + 2a\Delta^{-1}\partial_{\phi_-}, \quad N_- = -\partial_r. \tag{27}$$

Both metrics (24) extend analytically to $-\infty < r < \infty$.

The transformation $u = 2l\psi_+$ gives Kinnersley's[3] form of the Kerr–Taub–NUT metric. The degeneracies at $\theta = 0, \pi$ can be covered by transforming to $(w_\pm, x_\pm, y_\pm, z_\pm)$ coordinates defined by placing the subscript $\pm$ on w, x, y, z, ϕ, ψ in (15). The vector fields $X_{i_\pm}, Y_{i_\pm}, \omega^i_\pm$ in these coordinates are given by (10), (11), and (13) with the appropriate subscripts. The $\omega^i_\pm$ are the dual 1-forms of the left-invariant vector fields $X_{i_\pm}$ on $Q_\pm = \{q_\pm = w_\pm + ix_\pm + jy_\pm + kz_\pm \mid\mid q_\pm \mid \neq 0\}$.

In $(w_\pm, x_\pm, y_\pm, z_\pm)$ coordinates, metrics (24) for $a^2 < l^2$ are analytic and have Lorentz signature on $M_\pm = \{(w_\pm, x_\pm, y_\pm, z_\pm) \in R^4 \mid w_\pm^2 + x_\pm^2 + y_\pm^2 + z_\pm^2 \neq 0\}$. The hypersurfaces $r = r_\pm$ of $M_\pm$ are null and are Killing horizons[19,20] with respect to the Killing vector fields

$$K_\pm \equiv (2l)^{-1}\partial_{\psi_+} + a\rho_\pm^{-1}\partial_{\phi_+}, \tag{28}$$

where $\rho_\pm \equiv r_\pm^2 + l^2 + a^2$, i.e., $K_\pm$ are null at $r = r_\pm$, respectively. In $(r, \theta, \phi_-, \psi_-)$ coordinates

$$K_\pm = (2l)^{-1}\partial_{\psi_-} + a\rho_\pm^{-1}\partial_{\phi_-}, \tag{29}$$

and the null hypersurfaces $r = r_\pm$ of M_- are also Killing horizons with respect to $K_\pm$, respectively. Thus, M_+ and M_- are extensions of Kerr–Taub space that cover different branches of the Killing horizons. In M_+ the null geodesics which are the integral curves of N_+ have been extended across branches of the Killing horizons and are complete, but the null geodesics which are the integral curves of N_- are incomplete. In M_- the integral curves of N_- have been extended across different branches of the Killing horizons and are complete, but the integral curves of N_+ are incomplete. This type of behavior of null geodesics in Taub–NUT space has been discussed by Misner and Taub[11] and Geroch.[28]

Metric (24) with the upper sign defined on M_+ is isometric to metric (24) with the lower sign defined on M_-, and either one will be called Kerr–Taub–NUT space.

The isometry is given by $\psi_+ = -\psi_-, \phi_+ = -\phi_-$ or $(w_+, x_+, y_+, z_+) = (w_-, -x_-, -y_-, z_-)$. The regions $e^{r_+} < w_\pm^2 + x_\pm^2 + y_\pm^2 + z_\pm^2 < \infty$ and $0 < w_\pm^2 + x_\pm^2 + y_\pm^2 + z_\pm^2 < e^{r_-}$ of M_+ and M_- are isometric to the Kerr–NUT spaces $\bar{M}_{\rm I}$ and $\bar{M}_{\rm III}$, respectively. Thus Kerr–Taub–NUT space is also an analytic extension of either one of the Kerr–NUT spaces. When a is zero, Kerr–Taub–NUT space reduces to Taub–NUT space.[11]

It will be shown later that when both extensions $M_\pm$ of Kerr–Taub space are adopted simultaneously, the resulting extension is non-Hausdorff. To obtain a maximal (non-Hausdorff) analytic extension of Kerr–Taub space, an infinite number of copies of $M_\pm$ are patched together in the same way that Eddington–Finkelstein-like coordinate systems are patched together in the maximal analytic extension of the Kerr metric.[8,14] In this maximal analytic extension, M_+ and M_- overlap on a Kerr–Taub or a Kerr–NUT space, and the transformation on an overlap region is obtained from (23). A maximal (non-Hausdorff) extension of Taub space has already been given by Hajicek.[13]

C. Bifurcate Killing horizons

When a is zero, N_+ and N_- are surface-forming and can be "straightened out" simultaneously in one coordinate system.[14,15] A vector field is straightened out in a particular coordinate system if the images in the coordinate patch of the integral curves are straight lines. Such a coordinate system for the Schwarzschild metric which extends analytically to cover the bifurcate Killing horizon was given by Kruskal.[29] Although N_+ and N_- are not in general surface-forming, Boyer and Lindquist[8] have given a generalization of Kruskal's method for the Kerr metric which is also applicable to the Kerr–Taub–NUT metric. However, the fixed points of the bifurcate Killing horizons cannot be covered in an analytic extension when identification (9) is imposed. Therefore, we will consider an analytic extension of the Kerr–Taub–NUT metric with only part of identification (9). Rather than proceeding with the generalization of Kruskal coordinates, an alternative method is given which is based on the fact that the pairs of vector fields N_-, K_+ and N_-, K_- are surface-forming. (The pairs of vector fields N_+, K_+ and N_+, K_- are also surface-forming). The new method will be compared with the method of Boyer and Lindquist.

The transformation $t_\pm = 2l(\psi_\pm + \phi_\pm), \phi_\pm = \phi_\pm$ takes metric (24) into

$$ds^2 = \Sigma d\theta^2 + \Sigma^{-1}(\bar{\rho}^2 \sin^2\theta - \Delta\bar{A}^2)d\phi_\pm^2 - \Sigma^{-1}\Gamma dt_\pm^2 - 2(a\sin^2\theta - \Sigma^{-1}\Gamma\bar{A})dt_\pm d\phi_\pm \mp 2dt_\pm dr \pm 2\bar{A}d\phi_\pm dr, \tag{30}$$

where

$$\bar{A} \equiv a\sin^2\theta + 4l\sin^2\theta/2, \tag{31}$$

$$\bar{\rho} \equiv r^2 + (l+a)^2 = \Sigma + a\bar{A}, \tag{32}$$

and identification (25) becomes

$$(\phi_\pm, t_\pm) = (\phi_\pm + (n+m)2\pi, t_\pm + 8ln\pi). \tag{33}$$

Metric (30) with the upper sign is Demiański and Newman's[1] form of the Kerr–Taub–NUT metric. In (r, θ, ϕ_+, t_+) coordinates

$$\begin{aligned} N_+ &= \partial_r, \\ K_\pm &= \bar{\rho}_\pm\rho_\pm^{-1}\partial_{t_+} + a\rho_\pm^{-1}\partial_{\phi_+}, \\ X_3 - Y_3 &= -\partial_{\phi_+}, \\ X_3 + Y_3 &= 4l\partial_{t_+} + \partial_{\phi_+}, \end{aligned} \tag{34}$$

where $\bar{\rho}_\pm \equiv r_\pm^2 + (l+a)^2$, and in (r, θ, ϕ_-, t_-) coordinates

$$\begin{aligned} N_- &= -\partial_r, \\ K_\pm &= \bar{\rho}_\pm\rho_\pm^{-1}\partial_{t_-} + a\rho_\pm^{-1}\partial_{\phi_-}, \\ X_3 - Y_3 &= -\partial_{\phi_-}, \\ X_3 + Y_3 &= 4l\partial_{t_-} + \partial_{\phi_-}. \end{aligned} \tag{35}$$

Now consider metrics (30) with identification (33) for $n = 0$ and all intergers m. The degeneracy at $\theta = 0$ is like a polar coordinate singularity in these coordinates and since the $\phi_\pm$ coordinate is periodic with period 2π, the degeneracy can be covered. Consider the extensions of metrics (30) that cover the $\theta = 0$ degeneracy. These spaces can be described as the universal covering space of Kerr–Taub–NUT space with the $\theta = \pi$ cylinder deleted, and a maximal analytic extension of them will be given by forcing the Killing vector fields $K_\pm$ into the form of timelike rotations.[20] In these spaces the totally geodesic submanifold defined by $\theta = 0$ is topologically a plane and the $\theta = \pi$ singularity cannot be covered in an extension.[14] We need to consider only one of the metrics (30), say the one with the lower sign.

We will assume that $\bar{\rho}_\pm \neq 0$ and define

$$\bar{K}_\pm \equiv \bar{\rho}_\pm^{-1}\rho_\pm K_\pm = \partial_{t_-} + a\bar{\rho}_\pm^{-1}\partial_{\phi_-}. \tag{36}$$

The vector fields N_-, $\bar{K}_+$, and $X_3 - Y_3$ commute with each other and are linearly independent, and therefore all three can be represented as coordinate vector fields in one coordinate system. Similarly, N_-, $\bar{K}_-$, and $X_3 - Y_3$ can be represented as coordinate vector fields in one coordinate system. The transformation[8]

$$\bar{\varphi}_\pm = \phi_- - a\bar{\rho}_\pm^{-1}t_-, \quad t_- = t_- \tag{37}$$

takes metric (30) with the lower sign into the metric

$$\begin{aligned} ds^2 = {} & \Sigma d\theta^2 + \Sigma^{-1}(\bar{\rho}^2 \sin^2\theta - \Delta\bar{A}^2)d\bar{\varphi}_\pm^2 \\ & - \bar{\rho}_\pm^{-2}\Sigma^{-1}[\Sigma_\pm^2\Delta - a^2(\bar{\rho} - \bar{\rho}_\pm)^2 \sin^2\theta]dt_-^2 \\ & + 2\bar{\rho}_\pm^{-1}\Sigma^{-1}[\Sigma_\pm\Delta\bar{A} + a\bar{\rho}(\bar{\rho} - \bar{\rho}_\pm)\sin^2\theta]dt_-d\bar{\varphi}_\pm \\ & + 2\bar{\rho}_\pm^{-1}\Sigma_\pm dt_-dr - 2\bar{A}d\bar{\varphi}_\pm dr, \end{aligned} \tag{38}$$

where $\Sigma_\pm \equiv r_\pm^2 + (l + a\cos\theta)^2 = \bar{\rho}_\pm - a\bar{A}$. In $(r, \theta, \bar{\varphi}_+, t_-)$ coordinates

$$\begin{aligned} & N_- = -\partial_r, \quad \bar{K}_+ = \partial_{t_-}, \quad X_3 - Y_3 = -\partial_{\bar{\varphi}_+}, \\ & X_3 + Y_3 = 4l\partial_{t_-} + (1 - 4la\bar{\rho}_+^{-1})\partial_{\bar{\varphi}_+}, \end{aligned} \tag{39}$$

and in $(r, \theta, \bar{\varphi}_-, t_-)$ coordinates

$$\begin{aligned} & N_- = -\partial_r, \quad \bar{K}_- = \partial_{t_-}, \quad X_3 - Y_3 = -\partial_{\bar{\varphi}_-} \\ & X_3 + Y_3 = 4l\partial_{t_-} + (1 - 4la\bar{\rho}_-^{-1})\partial_{\bar{\varphi}_-}. \end{aligned} \tag{40}$$

Thus $\bar{\varphi}_\pm$ is periodic with period 2π.

We now seek coordinate systems $(\xi_\pm, \eta_\pm, \theta, \bar{\varphi}_\pm)$ such that the bifurcate Killing horizons are defined by $\xi_\pm\eta_\pm = 0$ and

$$\bar{K}_\pm = \bar{\kappa}_\pm(\eta_\pm\partial_{\eta_\pm} - \xi_\pm\partial_{\xi_\pm}), \tag{41}$$

where $\bar{\kappa}_\pm$ are constants. We want $\xi_\pm$ and $\eta_\pm$ to depend only on r and t_-. The vector field N_- is straightened out in $(r, \theta, \bar{\varphi}_\pm, t_-)$ coordinates, and r is an affine parameter of the integral curves of N_-. This property can be retained in $(\xi_\pm, \eta_\pm, \theta, \bar{\varphi}_\pm)$ coordinates by requiring N_- to be a product of the coordinate vector field $\partial_{\xi_\pm}$ by a function of $\eta_\pm$. The $\xi_\pm$ coordinate will then be an affine parameter of the integral curves of N_-. In $(\xi_\pm, \eta_\pm, \theta, \bar{\varphi}_\pm)$ coordinates, $N_- = -(\partial\xi_\pm/\partial r)\partial_{\xi_\pm} - (\partial\eta_\pm/\partial r)\partial_{\eta_\pm}$. Thus,

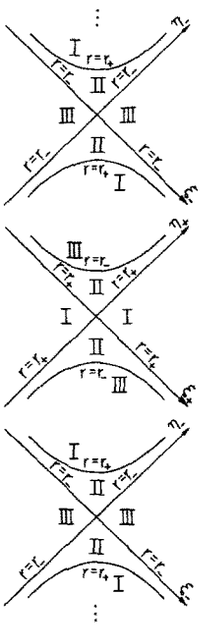


FIG. 1. Bifurcate Killing horizons. I and III label Kerr–NUT regions and II labels Kerr–Taub regions. Adjacent members of the infinite sequence of coordinate patches $(\xi_\pm, \eta_\pm, \theta, \bar{\varphi}_\pm)$ overlap on regions $\eta_+ > 0, \eta_- < 0$ or $\eta_+ < 0$, $\eta_- > 0$, and the transformations between these regions are given by Eqs. (47). The vector fields $\bar{N}_-$ and $\bar{K}_-$ are tangent to the submanifolds of constant θ and $\bar{\varphi}_-$, and the vector fields $\bar{N}_-$ and $\bar{K}_-$ are tangent to the submanifolds of constant θ and $\bar{\varphi}_-$.

we want $\eta_\pm$ to depend only on t_- and $\partial\xi_\pm/\partial r$ to depend only on $\eta_\pm$.

The transformations are restricted to the form

$$\xi_\pm = f_\pm r + g_\pm, \quad \eta_\pm = h_\pm, \tag{42}$$

where $f_\pm, g_\pm$, and $h_\pm$ are functions of t_-. In $(\xi_\pm, \eta_\pm, \theta, \bar{\varphi}_\pm)$ coordinates $\bar{K}_\pm = (\partial\xi_\pm/\partial t_-)\partial_{\xi_\pm} + (\partial\eta_\pm/\partial t_-)\partial_{\eta_\pm}$, and from (41) we see that $f_\pm' + \bar{\kappa}_\pm f_\pm = 0$, $g_\pm' + \bar{\kappa}_\pm g_\pm = 0$, and $h_\pm' - \bar{\kappa}_\pm h_\pm = 0$, where the prime denotes differentiation with respect to t_-. Thus $f_\pm = \alpha_\pm e^{-\bar{\kappa}_\pm t_-}$, $g_\pm = \beta_\pm e^{-\bar{\kappa}_\pm t_-}$, and $h_\pm = \gamma_\pm e^{\bar{\kappa}_\pm t_-}$, where $\alpha_\pm, \beta_\pm$, and $\gamma_\pm$ are constants. For the bifurcate Killing horizons to be defined by $\xi_\pm\eta_\pm = 0$, we must choose $\beta_\pm/\alpha_\pm = -r_\pm$. Without loss of generality, we arbitrarily set $\alpha_\pm = 1$ and $\gamma_\pm = \pm 1$ and finally arrive at the transformations

$$\xi_\pm = (r - r_\pm)e^{-\bar{\kappa}_\pm t_-}, \quad \eta_\pm = \pm e^{\bar{\kappa}_\pm t_-}. \tag{43}$$

The constants $\bar{\kappa}_\pm$ are still to be determined. The inverse transformations are

$$\begin{aligned} & r = \pm\xi_\pm\eta_\pm + r_\pm, \\ & t_- = \bar{\kappa}_\pm^{-1}\ln(\pm\eta_\pm), \end{aligned} \tag{44}$$

and metric (38) in $(\xi_\pm, \eta_\pm, \theta, \bar{\varphi}_\pm)$ coordinates becomes

$$\begin{aligned} ds^2 = {} & \Sigma d\theta^2 + \Sigma^{-1}(\bar{\rho}^2 \sin^2\theta - \Delta\bar{A}^2)d\bar{\varphi}_\pm^2 \\ & + \bar{\kappa}_\pm^{-2}\bar{\rho}_\pm^{-2}\Sigma^{-1}\xi_\pm^2(r - r_\pm)^{-2}[2\bar{\kappa}_\pm\bar{\rho}_\pm\Sigma_\pm\Sigma(r - r_\pm) \\ & - \Sigma_\pm^2\Delta + a^2(\bar{\rho} - \bar{\rho}_\pm)^2 \sin^2\theta]d\eta_\pm^2 \\ & \pm 2\bar{\kappa}_\pm^{-1}\bar{\rho}_\pm^{-1}\Sigma^{-1}\xi_\pm(r - r_\pm)^{-1}[\Sigma_\pm\Delta\bar{A} \\ & + a\bar{\rho}(\bar{\rho} - \bar{\rho}_\pm)\sin^2\theta]d\eta_\pm d\bar{\varphi}_\pm \\ & \pm 2\bar{\kappa}_\pm^{-1}\bar{\rho}_\pm^{-1}\Sigma_\pm d\xi_\pm d\eta_\pm \mp 2\bar{A}(\eta_\pm d\xi_\pm + \xi_\pm d\eta_\pm)d\bar{\varphi}_\pm. \end{aligned} \tag{45}$$

If we choose

$$\bar{\kappa}_\pm \equiv \tfrac{1}{2}\bar{\rho}_\pm^{-1}(r_\pm - r_\mp), \tag{46}$$

metrics (45) extend analytically to the entire $(\xi_\pm, \eta_\pm)$ plane. A maximal analytic (Hausdorff) extension is given by patching these metrics together as indicated in Fig. 1. [See also Fig. 1 of Ref. (7) or Fig. 1 of Ref. (14), where a two-dimensional submanifold is drawn rather than coordinate patches.] The transformation between overlap regions $\eta_+ > 0, \eta_- < 0$ or $\eta_+ < 0, \eta_- > 0$ is given by

$$\begin{aligned} & \xi_+\eta_+ + r_+ = -\xi_-\eta_- + r_-, \\ & \bar{\kappa}_+^{-1}\ln|\eta_+| = \bar{\kappa}_-^{-1}\ln|\eta_-|, \\ & \bar{\varphi}_\pm - \bar{\varphi}_\mp = a\bar{\kappa}_\pm^{-1}(\bar{\rho}_\mp^{-1} - \bar{\rho}_\pm^{-1})\ln|\eta_\pm|. \end{aligned} \tag{47}$$

In $(\xi_\pm, \eta_\pm, \theta, \bar{\varphi}_\pm)$ coordinates, $N_- = -|\eta_\pm|^{-1}\partial_{\xi_\pm}$ and is singular at $\eta_\pm = 0$, but the vector field $\bar{N}_- \equiv |\eta_\pm|N_- = -\partial_{\xi_\pm}$ is analytic and satisfies $\nabla_{\bar{N}}\bar{N}_- = 0$. The vector field N_+ is more complicated in $(\xi_\pm, \eta_\pm, \theta, \bar{\varphi}_\pm)$ coordinates. Although the integral curves of N_+ are not complete in any one of these coordinate patches, they are complete in the maximal analytic extension.

When l is zero, $\bar{\varphi}_\pm$ is also an azimuthal angle coordinate near $\theta = \pi$, and the singularities at $\theta = 0, \pi$ correspond to the usual degeneracies of spherical coordinates on the 2-sphere. Thus for $l = 0$, metrics (45) give a maximal analytic extension of the Kerr metric which is isometric to the one given by Boyer and Lindquist[8] when e is zero and by Carter[9] when e is nonzero. When a is zero,

metrics (45) give a maximal analytic extension of the Taub–NUT metric with the $\bar{\varphi}_\pm$ coordinate identified modulo 2π.[14] When l and a are both zero, these metrics give a maximal analytic extension of the Reissner–Nordström metric[30]; and when e is also zero, metric (45) with the upper sign becomes

$$ds^2 = 8m\xi_+^2(\xi_+\eta_+ + 2m)^{-1}d\eta_+^2 + 8m d\xi_+ d\eta_+ + (\xi_+\eta_+ + 2m)^2(d\theta^2 + \sin^2\theta d\bar{\varphi}_+^2), \quad (48)$$

and is isometric to Kruskal's[29] extension of the Schwarzschild metric. Similar coordinate systems for the Schwarzschild metric that cover the Kruskal diagram have previously been given by Israel[31] and Pajerski and Newman[32].

When a and e are zero and l is nonzero, Bonnor[12] has interpreted region I as the superposition of the Schwarzschild geometry and a massless source of angular momentum at $\theta = \pi$ characterized by the NUT parameter. We therefore suggest that metrics (45) with the $\bar{\varphi}_\pm$ coordinate identified modulo 2π give a maximal analytic extension of the superposition of the Kerr geometry and a massless source of angular momentum at $\theta = \pi$ characterized by the NUT parameter.

A generalization of Kruskal coordinates for the Kerr–Taub–NUT metric can be given by first applying the transformation[8]

$$\varphi_\pm = \phi - a\rho_\pm^{-1}t, \quad t = t, \quad (49)$$

to the Schwarzschild-like coordinates of Sec. II. For $(r, \theta, \varphi_+, t)$ coordinates on region II

$$N_\pm = \rho\Delta^{-1}\partial_t \pm \partial_r + a\Delta^{-1}(1 - \rho_+^{-1}\rho)\partial_{\varphi_+}, \quad K_+ = \partial_t, \quad (50)$$

and for $(r, \theta, \varphi_-, t)$ coordinates on region II

$$N_\pm = \rho\Delta^{-1}\partial_t \pm \partial_r + a\Delta^{-1}(1 - \rho_-^{-1}\rho)\partial_{\varphi_-}, \quad K_- = \partial_t. \quad (51)$$

We apply the methods for extending two-dimensional Lorentz manifolds[30,33] to the vector fields $F^{-1}\partial_t \pm \partial_r$, where $F \equiv \rho^{-1}\Delta$, and arrive at the transformation

$$u_\pm = \pm \exp(\kappa_\pm \int F^{-1}dr)\sinh\kappa_\pm t,$$
$$v_\pm = \pm \exp(\kappa_\pm \int F^{-1}dr)\cosh\kappa_\pm t \quad (52)$$
$$= \pm \left|\frac{r - r_\pm}{r_\pm}\right|^{1/2}\left|\frac{r - r_\mp}{r_\pm}\right|^{\kappa_\pm/2\kappa_\mp} e^{\kappa_\pm r}\cosh\kappa_\pm t,$$

where

$$F^{-1} = 1 + \tfrac{1}{2}\kappa_+^{-1}(r - r_+)^{-1} + \tfrac{1}{2}\kappa_-^{-1}(r - r_-)^{-1}, \quad (53)$$

$$\kappa_\pm \equiv \tfrac{1}{2}\rho_\pm^{-1}(r_\pm - r_\mp). \quad (54)$$

The transformation with the upper sign maps region II onto the quadrant $|u_+| < v_+$, and the transformation with the lower sign maps region II onto $|u_-| < -v_-$. In $(u_+, v_+, \theta, \varphi_+)$ coordinates

$$N_\pm = -\kappa_+^{-1}|v_+ \mp u_+|^{-1}f_+^{-2}(\partial_{u_+} \pm \partial_{v_+}) + a\Delta^{-1}(1 - \rho_+^{-1}\rho)\partial_{\varphi_+}, \quad (55)$$
$$K_+ = \kappa_+(v_+\partial_{u_+} + u_+\partial_{v_+}),$$

and in $(u_-, v_-, \theta, \varphi_-)$ coordinates

$$N_\pm = \kappa_-^{-1}|v_- \mp u_-|^{-1}f_-^{-2}(\partial_{u_-} \pm \partial_{v_-}) + a\Delta^{-1}(1 - \rho_-^{-1}\rho)\partial_{\varphi_-}, \quad (56)$$
$$K_- = \kappa_-(v_-\partial_{u_-} + u_-\partial_{v_-}),$$

where

$$f_\pm^2 \equiv \kappa_\pm^{-2}\,|F\exp(-2\kappa_\pm \int F^{-1}dr)|$$
$$= r_\pm^2\kappa_\pm^{-2}\rho^{-1}\left|\frac{r - r_\mp}{r_\pm}\right|^{1-\kappa_\pm/\kappa_\mp} e^{-2\kappa_\pm r}. \quad (57)$$

It is not necessary to compute metric (1) in Kruskal-like coordinates $(u_\pm, v_\pm, \theta, \varphi_\pm)$ and verify that it extends analytically to the entire $(u_\pm, v_\pm)$ plane and, in particular, covers the bifurcate Killing horizons defined by $u_\pm^2 - v_\pm^2 = 0$, since that has already been accomplished by (45). The transformation (49) was needed to ensure that $K_\pm$ would go into the form of a timelike rotation when transformation (52) was applied. The Kruskal-like coordinates treat the vector fields N_+ and N_- in a more symmetric fashion than the enlarged Eddington–Finkelstein-like patches $(\xi_\pm, \eta_\pm, \theta, \bar{\varphi}_\pm)$, but the latter has the advantage that the metric in these coordinates is given explicitly.

Although we only considered identifying $\bar{\varphi}_\pm$ modulo 2π in metric (45), we can impose the further identification (33). In $(\xi_\pm, \eta_\pm, \theta, \bar{\varphi}_\pm)$ coordinates

$$X_3 - Y_3 = -\partial_{\bar{\varphi}_\pm},$$
$$X_3 + Y_3 = 4l\bar{\kappa}_\pm(\eta_\pm\partial_{\eta_\pm} - \xi_\pm\partial_{\xi_\pm}) + (1 - 4la\bar{\rho}_\pm^{-1})\partial_{\bar{\varphi}_\pm}, \quad (58)$$

and identification (33) becomes

$$(\xi_\pm, \eta_\pm, \theta, \bar{\varphi}_\pm) = (\xi_\pm e^{-8l\bar{\kappa}_\pm n\pi},\ \eta_\pm e^{8l\bar{\kappa}_\pm n\pi}, \theta, \bar{\varphi}_\pm + (n + m)2\pi - 8la\bar{\rho}_\pm^{-1}n\pi). \quad (59)$$

Under this identification, metric (45) with the zeros of $\bar{K}_\pm$ deleted is compatible with the atlas of the non-Hausdorff extension of Kerr–Taub space of Sec. IIIB. Any two points on different branches of a Killing horizon and in the same subspace of constant θ violate the Hausdorff property. It can be seen from (59) that if $la\bar{\rho}_\pm^{-1}$ is rational, the orbits of $\bar{K}_\pm$ are closed (diffeomorphic to S^1) and the submanifolds of constant θ and $\bar{\varphi}_\pm$ (Fig. 1) are non-Hausdorff. However, if $la\bar{\rho}_\pm^{-1}$ is irrational, the orbits of $\bar{K}_\pm$ are open (diffeomorphic to R) and the submanifolds of constant θ and $\bar{\varphi}_\pm$ are Hausdorff, but their topology is not the relative topology.[34] If the zeros of $\bar{K}_\pm$ are not deleted, the space given by identification (59) is not even a manifold, because it is not a T_1 topological space.[14,15] Thus the zeros of $\bar{K}_\pm$ cannot be covered in an analytic extension of Kerr–Taub space.

The non-Hausdorff property occurs only on the horizons and causes no geodesics to bifurcate.[13] The causal violations (closed timelike curves) in Kerr–NUT space are more disturbing than the non-Hausdorff property. However, the Killing horizons are also Cauchy horizons with respect to any spacelike hypersurface of Kerr–Taub space, and the only way of establishing a relationship between Kerr–Taub space and a Kerr–NUT space is by an analytic extension.[11] The stability of Kerr–Taub space should also be considered.[35]

IV. GEODESICS AND ORBITS

We will follow the notation of Carter in Ref. 9 fairly closely in the discussion of geodesics and the orbits of charged particles. We will show that all geodesics in the non-Hausdorff extension of Sec. IIIB are complete except those approaching the zeros of $K_\pm$ and that all geodesics in the Hausdorff extension of Sec. IIIC are complete except those approaching the zeros of X_3 +

Y_3. This implies that the extensions are maximal analytic Lorentz manifolds.

The equations of motion of a test particle of mass μ and charge ϵ are given by

$$\mu(Du^i/d\tau) = \epsilon F^{ij}u_j, \tag{60}$$

where $u^i = dx^i/d\tau, D/d\tau$ denotes covariant differentiation along the curve $x^i(\tau)$ with respect to proper time τ, and F is the electromagnetic field tensor. If $\tilde{A}$ is a covariant vector potential satisfying

$$F = 2d\tilde{A}, \tag{61}$$

the equations of motion can be derived from the Lagrangian

$$L = \tfrac{1}{2}g_{ij}\dot{x}^i\dot{x}^j + \epsilon\tilde{A}_i\dot{x}^i, \tag{62}$$

where the dot denotes differentiation with respect to an affine parameter λ related to proper time by

$$\tau = \mu\lambda. \tag{63}$$

The condition (63) is equivalent to the normalization

$$g_{ij}\dot{x}^i\dot{x}^j = -\mu^2. \tag{64}$$

For $\epsilon = 0$ and $\mu^2 = \pm 1, 0$, we obtain the spacelike, timelike, and null geodesics.

The Hamiltonian

$$H = \tfrac{1}{2}g^{ij}(p_i - \epsilon\tilde{A}_i)(p_j - \epsilon\tilde{A}_j), \tag{65}$$

where the momenta are defined by

$$p_i \equiv \frac{\partial L}{\partial\dot{x}^i} = g_{ij}\dot{x}^j + \epsilon\tilde{A}_i, \tag{66}$$

is itself a constant of the motion, since it does not explicitly depend on λ.

From (64) we see that

$$H = -\tfrac{1}{2}\mu^2. \tag{67}$$

We will use metric (24) in Euler angle coordinates and the vector potentials

$$A_\pm \equiv -er\Sigma^{-1}(2ld\psi_\pm - Ad\phi_\pm), \tag{68}$$

which satisfy (61). Let G be the nonsingular linear map induced by the metric tensor field that takes vector fields into 1-forms. If X and Y are arbitrary vector fields, then GX is the 1-form defined by

$$GX(Y) = g(X, Y), \tag{69}$$

where g is the metric tensor field. In terms of the 1-forms $\omega_\pm \equiv GN_\pm$, we can write (68) as

$$A_\pm = er\Sigma^{-1}\omega_\pm. \tag{70}$$

From (24) the momenta are

$$\begin{aligned} p_r &= \pm A\dot{\phi}_\pm \mp 2l\dot{\psi}_\pm, \quad p_\theta = \Sigma\dot{\theta}, \\ p_{\phi_\pm} &= \Sigma^{-1}(\rho^2\sin^2\theta - \Delta A^2)\dot{\phi}_\pm - 2l\Sigma^{-1}(a\rho\sin^2\theta - \Delta A) \\ &\quad \times\dot{\psi}_\pm \pm A\dot{r} + \epsilon er\Sigma^{-1}A, \\ p_{\psi_\pm} &= -2l\Sigma^{-1}(a\rho\sin^2\theta - \Delta A)\dot{\phi}_\pm - 4l^2\Sigma^{-1}\Gamma\dot{\psi}_\pm \\ &\quad \mp 2lr^{\cdot} - 2l\epsilon er\Sigma^{-1}. \end{aligned} \tag{71}$$

The contravariant form of the metric (24) is

$$\begin{aligned} (\partial_s)^2 &= \Sigma^{-1}\Delta(\partial_r)^2 + \Sigma^{-1}(\partial_\theta)^2 \mp 2a\Sigma^{-1}(\partial_r)(\partial_{\phi_\pm}) \\ &\quad \mp (l\Sigma)^{-1}\rho(\partial_r)(\partial_{\psi_\pm}) + \Sigma^{-1}\csc^2\theta(\partial_{\phi_\pm})^2 \\ &\quad + (l\Sigma)^{-1}A\csc^2\theta(\partial_{\phi_\pm})(\partial_{\psi_\pm}) + (4l^2\Sigma)^{-1}A^2\csc^2\theta(\partial_{\psi_\pm})^2, \end{aligned} \tag{72}$$

and the Hamiltonian becomes

$$\begin{aligned} H = \tfrac{1}{2}\Sigma^{-1}\{ &\Delta p_r^2 + p_\theta^2 \mp 2[ap_{\phi_\pm} + (2l)^{-1}\rho p_{\psi_\pm} + \epsilon er]p_r \\ &+ \csc^2\theta[p_{\phi_\pm} + (2l)^{-1}Ap_{\psi_\pm}]^2\}. \end{aligned} \tag{73}$$

From the Killing vector fields X_3 and Y_3 we immediately obtain two constants of motion

$$p_{\phi_\pm} = \Phi, \quad p_{\psi_\pm} = \Psi, \tag{74}$$

the constants Φ and Ψ being independent of the plus-or-minus sign.

From (65) the Hamilton–Jacobi equation is

$$\frac{\partial S}{\partial\lambda} + \tfrac{1}{2}g^{ij}\left[\left(\frac{\partial S}{\partial x^i}\right) - \epsilon\tilde{A}_i\right]\left[\left(\frac{\partial S}{\partial x^j}\right) - \epsilon\tilde{A}_j\right] = 0, \tag{75}$$

where S is the Jacobi action. The action separates[2] in the form

$$S = \tfrac{1}{2}\mu^2\lambda + \Phi\phi_\pm + \Psi\psi_\pm + S_\theta + S_r, \tag{76}$$

where S_θ is a function of θ and S_r of r. Substituting (76) into (75), we find that

$$\begin{aligned} &\left(\frac{dS_\theta}{d\theta}\right)^2 + \csc^2\theta[\Phi + (2l)^{-1}\Psi A]^2 + \mu^2(l + a\cos\theta)^2 \\ &\quad = -\Delta(dS_r/dr)^2 \pm 2[a\Phi + (2l)^{-1}\Psi\rho + \epsilon er]\left(\frac{dS_r}{dr}\right) - \mu^2r^2. \end{aligned} \tag{77}$$

Both sides of (77) must be equal to a fourth constant of motion, which we denote by $\mathcal{K}$. It is positive when μ is real, i.e., for charged particle orbits and timelike and null geodesics. Since $p_\theta = \partial S/\partial\theta$ and $p_r = \partial S/\partial r$, $\mathcal{K}$ can be expressed in terms of the momenta:

$$\begin{aligned} &p_\theta^2 + \csc^2\theta[\Phi + (2l)^{-1}\Psi A]^2 + \mu^2(l + a\cos\theta)^2 = \mathcal{K}, \\ &\quad \Delta p_r^2 \mp 2[a\Phi + (2l)^{-1}\Psi\rho + \epsilon er]p_r + \mu^2r^2 = -\mathcal{K}. \end{aligned} \tag{78}$$

These equations along with (74) give a complete set of first integrals. Substituting (71) into (78), we can read off the covariant components K_{ij} of a Killing tensor field[36] satisfying

$$K_{ij}\dot{x}^i\dot{x}^j = \mathcal{K}. \tag{79}$$

The contravariant form of the Killing tensor field is

$$K \equiv \Delta N_+N_- + r^2(\partial_s)^2, \tag{80}$$

where $(\partial_s)^2$ is given by (72) and N_+N_- is the symmetric product of the vector fields N_+ and N_- defined by

$$N_+N_- \equiv \tfrac{1}{2}(N_+ \otimes N_- + N_- \otimes N_+). \tag{81}$$

This Killing tensor field has been obtained in an invariant way by Hughston *et al.*[37] From (77) we have

$$\frac{dS_\theta}{d\theta} = \sqrt{\Theta}, \quad \frac{dS_r}{dr} = \Delta^{-1}(\pm P + \sqrt{R}), \tag{82}$$

where the functions $\Theta(\theta), P(r), R(r)$ are defined by

$$\Theta \equiv \mathcal{K} - \csc^2\theta[\Phi + (2l)^{-1}\Psi A]^2 - \mu^2(l + a\cos\theta)^2,$$
$$P \equiv a\Phi + (2l)^{-1}\Psi\rho + \epsilon e r, \quad R \equiv P^2 - \Delta(\mu^2 r^2 + \mathcal{K}). \tag{83}$$

The choice of signs for $\sqrt{\Theta}$ and $\sqrt{R}$ is independent of each other and of the plus-or-minus sign.

Thus the Jacobi action is

$$S = \tfrac{1}{2}\mu^2\lambda + \Phi\phi_\pm + \Psi\psi_\pm + \int^\theta \sqrt{\Theta}d\theta + \int^r \Delta^{-1}(\pm P \pm \sqrt{R})dr, \tag{84}$$

and by differentiating with respect to $\mathcal{K}, \mu^2, \Phi, \Psi$, the equations of motion are solved by the quadratures

$$\int^\theta \frac{d\theta}{\sqrt{\Theta}} = \int^r \frac{dr}{\sqrt{R}}, \tag{85}$$

$$\lambda = \int^\theta \frac{(l + a\cos\theta)^2 d\theta}{\sqrt{\Theta}} + \int^r \frac{r^2 dr}{\sqrt{R}}, \tag{86}$$

$$\phi_\pm = \int^\theta \frac{\csc^2\theta[\Phi + (2l)^{-1}\Psi A]d\theta}{\sqrt{\Theta}} \mp \int^r \frac{a}{\Delta}\left(1 \pm \frac{P}{\sqrt{R}}\right)dr, \tag{87}$$

$$2l\psi_\pm = \int^\theta \frac{A\csc^2\theta[\Phi + (2l)^{-1}\Psi A]d\theta}{\sqrt{\Theta}} \mp \int^r \frac{\rho}{\Delta}\left(1 \pm \frac{P}{\sqrt{R}}\right)dr. \tag{88}$$

We immediately see from (85) and (86) that for $a^2 < l^2$ Kerr-Taub-NUT space is a maximal Hausdorff manifold since it has the property of distant boundaries,[11] i.e., every geodesic that is not contained in some compact set has infinite affine length.

From (74) and (78) or from (85) through (88) we obtain the first order system of differential equations:

$$\Sigma\dot{\theta} = \sqrt{\Theta}, \tag{89}$$

$$\Sigma\dot{r} = \sqrt{R}, \tag{90}$$

$$\Sigma\dot{\phi}_\pm = \csc^2\theta[\Phi + (2l)^{-1}\Psi A] \mp a\Delta^{-1}[(\sqrt{R}) \pm P], \tag{91}$$

$$2l\Sigma\dot{\psi}_\pm = A\csc^2\theta[\Phi + (2l)^{-1}\Psi A] \mp \rho\Delta^{-1}[(\sqrt{R}) \pm P]. \tag{92}$$

A geodesic will be incomplete in the coordinate patch $(r, \theta, \phi_+, \psi_+)$ only if it reaches $\theta = 0, \pi$ or approaches a branch of a Killing horizon in M_-. The latter occurs where Δ vanishes. If P is nonzero where Δ vanishes, then $|P(r_\pm)/\sqrt{R(r_\pm)}| = 1$. The second integral with the upper sign in (87) and (88) diverges at $r = r_\pm$ when $P(r_\pm)/\sqrt{R(r_\pm)} = 1$. In this case, the integrals with the lower sign do not diverge, and the incomplete geodesics can be continued in the $(r, \theta, \phi_-, \psi_-)$ coordinate patch. Similarly, when $P(r_\pm)/\sqrt{R(r_\pm)} = -1$, the incomplete geodesics in M_- approaching a branch of an horizon in M_+ can be continued in the $(r, \theta, \phi_+, \psi_+)$ coordinate patch.

Finally, there is the case in which P vanishes at a zero of Δ. From (83) we see that R has a zero where P and Δ both vanish. We need to consider only the case in which R has a simple zero, since the integral for λ diverges when R has a double zero. In this case $P/\sqrt{R}$ goes to zero as a zero of Δ is approached, and the integrals for $\phi_\pm$ and $\psi_\pm$ diverge. This means that the geodesic is incomplete in both coordinate patches $(r, \theta, \phi_\pm, \psi_\pm)$ and is approaching a point of the 2-surface where $K_\pm$ is zero. The geodesic can be continued through this surface in the extension of Sec. IIIC provided it is not also approaching a zero of $X_3 + Y_3$.

We conclude that for $a^2 < l^2$, all geodesics in the non-Hausdorff extension of Sec. IIIB are complete except those approaching the zeros of $K_\pm$, and all geodesics in the Hausdorff extension of Sec. IIIC are complete except those approaching the zeros of $X_3 + Y_3$. Thus both extensions are maximal analytic Lorentz manifolds. In case $a^2 \geq l^2$, it can be seen from (89) and (90) that both extensions have the additional incompleteness at the curvature singularity where Σ vanishes.

V. CONCLUSIONS

The Kerr-Taub-NUT metric has been represented in various coordinate systems for different purposes. It is easy to recognize the special cases of the metric when various parameters are zero in Schwarzschild-like coordinates. The Cartesian coordinates $(w_\pm, x_\pm, y_\pm, z_\pm)$ on $R^4 - \{0\}$ give a one-map atlas of Kerr-Taub-NUT space. However, computations in these coordinates are not very practical, and Euler angle coordinates are more desirable for this purpose. The coordinate systems $(\xi_\pm, \eta_\pm, \theta, \overline{\varphi}_\pm)$ were constructed to cover the bifurcate Killing horizons and also to show how the non-Hausdorff property arises when M_+ and M_- are patched together.

The Kerr-Taub-NUT metric is not particularly simple in any coordinate system. Fortunately, it was not necessary to work with the metric directly, since it was possible to carry out the global analysis as an analysis of the vector fields $N_\pm, K_\pm, X_3 - Y_3$ and $X_3 + Y_3$. The various coordinate transformations were derived by requiring some of these vector fields to have a specific form. The global problems that were encountered in the analytic extensions can be explained in terms of the Killing vector fields $K_\pm, X_3 - Y_3$, and $X_3 + Y_3$.

The types of zeros of a Killing vector field X (or fixed points of the local one-parameter group of isometries $\{U_\lambda\}$ generated by X) in a 4-dimensional Lorentz manifold have been invariantly characterized by Boyer.[20] Let $X(p) = 0$ for some point p and assume that X is not identically zero on a neighborhood of p. We define $\hat{\xi}_{ij} \equiv (d\omega)_{ji} = (\nabla_j\omega)_i$, where $\omega \equiv GX$. Then the invariants

$$\hat{\xi}_{ij}{}^*\hat{\xi}^{ij}, \tag{93}$$

$$\hat{\xi}_{ij}\hat{\xi}^{ij}, \tag{94}$$

evaluated at p are the invariants of the infinitesimal generator $\hat{\xi}^i{}_j(p)$ of the Lorentz transformation $U_{\lambda_*}(p)$, where U_{λ_*} is the differential of the map $U_\lambda : m \to U_\lambda(m)$. (${}^*\hat{\xi}^{ij}$ is the dual of $\hat{\xi}^{ij}$.) The invariant (93) evaluated at a fixed point is zero for spacelike and timelike rotations, and invariant (94) evaluated at a fixed point is positive for spacelike rotations and negative for timelike rotations. The fixed points of spacelike and timelike rotations are not isolated but form a totally geodesic 2-dimensional submanifold called the fixed point surface.

The invariant (94) of $X_3 - Y_3$ or $X_3 + Y_3$ evaluated at a fixed point is 2. In general, if X is a spacelike rotation near p and X is represented as a coordinate vector field ∂_ϕ, then ϕ must be periodic with period $2\pi/[\frac{1}{2}\hat{\xi}_{ij}(p)\hat{\xi}^{ij}(p)]^{1/2}$ in order to avoid a "conical" singularity at p. Misner[10] and Demiański and Newman[1] discovered identification (33) by seeking the identifications that permit the zeros of $X_3 - Y_3$ and $X_3 + Y_3$ to be covered. The invariant (94) of $K_\pm$ evaluated at a fixed point is $-2\kappa_\pm^2$.

If the zeros of $X_3 - Y_3$ and $X_3 + Y_3$ are covered, the

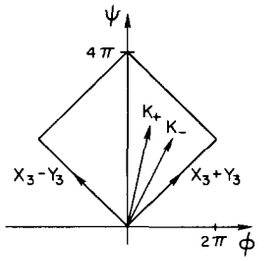


FIG. 2. Surfaces of transitivity for nonzero NUT parameter. For $\theta \neq 0, \pi$, the Killing vector fields $X_3 - Y_3$, $X_3 + Y_3$, and $K_\pm$ are tangent to the submanifolds of constant θ and r, and their directions at $(\phi, \psi) = (0, 0)$ for a typical choice of constants m, l, a, and e are indicated by the arrows. When identification (16) is imposed, the 2-torus is obtained by identifying opposite sides of the square. When a is zero, $K_+ = K_- = (2l)^{-1}\partial_\psi$.

orbits of $X_3 - Y_3$ and $X_3 + Y_3$ are closed and the surfaces of transitivity for $\theta \neq 0, \pi$ are 2-tori (see Fig. 2). Thus the orbits of $K_\pm$ are either closed or dense in a 2-torus. In either case the zeros of $K_\pm$ cannot be covered in an extension. In the universal covering space of Kerr–Taub–NUT space with the zeros of $X_3 + Y_3$ deleted, the orbits of $X_3 + Y_3$ are "unwrapped" and the zeros of $K_\pm$ may be covered in an extension provided $\bar{\rho}_\pm \neq 0$.

ACKNOWLEDGMENTS

I am deeply grateful to M. D. Kruskal and J. A. Wheeler for stimulating discussions during the course of this work.

*Work supported in part by the National Science Foundation under Grant No. GP 30799 X.

†Present address: Department of Mathematics, University of California, Los Angeles, California 90024.

[1]M. Demiański and E. T. Newman, Bull. Acad. Pol. Sci. Ser. Sci. Math. Astron. Phys. **14,** 653 (1966).
[2]B. Carter, Commun. Math. Phys. **10,** 280 (1968).
[3]W. Kinnersley, J. Math. Phys. **10,** 1195 (1969).
[4]D. Kramer and G. Neugebauer, Commun. Math. Phys. **10,** 132 (1968).
[5]I. Robinson, J. Robinson, and J. Zund, J. Math. Mech. **18,** 881 (1969).
[6]C. Talbot, Commun. Math. Phys. **13,** 45 (1969).
[7]B. Carter, Phys. Rev. **141,** 1242 (1966).
[8]R. H. Boyer and R. Lindquist, J. Math. Phys. **8,** 265 (1967).
[9]B. Carter, Phys. Rev. **174,** 1559 (1968).
[10]C. Misner, J. Math. Phys. **4,** 924 (1963).
[11]C. Misner and A. Taub, Zh. Eksp. Teor. Fiz. **55,** 233 (1968) [Sov. Phys.-JETP **28,** 122 (1969)].
[12]W. Bonnor, Proc. Camb. Philos. Soc. **66,** 145 (1969). See also A. Sackfield, Proc. Camb. Philos. Soc. **70,** 89 (1971).
[13]P. Hajicek, J. Math. Phys. **12,** 157 (1971); Commun. Math. Phys. **17,** 109 (1970).
[14]J. G. Miller, M. D. Kruskal, and B. Godfrey, Phys. Rev. D **4,** 2945 (1971).
[15]G. F. R. Ellis and S. W. Hawking, *Space-Time in the Large* (to be published).
[16]This metric can be obtained from metric [*A*] in Ref. 2 by applying the transformation $\lambda = r, \mu = l + a\cos\theta, \psi = \phi/a, \chi = t - (l^2 + a^2)\phi/a$, where $\Lambda \equiv 0, h \equiv 1, p \equiv a^2 - l^2, q \equiv l$.
[17]A. Taub, Ann. Math. **53,** 472 (1951); E. T. Newman, L. Tamburino, and T. Unti, J. Math. Phys. **4,** 915 (1963); D. Brill, Phys. Rev. B **133,** 845 (1964).
[18]R. P. Kerr, Phys. Rev. Lett. **11,** 237 (1963); E. T. Newman, E. Couch, K. Chinnapared, A. Exton, A. Prakash, and R. Torrence, J. Math. Phys. **6,** 918 (1965).
[19]B. Carter, J. Math. Phys. **10,** 70 (1969).
[20]R. H. Boyer, Proc. R. Soc. A **311,** 245 (1969).
[21]A. Z. Petrov, Sci. Notices, Kazan State Univ. **114,** 55 (1954); F. A. E. Pirani, Phys. Rev. **105,** 1089 (1957).
[22]W. Kundt and M. Trumper, Akad. Wiss. Lit. Mainz **12,** (1962); J. N. Goldberg and R. K. Sachs, Acta Phys. Pol. Suppl. **22,** 13 (1962).
[23]P. M. Cohn, *Lie Groups* (Cambridge, U. P., Cambridge, 1957), p. 65.
[24]C. Misner, in *Lectures in Applied Mathematics*, edited by J. Ehlers (American Mathematical Society, Providence, R. I., 1967), Vol. 8, p. 160.
[25]R. Gowdy, Phys. Rev. Lett. **27,** 826 (1971).
[26]R. Gowdy, (to be published).
[27]To be more precise, we should write $r_\pm = r$ and $\theta_\pm = \theta$, but this would conflict with the well-established notation in which the zeros of Δ are denoted by $r_\pm$. We find it convenient to use the plus-or-minus sign for two purposes, to label the horizons ($r = r_\pm$), and to label the principal null vector fields and coordinate systems adapted to these vector fields.
[28]R. Geroch, J. Math. Phys. **9,** 450 (1968).
[29]M. D. Kruskal, Phys. Rev. **119,** 1743 (1960).
[30]J. Graves and D. Brill, Phys. Rev. **120,** 1507 (1960).
[31]W. Israel, Phys. Rev. **143,** 1016 (1966).
[32]D. Pajerski and E. T. Newman, J. Math. Phys. **12,** 1929 (1971).
[33]M. Walker, J. Math. Phys. **11,** 2280 (1970); B. Godfrey, J. Math. Phys. **12,** 606 (1971).
[34]N. Hicks, *Notes on Differential Geometry* (Van Nostrand, Princeton, N. J., 1965), p. 13.
[35]S. Bonanos, Commun. Math. Phys. **22,** 190 (1971).
[36]M. Walker and R. Penrose, Commun. Math. Phys. **18,** 265 (1970).
[37]L. P. Hughston, R. Penrose, P. Sommers, and M. Walker, Commun. Math. Phys. **27,** 303 (1972).

The differentiable and causal structures of space-time*

N. M. J. Woodhouse

Department of Mathematics, King's College, London, England
Department of Physics, Joseph Henry Laboratories, Princeton University, Princeton, New Jersey
(Received 22 August 1972)

An improved version is presented of Ehlers, Schild, and Pirani's axiomatization of the differentiable and causal structures of space-time. It is shown that, under certain physically reasonable axioms, the properties of freely falling particles and light signals define a causal and differentiable structure on space-time. Further, it is shown how the causal structure define a causal and differentiable structure light signals propagate along conformal null geodesics.

INTRODUCTION

The normal development of general relativity begins with two hypotheses. The first is that space–time can be represented as a differentiable manifold and the second is the chronometric hypothesis: that there is a metric tensor (g_{ab} in local coordinates) defined on the space–time manifold and that the time interval between two events x^a and $x^a + dx^a$ on the world line of a standard clock is given by $(g_{ab}dx^a dx^b)^{1/2}$. From these two hypotheses all the geometric structures of space–time, its causal, conformal, and projective structures can be very easily derived. The two geodesic hypotheses are then introduced to define the behavior of light signals and the motions of free falling particles. However, this approach has been frequently criticized, in particular by Ehlers, Pirani, and Schild[1] on the following grounds:

(a) Whereas in this approach the whole mathematical structure of general relativity can be obtained from just two axioms, the physical origin and meaning of each of the derived geometric structures is correspondingly obscured. There is little scope for modifying the theory, by, for example, rejecting the metric structure but retaining the conformal and projective structures.

(b) Once the metric coefficients g_{ab} have been defined by the chronometric hypothesis, there is no overwhelming physical justification for postulating that these coefficients also determine the motion of freely falling particles and light signals, via the geodesic hypotheses.

(c) It has been shown by Marzke and Wheeler[2] and by Kundt and Hoffman[3] that it is possible to construct a standard clock from the paths of freely falling particles and light rays. Thus the geodesic hypotheses alone imply a physical interpretation of the metric in terms of time, independently of the chronometric hypothesis. Then either the chronometric hypothesis is redundant or else it is reduced to a statement of the equality of gravitational time, measured by geodesic clocks, and atomic time, measured by standard clocks: Such a statement is out of place in a theory which does not embrace both atomic and gravitational phenomena.

Another objection concerns the mathematical construction of the theory, rather than its physical interpretation. It is usual to impose certain additional restrictions on physically reasonable space–times. The least restrictive of these is that there should be no closed nonspacelike curves, but other causality conditions, such as the strong causality condition,[4] can also be justified in physical terms. Now the metric tensor g_{ab} cannot be defined independently of the background manifold, and yet even without any causality condition, the metric, which is an entirely local structure, imposes certain restrictions on the global topology of the manifold (for instance, there are very few manifolds on which it is possible to define a metric which is homogeneous and spatially isotropic). With the causality condition, the situation is even worse: For instance, it has been shown by Geroch[5] that, under certain conditions, a 4-geometry cannot have two slices (roughly, spacelike sections) with different topologies unless it is acausal. This is a restriction on the topology of space–time which cannot be formulated until the metric has been defined, and yet the metric cannot be defined until the background manifold has been specified. Of course, in practical terms, this is not a problem. The field equations for the metric can, in any case, only be solved locally: The global topology of space–time is then treated as an extension problem. It does, however, lead to certain difficulties for instance in the definition of a singularity. There is a perfectly clear intuitive idea of what is meant by a singularity and by such statements as "in the neighborhood of a singularity", but in general relativity, because a singularity is a place (in some sense) where the theory breaks down, singularities can only be discussed in terms of geodesic incompleteness; this is an inevitable consequence of the fact that general relativity is essentially a local theory; it only says anything about the geometry of space–time in the locality of nonsingular points. The situation is doubly unsatisfactory, firstly because singularities become a global problem whereas one feels intuitively that they should be a local phenomenon and secondly because at least the projective or the Weyl structure (and hence an idea of geodesic incompleteness) is needed even to define a singularity; and yet it seems that it should be possible to give a definition in terms of more primitive structures: that, for instance, a singularity should be a point where the differentiable structure of the background manifold breaks down. This is not a useful definition if the background manifold is fixed and independent of the physical processes happening in it.

Various attempts have been made at giving a local characterization of a singularity by adding the singular points as a boundary on a geodesically incomplete space–time. For instance, Geroch[5] does this by identifying various classes of incomplete timelike geodesics. The problem is that there are many such possible boundaries with no overiding physical reason for choosing one instead of another and that it seems scarcely logical to define the topological structure of space–time at a singular point in terms of, for instance, the world lines of freely falling particles but to take the differentiable and topological structure at nonsingular points as absolute and fixed. It would be more reasonable to define all the existing geometric structures of space–time, both at singular and at nonsingular points, in terms of more primitive concepts, such as the world lines of freely falling particles and light signals, and to define as singular a point of space–time where one of these geometric structures breaks down. Even if this procedure did not lead to any radically new techniques for analyzing singularities, it would at least provide a criterion for choos-

ing the most physically reasonable boundary on an incomplete space-time.

Ehlers, Pirani, and Schild[1] have shown how the projective, conformal, Weyl and metric structures of space-time can be built up successively from a few physically well-motivated axioms: Each axiom can be thought of as the reason for believing in a certain geometric structure which is normally derived from the chronometric hypothesis. The basic mathematical objects are taken to be two of the simplest objects in physics: light rays and the paths of freely falling particles (thought of as one-dimensional manifolds). The concept of a standard clock is abandoned completely.

The purpose of the present paper is to present a rigorous derivation of the differentiable and causal structure (in the sense of Kronheimer and Penrose[4]) of space-time from axioms concerning much the same primitive concepts as those used by Ehlers, Pirani, and Schild[1] (except that the idea of a light signal between two events, rather than a light ray in space-time, is taken as fundamental) and thus to complete the resolution of the difficulties stated above. The criterion for accepting a particular axiom is that it should have a simple and intuitively obvious physical interpretation and that it should involve only concepts introduced in previous axioms. At each stage the physical interpretation of the axiom is stressed: Thus the axioms together can be thought of as reasons for believing in the differentiable structure of space-time.

The only global restrictions imposed on space-time are introduced right at the beginning, in Axioms 1a and 1b, which deal, essentially, with the relation between the causal and topological structures (a local version of Axiom 2 can, in fact, be used without loss of anything but simplicity). Thus it is possible, at least in principle, to avoid the unsatisfactory feedback in the normal development, where local structures impose restrictions on a previously defined global topology. Further, Axioms 1a and 2b express properties which can reasonably be required of any physical space-time, even one with singularities. Thus the breakdown of any one of the properties expressed in the other axioms can be taken as the definition of a singularity in space-time: This definition will not depend on the global structure or the projective structure.

The main purpose of the paper is to give concrete expression to a point of view: that, even in classical relativity theory, the continuum structure of space-time should not be taken as absolute but should be derived from more primitive and more intuitively obvious concepts.

The basic objects in space-time are taken to be light signals between events and the world lines of spherically symmetric nonrotating freely falling particles, on which it is assumed there is some continuous idea of time. As far as the causal and conformal structures are concerned, there is nothing special about freely falling particles: Any class of particles could be taken, provided there is exactly one particle through each event in each direction. One important advantage of this approach over that of Ehlers, Pirani, and Schild is that no assumption is made, *ab initio*, about the paths in space-time along which light signals propagate: These paths are deduced from statements about the emission and absorbtion of light. All the considerations are entirely classical: For instance, it is assumed that, by some process of extrapolation, it is possible to define the world line of a freely falling particle through any event in space-time.

An outline of the argument is as follows:

Section 1: The causal structure of the event set is introduced and the global restrictions imposed. The topology is introduced.

Section 2: Light signals between nearby events are discussed.

Section 3: The differentiable structure of space-time is introduced.

Section 4: It is shown that the causal structure defines a conformal structure and that light signals must propagate along conformal null geodesics.

Throughout I rely heavily on the work of Kronheimer and Penrose[4] and that of Ehlers, Pirani, and Schild.[1]

1. DEFINITIONS AND CAUSAL STRUCTURE

Definition 1.1: A space-time is a pair (M, P) where M is a point set whose elements are called and P is a set of subsets of M, each with the structure of a c^0 one-dimensional manifold, homeomorphic with $\mathbb{R}$. The elements of P are called particles; it is assumed that there is at least one particle through each event.

The c^0 structure on a particle is interpreted as a continuous idea of time; the homeomorphism with $\mathbb{R}$ excludes the possibility of closed particle world lines. Each particle has two possible orientations. A particular choice of orientation of $p \in P$ defines an antireflexive linear ordering on p, denoted by $\ll_p$; if $x \in p$ and $y \in p$, then $x \ll_p y$ is interpreted: x chronologically precedes y on the world line of a freely falling particle.

Definition 1.2: (a) If a particular choice of orientation has been made for each particle in P and if x and y are any two events, then a *trip* from x to y is a sequence of events $x = z_0, z_1, \ldots, z_n = y$ together with a sequence of particles $p_1, p_2, \ldots, p_n$ such that, for $i = 0, 1, \ldots, n-1$,

$$z_i \in p_{i+1}, \quad z_{i+1} \in p_{i+1}, \quad \text{and} \quad z_i \ll_{p_{i+1}} z_{i+1}.$$

(b) x *chronologically precedes* y (written $x \ll y$) if there is a trip from x to y.

A trip is the world line of a massive particle which undergoes a finite number of collisions but which is otherwise freely falling. In order that this definition of chronological precedence should make sense physically it is necessary to postulate:

Axiom 1a: (first causality axiom): It is possible to choose the orientation of each $p \in P$ so that $\forall\, m \in M$, not $m \ll m$.

The interpretation of Axiom 1a is that on each particle there is a natural arrow of time and that no particle can enter its own past by undergoing a finite number of collisions. It is an immediate consequence of Axiom 1a that, with this choice of orientation, $\ll$ is an antireflexive partial ordering on M. Following Kronheimer and Penrose,[4] I make the following definitions:

Definitions 1.3: (a) if $x \in M$ then $I^+(x) = \{y \in M \mid x \ll y\}$ and $I^-(x) = \{y \in M \mid y \ll x\}$.

(b) $x < y$ ("x causally precedes y") if $I^+(x) \supset I^+(y)$ and $I^-(x) \subset I^-(y)$.

(c) The *Alexandrov topology*, $\mathcal{T}$ on M is the topology generated by $\{I^+(y) \mid y \in M\} \cup \{I^-(y) \mid y \in M\}$.

(d) $x \uparrow y$ *(horismos relation)* if $x < y$ but not $x \ll y$.

Axiom 2: If $x \in M$ and $p \in P$, then $p \cap I^+(x)$ and $p \cap I^-(x)$ are open in p.

Roughly speaking, Axiom 2 says that if there is one trip from x to an event on p then there is another faster one: Though there is an upper bound to the (average) speed of massive particles, this is not actually attained. The topology $\mathcal{T}$ now not only induces the right topology on every particle, but is also coincides with one idea of space-time topology: Suppose a sequence of events $\{x_n\}$ converges to an event x in the $\mathcal{T}$ topology. If p is some particle through x with fixed parametrization (i.e., time scale) and (t_1, t_2) is an arbitrarily small time interval in p, containing x, then an infinite number of the x_n are to the future of t_1 and to the past of t_2: The time interval (on p) needed to travel from p to x_n and back to p goes to zero as n goes to infinity. Thus there is an obvious physical interpretation of the convergence.

A further axiom is necessary to make M a causal space and to ensure that $\mathcal{T}$ has reasonable properties: this is a stronger version of Axiom 1a. It has been separated from Axiom 1a for the sake of clarity. But first, another definition:

Definition 1.4: If $x \in M$ and $y \in M$, then x *almost causally precedes* y (written x A y) if, $\forall\ z \in I^-(x)$, $I^+(z) \supset I^+(y)$. Equivalently, $\forall\ z \in I^+(y)$, $I^-(z) \supset I^-(x)$.

The point of this is that no real physical measurement is ever made at a single event: An event should really be thought of as the limit of its neighborhoods. If x almost causally precedes y, then every neighborhood of x contains events which chronologically precede events in any neighborhood of y, and no real physical experiment would reveal that x and y are not causally related. It is, therefore, reasonable to accept:

Axiom 1b: (second causality axiom): If $x \in M$ and $y \in M$, then x A y and y A $x \Rightarrow x = y$.

Lemma 1.1: $\mathcal{T}$ is Hausdorff.

If x and y are distinct events, then either not x A y or not y A x. Suppose not x A y. Then $\exists\ z_1 \in I^-(x)$ and $z_2 \in I^+(y)$ with not $z_1 \ll z_2$ and not $z_2 \ll z_1$. Then $I^+(z_1)$ and $I^-(z_2)$ are disjoint open neighborhoods of x and y, respectively. ■

Lemma 1.2: $\ll$ is future and past distinguishing, that is, $I^+(x) = I^+(y) \Rightarrow x = y$, and $I^-(x) = I^-(y) \Rightarrow x = y$.

If $I^+(x) = I^+(y)$, then x A y and y A x, so $x = y$. ■

Lemma 1.3: $\ll$ is full: That is, $\forall\ x \in M\ \exists\ y \in M$ such that $y \ll x$, and, if $y_1 \ll x$ and $y_2 \ll x$, then $\exists\ z \ll x$ with $y_1 \ll z$ and $y_2 \ll z$ (and the statement obtained by reversing $\ll$ also holds).

Let $x \in M$ and choose $p \in P$ with $x \in p$. p is homeomorphic with $\mathbb{R}$, so certainly $\exists\ y \in p$ with $y \ll x$. If $y_1 \ll x$ and $y_2 \ll x$, then $I^+(y_1) \cap I^+(y_2) \cap p$ is open in p and contains x. Hence $\exists\ z \in I^+(y_1) \cap I^+(y_2) \cap p$ such that $z \ll x$. ■

It is now clear that $(M, \ll, <, \uparrow)$ is a full causal space, according to the definition of Kronheimer and Penrose (Ref. 4, p. 485). The following conditions are fulfilled ($\forall\ x, y, z \in M$):

(I) $x < x$,

(II) $x < y$ and $y < x \Rightarrow x < z$,

(III) $x < y$ and $y < x \Rightarrow x = y$,

(IV) not $x \ll x$,

(V) $x \ll y$ and $y \ll z \Rightarrow x \ll z$,

(VI) (a) $x < y$ and $y \ll z \Rightarrow x \ll z$,

(b) $x \ll y$ and $y < z \Rightarrow x \ll z$,

(VII) $x \uparrow y$ if $x < y$ and not $x \ll y$.

2. LIGHT PROPAGATION BETWEEN NEARBY EVENTS

The crucial property of the light signals emitted at an event x is that they can be seen on the boundary of the future of x, $\dot{I}^+(x)$. It would seen natural, therefore, to introduce light propagation into this scheme by saying that there is a light signal from x to y if $y \in \dot{I}^+(x)$, equivalently $I^+(x) \supset I^+(y)$ and not $x \ll y$ (emission definition). However, it seems equally natural to say that there is a light signal from x to y if $x \in \dot{I}^-(y)$ (absorbtion definition): The light signals absorbed at y were emitted on the boundary of the past of y. The symmetry between emission and absorbtion could, possibly, be maintained by adopting the following definition: There is a light signal from x to y if either $I^+(x) \supset I^+(y)$ or $I^-(x) \subset I^-(y)$ but not $x \ll y$. But then suppose that $x \in M$ and $y \in M$, with $I^+(x) \supset I^+(y)$ but not $x \ll y$; choose a particle p with $x \in p$. In general $I^-(y)$ will intersect p in an open set of which x is *not* a limit point. There will thus be a finite interval on p, which is visible at y. Equally there may be single events which can be seen by an observer on a particle over a finite length of time. If either the emission or the absorption definition alone is adopted, then one is led by a similar argument to a situation where a clock (that is admissible parametrization) on one particle is seen to vary discontinuously by an observer on another particle: This situation is unacceptable, at least for nearby particles. The only way out of this dilemma, if the structure of M is not to be too severely restricted, is to adopt the following axiom:

Axiom 3: Every event has a future and past reflecting neighborhood; that is, $\forall\ x \in M\ \exists\ N_x$, an open neighborhood of x, such that if $y \in N_x$ and $z \in N_x$, then

$$I^+(y) \supset I^+(z) \Rightarrow I^-(z) \supset I^-(y)$$

and

$$I^-(y) \supset I^-(z) \Rightarrow I^+(z) \supset I^+(y)$$

[from now on N_x will always denote a future and past reflecting neighborhood of x].

Now it is possible to introduce at least a local definition of light signals: If $y \in N_x$, for some N_x, then there is a light signal from x to y if and only if $x \uparrow y$. Nothing more will be said, at least at this stage, about light signals between events which do not belong to each others future or past reflecting neighborhoods, or about the paths along which light signals travel.

Let p be a particle; for each $x \in p$ take U_x to be a future and past reflecting neighborhood of x of the form $I^-(m_1) \cap I^+(m_2)$, where $m_1 \in p$ and $m_2 \in p$, and let U_p be the union of all such neighborhoods U_x for all events $x \in p$. U_p is an open neighborhood of p. If $z \in U_p$, then $z \in U_{x_0}$ for some $x_0 \in p$. Define $f^+(z) = \inf_{x \in p}\{x \mid x \ll_p z\}$ and $f^-(z) = \sup_{x \in p}\{x \mid x \ll_p z\}$; sup and inf refer to the linear ordering $\ll_p$ on p. $f^+(z)$ and $f^-(z)$ certainly both exist and belong to U_{x_0}. Also $I^+(z) \supset I^+(f^+(z))$ and $I^-(z) \supset I^-(f^-(z))$, but not $^0 z \ll f^+(z)$ and not $z \gg f^-(z)$. U_{x_0} is a

future and past reflecting neighborhood; hence $z \uparrow f^+(z)$ and $f^-(z) \uparrow z$. This proves:

Lemma 2.1: For any particle p there exist two functions $f^+: U_p \to p$ and $f^-: U_p \to p$ such that $z \uparrow f^+(z)$ and $f^-(z) \uparrow z$, $\forall\ z \in U_p$. These properties uniquely define f^+ and f^-.

These functions are called message functions: They describe the process whereby an observer moving with p sees and is seen at nearby events.

Lemma 2.2: $\forall\ p \in P$, $f^+: U_p \to p$ and $f^-: U_p \to p$ are open and continuous maps.

Take, as a basis for the topology of U_p, open subsets of the form $\langle y_1, y_2\rangle = \{z \in M \mid y_1 \ll z \ll y_2\}$, where $y_1 \in U_p$, $y_2 \in U_p$ and $\exists\ p_y \in P$ with $y_1 \in p_y$ and $y_2 \in p_y$.

Certainly if $y_3 \in \langle y_1, y_2\rangle$, then $f^+(y_3)$ lies strictly between $f^+(y_1)$ and $f^+(y_2)$ on p. Also, by the same argument as that used in the proof of Lemma 2.1, every event on p, lying strictly between $f^+(y_1)$ and $f^+(y_2)$, is the image under f^+ of some point on p_y, lying strictly between y_1 and y_2. Thus the image under f^+ of $\langle y_1, y_2\rangle$ is the open interval in p bounded by $f^+(y_1)$ and $f^+(y_2)$. This proves f^+ is an open map. Similarly f^- is also open.

Let (x_1, x_2) be an open interval in p. Then $f^{+-1}((x_1, x_2)) = U_p \cap (I^-(x_2) \sim \overline{I^-(x_1)})$, which is an open set in U_p. For if $z \in U_p \cap (I^-(x_2) \sim \overline{I^-(x_1)})$, then $z \in U_{x_0}$ for some $x_0 \in p$. $f^+(z)$ certainly lies in the closed interval $[x_1, x_2]$ in p. Suppose $f^+(z) = x_1$; then $I^-(x_1) \supset I^-(z)$ and every neighborhood of z intersects $I^-(x_1)$, contradicting not $z \in \overline{I^-(x_1)}$. Suppose $f^+(z) = x_2$; then not $z \in I^-(x_2)$, which is also a contradiction. Hence $f^+(U_p \cap (I^-(x_2) \sim \overline{I^-(x_1)})) \subset (x_1, x_2)$. Conversely, if $z \in U_p$ and $f^+(z) \in (x_1, x_2)$, then $z < f^+(z)$ and $f^+(z) \ll x_2$, so $z \in I^-(x_2)$. If $z \in \overline{I^-(x_1)}$, then $I^-(x_1) \supset I^-(z)$ and so not $f^+(z) \gg x_1$: a contradiction. Thus $z \in U_p$ and $f^+(z) \in (x_1, x_2)$ implies that $z \in (I^-(x_2) \sim \overline{I^-(x_1)}) \cap U_p$. This proves that $f^{+-1}((x_1 x_2)) = U_p \cap (I^-(x_2) \sim I^-(x_1))$ and that f^+ is continuous (similarly f^- is also continuous). ■

It is easy to show that the $\mathcal{T}$ topology is the smallest topology which makes all the functions f^+ and f^- continuous: That it is the smallest topology which makes space-time "look" continuous for every observer in free fall. This is a further confirmation of the fact that the Alexandrov topology is the right topology for space-time: That is, it is the topology which embodies the most physically reasonable idea of continuity.

3. DIFFERENTIABLE STRUCTURE

So far nothing has been said about the dimension of space-time. There are a number of ways in which the dimension of a causal space can be fixed, using variations of the statements: In general, n light cones meet in a single point (cf. Woronowicz[6]). The simplest, for the present purpose, is embodied in axiom 4a (remember that the message functions are open maps).

Axiom 4a: (first dimension axiom): Given $p_1 \in p$ such that the four message functions defined by p_1 and p_2, together with homeomorphisms $p_1 \to \mathbb{R}$ and $p_2 \to \mathbb{R}$, define a one-to-one map from a neighborhood of every point in $(U_{p_1} \cap U_{p_2}) \sim (p_1 \cup p_2)$ onto an open set in $\mathbb{R}^4$. Further every event belongs to $(U_{p_1} \cap U_{p_2}) \sim (p_1 \cup p_2)$ for some such pair of particles p_1 and p_2.

Not every pair of particles can be expected to define such a map: In Minkowski space, for example, two coplanar timelike straight lines do not define a one-to-one map into $\mathbb{R}^4$, but a pair of skew timelike lines do. (I am grateful to Professor Penrose for pointing this out). Combining Axiom 4 with Lemma 1.1 and Lemma 2.2, we have the following:

Theorem 1: M is a c^0 four-dimensional real manifold.

The local coordinates are defined by the four message functions of two nearby particles. These are the "radar coordinates" [1]: Each event is parametrized by the time at which a signal is sent from one of the particles to the event and the time at which the echo from the event is received back at the particle.

Consider a particle p_1, and suppose that a second particle, p_2, intersects U_{p_1}. Then the two functions $f^+|_{p_2}: p_2 \to p_1$ and $f^-|_{p_2}: p_2 \to p_1$ are continuous, and, in fact, strictly monotonic; they describe the relationship between a clock on p_1, say, and a clock on p_2, as seen at p_1. Now each particle is homeomorphic with $\mathbb{R}$, that is, on each particle there is a continuous idea of time: All clocks used to measure time are continuously related. Each homeomorphism to $\mathbb{R}$ induces a differentiable structure on the particle; each clock defines a preferred class of clocks which are differentiably related. Suppose that it is possible to choose a differentiable structure on each particle, by, for instance, using the same type of clock on each particle, in such a way that the message functions between every pair of nearby particles are differentiable, all of class C^N say. This situation is one possible interpretation of the statement that "space-time looks as if it has the N^{th} degree of smoothness." The aim of this section is to show that, under certain additional axioms, space-time is a differentiable manifold which is as smooth as it looks.

Intuitively, at any event each particle defines a (timelike) direction in space-time. It is possible to introduce a notion of continuously varying direction by the following definition:

Definition 3.1: A set $\mathcal{C} = \{p_\lambda, t_\lambda \mid \lambda \in \Lambda\}$ of particles with given parametrization, $t_\lambda: p_\lambda \to \mathbb{R}$, is a c^0-congruence on $U\ M$ if:

(I) $\forall\ x \in U$ there is exactly on $(p_x, t_x) \in \mathcal{C}$ such that $x \in p_x$ and $t_x(x) = 0$.

(II) For each r in some neighborhood of 0 in $\mathbb{R}$, the map $U \to M: x \to t_x^{-1}(r)$ is a c^0-morphism.

Physically, a c^0-congruence is a fluid. Now suppose a preferred differentiable structure has been chosen on each particle. Let $\mathcal{C}$ be a c^0-congruence, each of whose elements has a differentiable parametrization; it is reasonable to suppose that, in addition to being continuous in the above sense, $\mathcal{C}$ also "looks" continuous. Any observer on a nearby particle p, with parametrization $t: p \to \mathbb{R}$, can assign two parameters to each element p_x of $\mathcal{C}$ which describe the relative velocity of p_x (at $x \in p_x$) with respect to p; physically these are the red shift of p_x and the doppler shift of signals bounced off p_x. The information is contained in the derivatives V_1 and V_2 of the two functions: $t \circ f_x^+ \circ t_x^{-1}$ and $t \circ f_x^- \circ t_x^{-1}$ at 0, where $f_x^+: p_x \to p$ and $f_x^-: p_x \to p$ are the message functions defined by p_x. If $\mathcal{C}$ is to appear continuous to the observer on p, then these two parameters must vary continuously over U_p. Hence:

Axiom 5: If $p \in P$ and $\mathcal{C}$ is a c^0-congruence on $V \subset U_p$, then the two functions

$$V_1 : V \to \mathbb{R} : x \to \{t \circ f_x^+ \circ t_x^{-1}\}'|_0,$$
$$V_2 : V \to \mathbb{R} : x \to \{t \circ f_x^- \circ t_x^{-1}\}'|_0$$

are continuous.

(It may be necessary to take U_p to be smaller than the neighborhood considered previously so as to avoid the possibility of U_p containing caustics in the null cones of points on p.)

A second definition of dimension can be given: This is, roughly, that there should be four independent directions at each point. A set $\{e_1 \cdots e_n\}$ of c^0-congruences on a neighborhood U of x is independent at x if the set of events which can be reached from x by going first along an element of e_1 then along one of e_2, etc., is a c^0-submanifold of U of dimension n. A set of directions at x, that is, particles through x, is independent if each particle can be extended to a c^0-congruence on a neighborhood of x, and these c^0-congruences are independent. The set of directions at each point must be four dimensional, that is:

Axiom 4b: (second dimension axiom): There are four independent directions at every event.

Axiom 4b also implies the existence of c^0-congruences: The fact that a particle through an event can be extended to a c^0-congruence on a neighborhood of that event can be thought of as stating that there are no gravitational singularities at the event, as the gravitational field can be measured, in some unspecified way, by the divergence of a congruence of particles through x.

It now follows that:

Theorem 2: M is a C^1-manifold, if all the message functions between particles are of class C^1.

Define the set ${C^1}_M$ of differentiable functions on M by $f : M \to$ is in ${C^1}_M$ if:

(i) $f|_p : p \to \mathbb{R}$ is of class C^1 for every particle p.

(ii) The derivatives of f along the particles of any c^0 congruence on U (with respect to the parameters on the particles of the congruence) define a c^0 function on U.

As local coordinates, use radar coordinates together with differentiable parametrizations. Then it follows from Axiom 5 that the local coordinates belong to ${C^1}_M$ (locally). In local coordinates, a coordinate patch U is mapped onto an open subset $\tilde{U}$ of $\mathbb{R}^4$, and a c^0-congruence on U is mapped onto a family of C^1 curves in $\tilde{U}$, with continuously varying tangent vectors. Using Axiom 4b, together with arguments from elementary calculus, it can be shown that any function $f \in {C^1}_M$ is represented on $\tilde{U}$ by a function of class C^1 (mapping $\tilde{U}$ to $\mathbb{R}$). Further any function of class C^1 on $\tilde{U}$ trivially induces a function on U which belongs to ${C^1}_M$ (locally). Lastly, as the definition of ${C^1}_M$ is entirely local, any function which agrees with an element of ${C^1}_M$ on a neighborhood of each point of M must also belong to ${C^1}_M$. ■

The C^N structure on M can now be defined inductively: C^1-congruences are defined on M and modified versions of Axioms 5 and 4b are introduced, and hence the C^2 structure is introduced. The process can be repeated successively to show that M is as smooth as the message functions between particles: that M is as smooth as it looks.

4. CONFORMAL STRUCTURE AND LIGHT PROPAGATION

The causal structure on M can be used to define a conformal metric (M is now taken to be as smooth as is necessary). Consider a point $x \in M$ and a particle p through x. Choose a particular differentiable parametrization t: $p \to \mathbb{R}$ on p, such that $t(x) = 0$. Now define the function g: $U_p \to \mathbb{R}$ by $g(z) = t(f^+(z)) \times t(f^-(z))$. Thus $g(z) = 0$ if and only if $z \in U_p$ and $z \uparrow x$ or $x \uparrow z$, so $\{z \in U_p | g(z) = 0\} = \overset{\circ}{I}{}^+(x) \cap \overset{\circ}{I}{}^-(x) \cap U_p$. But $\overset{\circ}{I}(x) \cap U_p$ and $\overset{\circ}{I}{}^-(x) \cap U_p$ are three-dimensional submanifolds of M, everywhere except at x (because they are everywhere surfaces on which the radar coordinates defined by p are constant). $g(z) = 0$ cannot define a hypersurface at x itself, or there would be points in every neighborhood of x which belong to both $\overset{\circ}{I}{}^+(x)$ and $\overset{\circ}{I}{}^-(x)$ (in addition to x itself). This would contradict Axiom 3. Hence, in local coordinates at x, $g_{,a}(x) = 0$ and $g_{,ab}(x)$ defines a tensor. Differentiating g twice along p gives $g_{ab} V^a V^b = 2$, where V^a is the tangent vector to p at x. Hence $g_{ab} \neq 0$. Further, it can easily be seen that, up to a scalar factor, g_{ab} is independent of both p and t.

Any particle (with given parametrization) through x is a curve whose tangent vector V^a at x must satisfy $g_{ab} V^a V^b > 0$. Any curve through x which lies in $\overset{\circ}{I}{}^+(x)$ or $\overset{\circ}{I}{}^-(x)$ must have a tangent vector n^a which satisfies $g_{ab} n^a n^b = 0$. Any point in a curve through x in $\overset{\circ}{I}{}^+(x)$ is the limit of a sequence of events which can be reached from x by trips, that is, in physical terms, by massive particles undergoing a finite number of collisions. Thus a vector at x which satisfies $g_{ab} n^a n^b = 0$ is either a limiting (average) velocity for trips through x or minus such a limiting velocity. Now it seems to be true that in the real world the limiting speed at which massive particles can travel is finite. So it is reasonable to suppose that no vector n^a can be a limiting average velocity for both future and past directed trips. Hence:

Axiom 6: g_{ab} is everywhere nonsingular.

g_{ab} is the conformal metric on M: It is defined at all nonsingular points, up to a scalar factor. It is easy to see that the only possible signature for g_{ab} is $(+ - - -)$. (The above contruction is given in Ref. 1.)

It will now be shown that, with an obvious definition of light paths, light must propagate along smooth curves which are conformal null geodesics.

Definition: (a) (nonlocal definition of light signals): If $x \in M$ and $y \in M$, then there is a light signal from x to y if there is a sequence of events $z_0 = x, z_1, z_2, \ldots, z_n = y$ such that, for $i = 1, \ldots, n-1$,

(1) $z_{i-1} \in N_{z_i}$ and $z_{i+1} \in N_{z_i}$, for some N_{z_i},

(2) $z_{i-1} \uparrow z_i$ and $z_i \uparrow z_{i+1}$ and $z_{i-1} \uparrow z_{i+1}$.

(b) A light path L is a connected set of events such that $\forall\, z \in L$ there is a neighborhood U_z of z such that $U_z \cap L$ is linearly ordered by $\uparrow$ and is not contained in any larger subset of U_z linearly ordered $\uparrow$.

Thus there is a light signal from x to y if there is a sequence of points between x and y, with light signals between them according to the local definition and if this

sequence does not bend (that is, there is a light signal—according to the local definition—between one point of the sequence and the next point along but one). A light path from x to y is, roughly speaking, the set of all events between x and y through which the light signal can travel. Both these definitions make good sense physically.

[Note: This is slightly more restrictive than Kronheimer and Penrose's[5] treatment of light signals in terms of girders and beams. The reason for this is that, in general, M is not globally future and past reflecting and so its causal structure does not coincide with its natural causal structure (in the sense of Kronheimer and Penrose[5] as a conformal manifold: The chronology relation is the same but the horismos relation is not. Consider, for example, the pseudo-Riemannian manifold obtained by removing the set $\{(x_0, x_1) \mid x_1 = 0,\ 0 \leq x_0 \leq 1\}$ from Minkowski space. Taking the particles as timelike straight lines the procedures given above lead to the correct conformal structure on M. However, if z_1 is the point $(-1,-1)$ and z_2 is the point $(1, 2)$, then the points z_1 and z_2 satisfy $z_1 \uparrow z_2$, even though there is certainly no light signal from z_1 to z_2 (and z_1 does not belong to a future reflecting neighborhood of z_2). The rather cumbersome definition given above avoids saying that there is a light signal from z_1 to z_2 whenever $z_1 \uparrow z_2$. With this causal structure (but *not* with the natural causal structure) there would be a light signal from z_1 to z_2 according to Kronheimer and Penrose's definition.]

Lemma 4.1: If $y \in M$, $x \in N_y$, for some N_y and $x \uparrow y$ and if U is any neighborhood of y, then there exists $z \neq y$ in U such that $x \uparrow z$ and $z \uparrow y$. Similarly, if W is any neighborhood of x, then there exists $\tilde{z} \neq x$ in W such that $x \uparrow \tilde{z}$ and $\tilde{z} \uparrow y$.

M is locally Euclidean and therefore locally compact. Choose $V_y \subset U \cap N_y$ with $y \in V_y$ and $\overline{V_y} = \{z \in M \mid m_1 \ll z \ll m_2\}$ such that $\overline{V_y}$ is compact. Pick a particle p through y and a point n of p with $n \ll y$ and $n \in V_y$. (Without loss of generality $m_1 \in p$ and $m_2 \in p$).

Put $I = \overline{I^+(x)} \cap \mathring{I}^+(n) \cap \overline{V_y}$. I is a closed subset of $\overline{V_y}$ and hence is compact. I is also nonempty, because $x \ll m_2$ and a trip from x to m_2 must contain one point of I.

Let $\{z_n\}$ be a sequence on p such that $z_n \gg y$ and $z_n \to y$ as $n \to \infty$. For each n, $I \cap \overline{I^-(z_n)} \cap \overline{V_y}$ is compact and nonempty (because $x \ll z_n$ and a trip from x to z_n must contain a point of $I \cap \overline{I^-(z_n)} \cap V_y$).

$\{\overline{I^-(z_n)} \cap \overline{V_y} \cap I\}$ is thus a nested sequence of nonempty compact sets. Hence there is at least one point $z \in \cap_{n=0}^{\infty} \overline{I^-(z_n)} \cap \overline{V_y} \cap I$).

Certainly $z \in U$ and $z > x$, since $z \in Z$. Let f^+ be a message function from $U_p \supset \overline{V_p}$ to p. Suppose $f^+(z) \gg y$; then $f^+(z) \gg z_n$ for some n and so not $z \in \overline{I^-(z_n)}$, a contradiction. Hence not $f^+(z) \gg y$ and so $z < y$. Further $z \in Z$ implies $z \neq y$. But $x < z < y$ and $x \uparrow y$ implies $x \uparrow z$ and $z \uparrow y$.

The proof of the second part is similar. ■

Lemma 4.2: If $x \in M, y \in N_x$, for some N_x, and $x \uparrow y$ then $\mathring{I}^+(x) \cap \mathring{I}^-(y) \cap V$ is a smooth path with end point y for some neighborhood V of y. Furthermore, light paths are one-dimensional submanifolds of M.

Remarks: (1) $\mathring{I}^+(x) \cap \mathring{I}^-(y) = \{z \in M \mid x \uparrow z \text{ and } z \uparrow y\}$.
(2) If $z \in \mathring{I}^+(x) \cap \mathring{I}^-(y)$ and $x \uparrow \tilde{z} \uparrow z$ or $z \uparrow \tilde{z} \uparrow y$, then $\tilde{z} \in \mathring{I}^+(x) \cap \mathring{I}^-(y)$.

Proof of Lemma 4.2: Choose a particle p through y with parametrization $t : p \to \mathbb{R}$ such that $t(y) = 0$. Define the function $\tilde{g}$: $\mathring{I}^+(x) \to \mathbb{R}$ by $\tilde{g}(z) = t(f^+(z))t(f^-(z))$, $z \in \mathring{I}^+(x)$, where f^+ and f^- are the message functions U_p to p. Then $\mathring{I}^-(y) \cap \mathring{I}^+(x) = \{z \in \mathring{I}^+(x) \mid \tilde{g}(z) = 0 \text{ and } z < y\}$.

$\mathring{I}^+(x)$ and $\mathring{I}^-(y)$ can never cross, so $\mathring{I}^-(y) \cap \mathring{I}^+(x) \subset \{z \in \mathring{I}^+(x) \mid g_{,\alpha}(z) = 0\}$, where $\{x^\alpha\}$ are local coordinates on $\mathring{I}^+(x)$ at y.

Now $g_{,\alpha\beta}(y)$ (a 3×3 matrix) must have rank at least 2, since $g_{ab}(y)$ is nonsingular. So $g_{,1} = 0$ and $g_{,2} = 0$, say, define a one-dimensional submanifold L of $\mathring{I}^+(x)$ in a neighborhood U of y; L contains $\mathring{I}^+(x) \cap \mathring{I}^-(y) \cap U$. By making U small enough, L can be taken to be homeomorphic with an open interval. By Lemma 4.2 there is a point $z \neq y$ of $L \cap U$ such that $x \uparrow z \uparrow y$.

Suppose there is an open interval Z in L between z and y with $w \in Z \Rightarrow$ not $w \uparrow y$. Let m be an end point of this interval. By Lemma 4.1 there are two sequences $\{u_n\}$ and $\{v_n\}$ in L which both converge to m and which satisfy, $\forall\ n, x \uparrow u_n \uparrow m, m \uparrow v_n \uparrow y, m \neq u_n$ and $m \neq v_n$. It is not possible that, but for a finite number of elements, both sequences lie on the same side of m in L or L would have a null tangent at m which would be both future and past directed. Thus it is not possible that m is an end point of Z, so Z cannot exist. Hence, if $\tilde{L}$ is the part of L between z and y, then $\mathring{I}^+(x) \cap \mathring{I}^-(y) \cap \overline{\tilde{L}}$ is dense in $\tilde{L}$. But $\mathring{I}^+(x) \cap \mathring{I}^-(y)$ is closed so $\tilde{L} \subset \mathring{I}^+(x) \cap \mathring{I}^-(y)$ and the first part is proved.

A slightly extended version of this argument shows that if L is a light path and $z \in L$, with $x \uparrow z$, say, and $x \in N_z \cap L$, then in some neighborhood W of z, L is exactly $W \cap \mathring{I}^+(x) \cap (\mathring{I}^+(z) \cup \mathring{I}^-(z))$ and that this set is a one-dimensional submanifold of W. This proves the second part. ■

The first part means that nearby events are "seen" in a definite direction.

Lemma 4.3: $\forall\ x \in M$ $(N_x \cap \mathring{I}^+(x)) \sim x$ and $(N_x \cap \mathring{I}^-(x)) \sim x$ are null hypersurfaces in N_x (with respect to the conformal structure for sufficiently small N_x).

As above, it can be shown that if $z \in (N_x \cap \mathring{I}^+(x)) \sim x$, then $\mathring{I}^-(z) \cap \mathring{I}^+(x)$ is a smooth curve in some neighborhood of z: so $N_x \cap \mathring{I}^+(x)$ must be null at z (provided that N_x is small enough for $\mathring{I}^+(x)$ to be differentiable everywhere in $N_x \sim x$).

Lemma 4.4: A light path is null curve.

If L is a light path and $z \in L$, then by Lemma 4.1 there is a sequence $\{x_n\} \subset L$ with $z \uparrow x_n$ and $x_n \to z$ as $n \to \infty$. Hence L intersects $\mathring{Z}^+(z)$ at points arbitrarily close to z and so must be tangent to $\mathring{Z}^+(z)$ at z: Thus L must be a null curve.

And so, finally:

Theorem 3: Light rays are conformal null geodesics.

[If a new horismos relation is defined on M by $x \tilde{\uparrow} y$ if and only if there is a light signal from x to y but not $x \ll y$, then the causal structure of M will coincide with its natural causal structure as a conformal manifold. Further light rays will be beams and the causal space will be regular.]

Conclusion: From this point the treatment of Ehlers, Pirani, and Schild[1] can be used to build up successively

the projective, Weyl and metric structures of space-time. (They take as their fundamental concepts the worldlines of freely falling particles and light *rays*, as opposed to light *signals*: One of their earlier axioms, which is redundant in the light of Theorem 3, is that light rays are smooth paths in space-time). An axiom equivalent to the law of inertia defines the projective structure of space-time and an axiom concerning the relationship between nearby gravitational clocks leads to the metric structure.

One possible interpretation of the results of this paper is that the kinematical behavior of particles and light signals determines the conformal structure of space-time, and hence the free gravitational field. The Riemannian structure of space-time depends on the more detailed structure of the matter in space-time, as expressed by its energy-momentum tensor.

Of course, throughout this approach the axioms become more and more intricate as the geometric structures they relate to become more and more complex: But what is really happening is that primitive geometric structures are being related to simple physical concepts and more specialized geometric structures are related to more sophisticated concepts. The process should not be thought of as adding unnecessary complications to a previously simple theory; it is rather one of bringing to light, that is, interpreting physically, complications that are inherent in the appearent simplicity of the chronometric and geodesic hypotheses.

---

*This work was carried out with the support of a Science Research Council Studentship.

[1]J. Ehlers, F. A. E. Pirani, and A. Schild, "The Geometry of Free Fall and Light Propagation", in *General Relativity, Papers in Honour of J. L. Syng* (Oxford U.P., Oxford, 1972).

[2]R. Marzke and J. A. Wheeler, in *Gravitation and Relativity*, edited by H.-Y. Chiu and W. F. Hoffman (Benjamin, New York, 1964).

[3]W. Kundt and B. Hoffman, in *Recent Developments in General Relativity* (Pergamon, New York, 1962), p. 303.

[4]E. Kronheimer and R. Penrose, Proc. Camb. Philos. Soc. **63**, 481 (1967).

[5]R. P. Geroch, thesis (Princeton University, 1967).

[6]S. L. Woronowicz, "A Theory of Causal Spaces" (unpublished).

A generalization of the equations governing the evolution of a particle distribution in a random force field

C. E. Newman

Institute for Plasma Research, Stanford University, Stanford, California
(Received 22 June 1972; revised manuscript received 9 November 1972)

The problem of a system of charged particles moving in a random force field and its application to the study of turbulent plasmas are discussed. We point out the inapplicability of existing formalisms to many cases of astrophysical interest. A generalization of this formalism is developed from two different points of view—a Fokker–Planck approach and a quasilinear approach. Both approaches lead to the same equation for the evolution of the particle distribution; this equation has the form of a Fokker–Planck equation, but the terms describing the effects of the random field do not retain the interpretation which they have in usual Fokker–Planck development. It is shown that this equation reduces to a quasilinear diffusion equation when the random field is electromagnetic.

I. INTRODUCTION

The study of the evolution of a distribution of charged particles moving in a random, or "turbulent," force field is very relevant to many problems of current interest in plasma physics and astrophysics. A proposal by Fermi that cosmic rays could be accelerated in a stochastic manner by moving magnetic "walls" distributed throughout the galaxy marked the historical origin of interest in this general area.[1,2] Since then various authors have proposed that the turbulent magnetic field of the solar wind is more important than scattering by collisions between individual particles; Jokipii[3] has calculated the effects of a turbulent magnetic field on the propagation of energetic particles through the solar wind, and Newman and Sturrock[4] have shown that turbulence is more important than collisions in determining the electrical conductivity of the solar wind. In addition, Manley and Olbert[5] have proposed that the observed electron energy spectra of x-ray stars arise due to the combined effects of stochastic acceleration and synchrotron radiation. Acceleration by stochastic fields has also been put forward as a mechanism for the production of high-energy particles produced in laboratory plasmas subject to the beam-plasma instability.[6]

In order to analyze problems of this general nature, two alternative methods have been employed. The first is the description of the evolution of the particle distribution in terms of a Fokker–Planck equation as discussed by Sturrock[7,8] and others[3,9]; the second method employs the quasilinear diffusion equation, as developed by Hall and Sturrock[10] and Kennel and Engelmann[11] from a classical point of view and by Melrose[12] and various Russian authors[13,14] by a semi-quantum-mechanical approach.[15] As we show later in this article, these two approaches are no longer valid if the problem at hand entails the consideration of nonfluctuating macroscopic fields and/or spatial gradients in addition to the small-scale random fields. Our aim in this article is then two-fold: First, we wish to derive the correct equations governing the evolution of the particle distribution function from both approaches without making this restriction; second, we must then compare the results of the two approaches and, if any differences arise, explain which method has more validity and why. As we will see, this latter task is easily dispatched since the final equations in both cases are identical. This in itself is an important result as it enhances our confidence in the validity of this equation.

In Sec. II we present a brief summary of Sturrock's derivation of the "traditional" Fokker–Planck equation[8] and show the difficulties which arise when one tries to apply the result to the general case described above. Section III gives a satisfactory method for generalizing the Fokker–Planck equation; the resultant expression is of the Fokker–Planck form, but the terms describing the effects of the stochastic forces have a slightly different interpretation. In Sec. IV we give a new derivation of the quasilinear equation which is compared with our generalization of the Fokker–Planck equation; it is shown that the two approaches lead to identical equations for the particle distribution function. In Sec. V, we restrict ourselves to electromagnetic fluctuations and show that in this case either approach leads to a diffusion-type equation in phase space of the form given by the quasilinear development of Hall and Sturrock,[10] but with slightly changed expressions for the diffusion coefficients.

II. THE TRADITIONAL FOKKER-PLANCK EQUATION

The Fokker–Planck equation has been derived in various ways by many authors.[8,9] In this section we give that of Sturrock[9] because it lends itself readily to the generalization given in Sec. III. However, before beginning, a few preliminary comments on notation and assumptions are necessary.

In the remainder of this article we concern ourselves with the behavior of a system of identical particles. We assume that this system is adequately described by a distribution function $F(Z_\mu, t)$, where Z_μ $(\mu = 1, 2, \cdots, 6)$ denotes the six phase-space coordinates x, y, z, p_x, p_y, p_z (Cartesian coordinates and their respective momentum components) and t is the time. F is so defined that $F^Z(Z_\mu, t)d^6Z$ is the number of particles in the phase-space volume d^6Z about Z_μ at time t.

We suppose further that a particle at Z_μ at time t obeys the equation of motion

$$\frac{dZ_\mu}{dt} = \mathcal{G}_\mu^Z(Z_\sigma, t), \tag{1}$$

where $\mathcal{G}_\mu^Z$ is a function of Z_σ and t only, as indicated. From (1) we see that knowledge of $\mathcal{G}_\mu$ for all Z_σ and t and of the position in phase space of a particle at some given time completely determines the position of that particle as a function of time.

Using the notation originally employed by Hall and Sturrock,[10] we separate the right-hand side of (1) into two parts as follows:

$$\frac{dZ_\mu}{dt} = G_\mu^Z(Z_\sigma, t) + g_\mu^Z(Z_\sigma, t), \tag{2}$$

where G_μ^Z denotes terms due to known external, non-fluctuating fields, and g_μ those due to the small-scale stochastic fields. Throughout this development we will be assuming that the g_μ^Z are small enough, in some sense, to render possible a perturbation, or weak-turbulence, expansion in orders of g.

The distribution function F^Z will then satisfy the phase-space continuity equation:

$$\frac{\partial F^Z}{\partial t} + \frac{\partial}{\partial Z_\mu}\left[(G_\mu^Z + g_\mu^Z)F^Z\right] = 0, \tag{3}$$

where the summation convection for repeated indices is used. The external forces with which we will be concerned will always be electromagnetic so that G_μ^Z will satisfy

$$\frac{\partial}{\partial Z_\mu} G_\mu^Z = 0; \tag{4}$$

hence (3) takes the form

$$\frac{\partial F^Z}{\partial t} + G_\mu^Z \frac{\partial F^Z}{\partial Z_\mu} + \frac{\partial}{\partial Z_\mu}(g_\mu^Z F^Z) = 0. \tag{5}$$

We do not assume that we know the behavior of the quantities g_μ^Z in detail. What we do assume is that there are many possible realizations of g_μ^Z consistent with the "macroscopic" behavior of the system (i.e., behavior smoothed out over time and length scales much larger than those of the fluctuating g_μ^Z); we then seek an equation governing the evolution of the system in terms of quantities of the form $\langle g_\mu^Z(Z_\sigma, t)\rangle$, $\langle g_\mu^Z(Z_\sigma, t) g_\nu^Z(Z'_\sigma, t')\rangle$, etc., where "$\langle\ \rangle$" denotes an average over an ensemble of possible realizations of g_μ^Z. We further assume that we may assign a coherence length L_c and a coherence time τ_c to the fluctuating field, such that the quantity

$$\left\langle \prod_{n=1}^{N} g_{\alpha_n}(Z_\sigma^{(n)}, t_n) \right\rangle$$

becomes negligibly small whenever any two of the times denoted by t_n, say t_i and $t_j (i, j < N)$, satisfy $|t_i - t_j| \gg \tau_c$ or whenever the spatial separation of any two of the phase-space positions denoted by $Z_\sigma^{(n)}$ significantly exceeds L_c. L_c and τ_c then provide measures for the length and time scales of the fluctuating field, or, alternatively, they may be thought of as the length and time over which a particle is scattered once by the fluctuations.

Now, in our notation the technique employed by Sturrock proceeds as follows: Consider the particles in a volume of phase space d^6Z about the point Z_μ at time t. At time t', these particles will occupy a volume d^6Z' about a new point Z'_μ; but Eq. (5) just expresses the conservation of particles in phase space, thus we must have

$$F^Z(Z_\mu, t) d^6Z = F^Z(Z'_\mu, t') d^6Z', \tag{6}$$

where Z'_μ is found by solving (1) [or (2)], with Z_μ and t replaced by Z'_μ and t', subject to the initial condition that $Z'_\mu = Z_\mu$ when $t' = t$. Sturrock then showed that, if we let

$$\begin{aligned} t' &= t + \Delta t, \\ Z'_\mu &= Z_\mu + \Delta Z_\mu(Z_\sigma, t, \Delta t) \end{aligned} \tag{7}$$

$F^Z(Z_\mu, t + \Delta t)$ can be found in terms of $F^Z(Z_\mu, t)$ via the following formula:

$$F^Z(Z_\mu, t + \Delta t) = F^Z(Z_\mu, t) + \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial Z_{\alpha_1} \cdots \partial Z_{\alpha_n}} \times \left[\Delta Z_{\alpha_1} \cdots \Delta Z_{\alpha_n} F^Z(Z_\mu, t)\right]. \tag{8}$$

The expansion given in (8) is valid for sufficiently small Δt if F^Z is well-behaved since $\Delta Z_\alpha \to 0$ as $\Delta t \to 0$. Thus, we may write (8) as

$$\frac{\Delta F^Z}{\Delta t} = -\frac{\partial}{\partial Z_\alpha}\left(\frac{\Delta Z_\alpha}{\Delta t} F^Z\right) + \frac{1}{2}\frac{\partial^2}{\partial Z_\alpha \partial Z_\beta}\left(\frac{\Delta Z_\alpha \Delta Z_\beta}{\Delta t} F^Z\right) - \cdots, \tag{9}$$

where

$$\Delta F^Z = F^Z(Z_\mu, t + \Delta t) - F^Z(Z_\mu, t). \tag{10}$$

We are actually looking for an equation for $\langle F^Z\rangle$, where the averaging process denoted by "$\langle\ \rangle$" has been defined above. We also wish to know the behavior of $\langle F^Z\rangle$ smoothed out over many collisions. To this end we must make two assumptions. First, we assume that there is a time interval Δt short enough that $\langle F^Z\rangle$ and G_μ^Z do not change very much but long enough that $\Delta t \gg \tau_c$. In such a case,

$$\begin{aligned} \left\langle \frac{\Delta Z_\mu}{\Delta t} \right\rangle &= G_\mu^Z + \left\langle \frac{\Delta Z_\mu}{\Delta t} \right\rangle_c, \\ \left\langle \frac{\Delta Z_\mu \Delta Z_\nu}{\Delta t} \right\rangle &= G_\mu^Z G_\nu^Z \Delta t + \left\langle \frac{\Delta Z_\mu \Delta Z_\nu}{\Delta t} \right\rangle_c \\ &\quad + G_\mu \Delta t \left\langle \frac{\Delta Z_\nu}{\Delta t} \right\rangle_c + G_\nu \Delta t \left\langle \frac{\Delta Z_\mu}{\Delta t} \right\rangle_c, \end{aligned} \tag{11}$$

and all such cumulants of higher order in ΔZ_μ are at least of order Δt, i.e., go to zero as Δt goes to zero. The quantities $\langle \Delta Z_\mu / \Delta t\rangle_c$ and $\langle \Delta Z_\mu \Delta Z_\nu / \Delta t\rangle_c$ are due to the terms containing g_μ^Z in the equations of motion (2); both will in general yield a contribution to zero order in Δt. Our second assumption is that the process under consideration is a Markov process, i.e., that the distribution function $\langle F^Z\rangle$ does not "remember" any of its past history beyond a few coherence times. This is tantamount to assuming that $\langle g_\alpha^Z(Z_\sigma, t) F^Z(Z_\sigma, t + \Delta t)\rangle = \langle g_\alpha^Z(Z_\sigma, t)\rangle \langle F^Z(Z_\sigma, t + \Delta t)\rangle = 0$ for $\Delta t \gg \tau_c$, i.e., that $F^Z(Z_\sigma, t + \Delta t)$ and $g_\alpha^Z(Z_\sigma, t)$ are uncorrelated. In order to determine how the distribution function evolves from any time t to $t + \Delta t$, we may then take $F^Z(Z_\sigma, t) = \langle F^Z(Z_\sigma, t)\rangle$ for all members of the ensemble since $F^Z(Z_\sigma, t + \Delta t)$ is uncorrelated with any particular realization of the g_α. Under these assumptions, if we then average (9) over ensembles and keep terms only to lowest order in Δt, the Fokker–Planck equation results:

$$\frac{\partial \langle F^Z\rangle}{\partial t} + G_\mu \frac{\partial \langle F^Z\rangle}{\partial Z_\mu} = -\frac{\partial}{\partial Z_\mu} \times \left(\left\langle \frac{\partial Z_\mu}{\partial t} \right\rangle_c \langle F^Z\rangle\right) + \frac{1}{2}\frac{\partial^2}{\partial Z_\mu \partial Z_\nu}\left(\left\langle \frac{\partial Z_\mu \partial Z_\nu}{\partial t} \right\rangle_c \langle F^Z\rangle\right), \tag{12}$$

where we have denoted

$$\begin{aligned} \left\langle \frac{\Delta Z_\mu}{\Delta t} \right\rangle_c &= \left\langle \frac{\partial Z_\mu}{\partial t} \right\rangle_c + O(\Delta t), \\ \left\langle \frac{\Delta Z_\mu \Delta Z_\nu}{\Delta t} \right\rangle_c &= \left\langle \frac{\partial Z_\mu \partial Z_\nu}{\partial t} \right\rangle_c + O(\Delta t). \end{aligned}$$

In retaining only lowest order terms in Δt, we may then

write

$$\left\langle \frac{\partial Z_\mu}{\partial t} \right\rangle_c = \lim_{"\Delta t \to 0"} \left\langle \frac{\Delta Z_\mu}{\Delta t} \right\rangle_c,$$

$$\left\langle \frac{\partial Z_\mu \partial Z_\nu}{\partial t} \right\rangle = \lim_{"\Delta t \to 0"} \left\langle \frac{\Delta Z_\mu \Delta Z_\nu}{\Delta t} \right\rangle_c. \tag{13}$$

Equation (12) governs the evolution of $\langle F^z \rangle$ from time t to $t + \Delta t$ where $\Delta t \gg \tau_c$. However, no restriction was made on the particular time t chosen; thus (12) is valid at all times.

Up to this point we have not done anything new. However, before continuing, let us look at our assumptions a little more closely and see what conditions must be met in order for these assumptions to be valid. First of all, we have assumed that some $\Delta t \gg \tau_c$ is a good expansion parameter in the sense that higher order terms in ΔZ_μ in Eq. (12) may be dropped; since ΔZ_μ depends on both G_μ^Z and g_μ^Z through Eq. (2), this requires

$$\frac{dG_\mu}{dt} \Delta t^2 \ll G_\mu \Delta t \ll L_\mu, \tag{14}$$

where L_μ is a typical "length" scale for variation of $\langle F^z \rangle$ with the coordinate Z_μ. This is then a restriction on the size of G_μ^Z in terms of parameters involving the fluctuations, since the "collision" or coherence time τ_c gives a lower bound for Δt and, thus, $L_\mu / \Delta t$ gives an upper bound for G_μ^Z. In the case where g_μ^Z just describes binary collisions between particles in laboratory plasmas, this restriction on G_μ^Z is not too severe, since collision times in such cases are sufficiently short and collisions sufficiently frequent that (14) may be satisfied for most external field configurations of interest. However, in turbulent plasmas, the scattering process is typically much slower and the restriction on G_μ^Z demanded by (14) eliminates all but very weak external fields from our consideration.

Another problem is the calculation of $\langle \partial Z_\mu / \partial t \rangle_c$ and $\langle \partial Z_\mu \partial Z_\nu / \partial t \rangle_c$ given in (13). How does the presence of G_μ^Z effect the calculation of these terms? Again, in the binary collision case, this question is easily resolved; collisions are typically of sufficiently short range that the interparticle forces completely dominate the external forces during the short time the collision occurs, and the dynamics of a single collision can be analyzed as if the external forces were not present. In a weakly turbulent plasma, however, we typically have $g_\mu^Z \ll G_\mu^Z$ at all points in phase space and time, so that it is clear that the presence of an external field will have a profound effect on the collision process.

In light of this discussion it becomes obvious that for application to the general problem involving a turbulent plasma, (12) is unsatisfactory as it stands. An outstanding example of this difficulty is a homogeneous plasma in a steady magnetic field taken along the z-axis but also subject to a weak fluctuating electromagnetic field. If the electron gyrofrequency, Ω_0, associated with the steady magnetic field is such that $\Omega_0 \tau_c \gg 1$, the first inequality of (14) is not satisfied for $Z_\mu = p_x$, since $|\Delta t^2 \, dG_\mu^z/dt| \approx |\Delta t^2 \, d^2 p_x/dt^2| \approx |-\Omega_0^2 p_x \Delta t^2| \gg |\Omega_0 p_y \Delta t| = G_\mu^z \Delta t$ for most of a particle's orbit. It is also clear that the external magnetic field will have to be considered in the calculation of the collision terms $\langle \partial Z_\mu / \partial t \rangle_c$ and $\langle \partial Z_\mu \partial Z_\nu / \partial t \rangle_c$. This particular case has been considered by Manheimer,[16] who succeeded in finding a suitable Fokker–Planck-type equation for this problem by parametrizing the distribution function in terms of unperturbed particle orbits. The method used by Manheimer is just a special case of the approach we develop in the next section.

III. EXTENDING THE FOKKER-PLANCK EQUATION TO MORE GENERAL CASES

Our aim is to modify Eq. (12) so that any external field configuration is allowed. To do this let us look at a generalized set of coordinates χ_μ related to Z_μ by a transformation of the form

$$\chi_\mu = \chi_\mu(Z_\sigma, t), \tag{15}$$

i.e., a transformation which, in general, has explicit time dependence.

Assuming we know the equation of motion (2) in the Z-system, we may find the equation of motion in the χ-system by transforming (2) according to (15). We find that χ_μ obeys the equation of motion

$$\frac{d\chi_\mu}{dt} = \frac{\partial \chi_\mu}{dt} + \frac{dZ_\sigma}{dt} \frac{\partial \chi_\mu}{\partial Z_\sigma} = \frac{\partial \chi_\mu}{\partial t} + G_\sigma{}^z \frac{\partial \chi_\mu}{\partial Z_\sigma} + g_\sigma{}^z \frac{\partial \chi_\mu}{\partial Z_\sigma}; \tag{16}$$

if the transformation (15) does not depend on the stochastic fields, we may separate (16) into two parts in analogy with (2) to find

$$\frac{d\chi_\mu}{dt} = G_\mu(\chi_\sigma, t) + g_\mu(\chi_\sigma, t), \tag{17}$$

where

$$G_\mu = \frac{\partial \chi_\mu}{\partial t} + G_\sigma{}^z \frac{\partial \chi_\mu}{\partial Z_\sigma} \quad \text{and} \quad g_\mu = g_\sigma{}^z \frac{\partial \chi_\mu}{\partial Z_\sigma} \tag{18}$$

describe the respective effects of the external and stochastic forces, as before. Defining the quantities

$$F(\chi_\mu, t) = F^z(Z_\mu, t),$$
$$h = \frac{\partial(Z_1, \cdots, Z_6)}{\partial(\chi_1, \cdots, \chi_6)}, \tag{19}$$

we may transform (6) to read

$$h(\chi_\mu, t) F(\chi_\mu, t) d^6\chi = h'(\chi_\mu', t) F(\chi_\mu', t') d^6\chi', \tag{20}$$

where, because of the explicit time dependence of the transformation (15), we have allowed h to depend explicitly on time as well. χ_μ' is point in the χ-system at time t' of a particle whose orbit passes through χ_μ at time t and which obeys the equation of motion (17).

To find an equation for $F(\chi_\mu, t)$ from (20), we may use the same procedure employed in Sec. II for F^z starting with (6). All that is necessary is to replace Z_μ by χ_μ and F^z by hF; the rest of the argument is identical. Hence our equation for hF corresponding to (9) is

$$\frac{\Delta(hF)}{\Delta t} = -\frac{\partial}{\partial \chi_\mu} \left(\frac{\Delta \chi_\mu}{\Delta t} hF \right) + \frac{1}{2} \frac{\partial^2}{\partial \chi_\mu \partial \chi_\nu} \left(\frac{\Delta \chi_\mu \Delta \chi_\nu}{\Delta t} hF \right) - \cdots. \tag{21}$$

In the Appendix we show that we can always find a choice of coordinates $\chi_\mu = Y_\mu$ related to Z_μ by a trans-

formation of the form (15) which has the property $G_\mu = 0$. Making this choice for χ_μ in (21), we see from (17) and (18) that ΔY_μ depends only on $\hat{g}_\mu$, where $\hat{g}_\mu = g^Z_0(\partial Y_\mu/\partial Z_\alpha)$. All explicit dependence on the external field configuration has been removed. Letting $\hat{F}(Y_\sigma, t)$ and $\hat{h}$ be defined by (19) (with χ_μ replaced by Y_μ) and noting that $(\partial/\partial t)\hat{h}(Y_\sigma, t) = 0$, we find that in this special coordinate system the distribution function evolves according to

$$\frac{\Delta\hat{F}}{\Delta t} = -\frac{1}{\hat{h}}\frac{\partial}{\partial Y_\alpha}\left(\hat{h}\frac{\Delta Y_\alpha}{\Delta t}\hat{F}\right) + \frac{1}{2\hat{h}}\frac{\partial^2}{\partial Y_\alpha \partial Y_\beta}\left(\hat{h}\frac{\Delta Y_\alpha \Delta Y_\beta}{\Delta t}\hat{F}\right) - \cdots. \tag{22}$$

Now, for the coordinate system Y_μ, we find from (17) that ΔY_α is found by integrating the equation of motion

$$\frac{d}{d\Delta t}[\Delta Y_\alpha(Y_\mu, t; \Delta t)] = \hat{g}_\alpha[Y_\mu + \Delta Y_\mu(Y_\sigma, t; \Delta t), t + \Delta t]. \tag{23}$$

If we take $\langle \hat{g}_\alpha(Y_\mu, t)\rangle = 0$ we can evaluate $\langle \Delta Y_\alpha\rangle$ by expanding (23) about the point Y_μ and integrating:

$$\begin{aligned}\langle \Delta Y_\alpha\rangle &= \int_0^{\Delta t} dt' \left(\langle \hat{g}_\alpha(Y_\mu, t + t')\rangle + \langle \Delta Y_\beta(Y_\mu, t; t')\frac{\partial}{\partial Y_\beta}\hat{g}_\alpha(Y_\mu, t + t')\rangle + O(g^3)\right)\\ &= \int_0^{\Delta t} dt' \int_0^{\Delta t} dt'' \langle \hat{g}_\beta(Y_\mu, t + t'') \times \frac{\partial}{\partial Y_\beta}\hat{g}_\alpha(Y_\mu, t + t')\rangle + O(g^3).\end{aligned} \tag{24}$$

Similarly,

$$\langle \Delta Y_\alpha \Delta Y_\beta\rangle = \int_0^{\Delta t} dt' \int_0^{\Delta t} dt'' \times \langle \hat{g}_\alpha(Y_\mu, t + t')\hat{g}_\beta(Y_\mu, t + t'')\rangle + O(g^3). \tag{25}$$

As in Sec. II, we seek an equation for $\langle \hat{F}\rangle$. However, instead of using Δt as an expansion parameter, we assume $\hat{g}_\mu$ is small enough to justify a perturbation expansion in orders of $\hat{g}$. Averaging (22) over ensembles and again assuming a Markov process to allow us to take $\langle \hat{F}(Y_\mu, t)\rangle = \hat{F}(Y_\mu, t)$, we find

$$\frac{\Delta\langle \hat{F}\rangle}{\Delta t} = -\frac{1}{\hat{h}}\frac{\partial}{\partial Y_\alpha}\left(\hat{h}\left\langle\frac{\Delta Y_\alpha}{\Delta t}\right\rangle\langle \hat{F}\rangle\right) + \frac{1}{2\hat{h}}\frac{\partial^2}{\partial Y_\alpha \partial Y_\beta}\left(\hat{h}\left\langle\frac{\Delta Y_\alpha \Delta Y_\beta}{\Delta t}\right\rangle\langle \hat{F}\rangle\right) + O(g^3). \tag{26}$$

But from (24) and (25), we note that (26) may be written as

$$\frac{1}{\Delta t}\int_0^{\Delta t} dt'\left(\frac{\partial\langle \hat{F}(t + t')\rangle}{dt'} + \frac{1}{\hat{h}}\frac{\partial}{\partial Y_\alpha}[h\hat{\Gamma}_\alpha(t + t')\langle \hat{F}(t)\rangle] - \frac{1}{2\hat{h}}\frac{\partial^2}{\partial Y_\alpha \partial Y_\beta}[\hat{h}\hat{\Lambda}_{\alpha\beta}(t + t')\langle \hat{F}(t)\rangle]\right) = O(g^3), \tag{27}$$

where we have suppressed the dependence of $\hat{F}$ on Y_σ and

$$\hat{\Gamma}_\alpha(t + t') = \int_0^{t'} dt''\left\langle \hat{g}_\beta(Y_\mu, t + t'')\frac{\partial}{\partial Y_\beta}\hat{g}_\alpha(Y_\mu, t + t')\right\rangle + O(g^3) \tag{28}$$

and

$$\Lambda_{\alpha\beta}(t + t') = 2\int_0^{t'} dt''\langle \hat{g}_\beta(Y_\mu, t + t'')\hat{g}_\alpha(Y_\mu, t + t')\rangle + O(g^3). \tag{29}$$

$\hat{\Gamma}_\alpha$ and $\hat{\Lambda}_{\alpha\beta}$ are found from (24) and (25) and the equality of symmetric parts (29) and (25) [the antisymmetric part does not contribute to (26)]. But (26), (25), and (24) also give $\langle \hat{F}(t + t')\rangle - \langle \hat{F}(t)\rangle = O(g^2)$ for $t' \le \Delta t$. Thus, we may replace $\langle \hat{F}(t)\rangle$ by $\langle \hat{F}(t + t')\rangle$ in (27) and make an error of only order g^3. Thus, (27) becomes

$$\frac{1}{\Delta t}\int_0^{\Delta t} dt'\left(\frac{\partial\langle \hat{F}\rangle}{\partial t} + \frac{1}{\hat{h}}\frac{\partial}{\partial Y_\alpha}[\hat{h}\hat{\Gamma}_\alpha\langle \hat{F}\rangle] - \frac{1}{2\hat{h}}\frac{\partial^2}{\partial Y_\alpha \partial Y_\beta}[\hat{h}\hat{\Gamma}_{\alpha\beta}\langle \hat{F}\rangle]\right)\Bigg|_{t+t'} = O(g^3). \tag{30}$$

But (30) must be valid for a range of $\Delta t \gg \tau_c$, thus the integrand must satisfy

$$\frac{\partial\langle \hat{F}\rangle}{\partial t}\Bigg|_{t+t'} = -\frac{1}{\hat{h}}\frac{\partial}{\partial Y_\alpha}[\hat{h}\hat{\Gamma}_\alpha(t + t')\langle \hat{F}(t + t')\rangle] + \frac{1}{2\hat{h}}\frac{\partial^2}{\partial Y_\alpha \partial Y_\beta}[\hat{h}\hat{\Lambda}_{\alpha\beta}\langle \hat{F}(t + t')\rangle] + O(g^3) \tag{31}$$

for all times $t + t'$. Now if we replace $t + t'$ by t and $\Lambda_{\alpha\beta}$ by $\hat{\Delta}_{\alpha\beta} = \frac{1}{2}\hat{\Lambda}_{\alpha\beta} + \frac{1}{2}\hat{\Lambda}_{\beta\alpha}$ in (31), we find that $\hat{\Gamma}_\alpha$ and $\hat{\Delta}_{\alpha\beta}$ may be found from $\langle \Delta Y_\alpha/\Delta t\rangle$ and $\langle \Delta Y_\alpha \Delta Y_\beta/\Delta t\rangle$ [(24) and (25)] merely by eliminating the operation $(1/\Delta t)\int_0^{\Delta t} dt'$. Thus

$$\hat{\Gamma}_\alpha = \lim_{\text{“}\Delta t\to 0\text{”}}\left\langle\frac{\Delta Y_\alpha}{\Delta t}\right\rangle = \left\langle\frac{\partial Y_\alpha}{\partial t}\right\rangle,$$

$$\hat{\Delta}_{\alpha\beta} = \lim_{\text{“}\Delta t\to 0\text{”}}\left\langle\frac{\Delta Y_\alpha \Delta Y_\beta}{\Delta t}\right\rangle = \left\langle\frac{\partial Y_\alpha \partial Y_\beta}{\partial t}\right\rangle.$$

We have, therefore, shown that we may write the equation for $\langle \hat{F}\rangle$ in the coordinate system Y_μ as

$$\frac{\partial\langle \hat{F}\rangle}{\partial t} = -\frac{1}{\hat{h}}\frac{\partial}{\partial Y_\mu}\left(\hat{h}\left\langle\frac{\partial Y_\mu}{\partial t}\right\rangle\langle \hat{F}\rangle\right) + \frac{1}{2\hat{h}}\frac{\partial^2}{\partial Y_\mu \partial Y_\nu}\left(\hat{h}\left\langle\frac{\partial Y_\mu \partial Y_\nu}{\partial t}\right\rangle\langle \hat{F}\rangle\right); \tag{32}$$

the above equation for $\langle \hat{F}\rangle$ is the Fokker-Planck equation as it generally appears in the literature for a coordinate system in which $G_\mu = 0$.

In a system where $G_\mu \neq 0$ the equation for $\langle \hat{F}\rangle$ is given by transforming (32) to the general system described by χ_μ:

$$\frac{\partial\langle F\rangle}{\partial t} + G_\mu\frac{\partial\langle F\rangle}{\partial \chi_\mu} = -\frac{1}{\hat{h}}\frac{\partial}{\partial \chi_\mu}(h\Gamma_\mu\langle F\rangle) + \frac{1}{2h}\frac{\partial^2}{\partial \chi_\mu \partial \chi_\nu}(h\Delta_{\mu\nu}\langle F\rangle), \tag{33}$$

where

$$\Delta_{\mu\nu} = \tfrac{1}{2}\Lambda_{\mu\nu} + \tfrac{1}{2}\Lambda_{\nu\mu}, \tag{34}$$

$$\Gamma_\mu = \hat{\Gamma}_\alpha\frac{\partial\chi_\mu}{\partial Y_\alpha} + \frac{1}{2}\hat{\Lambda}_{\mu\nu}\frac{\partial^2\chi_\mu}{\partial Y_\alpha \partial Y_\beta} = \int_{-\infty}^{t} dt'\left\langle g_\lambda(\chi'_\sigma, t')\frac{\partial}{\partial \chi_{\lambda'}}g_\mu(\chi_\sigma, t)\right\rangle \tag{35}$$

and

$$\Lambda_{\mu\nu} = \hat{\Lambda}_{\alpha\beta} \frac{\partial \chi_\mu}{\partial Y_\alpha} \frac{\partial \chi_\nu}{\partial Y_\beta} = 2\int_{-\infty}^{t} dt' \langle g_\lambda(\chi'_\sigma, t') g_\mu(\chi_\sigma, t)\rangle \frac{\partial \chi_\nu}{\partial \chi_\lambda'}. \tag{36}$$

In (35) and (36) $\chi'_\sigma(t')$ is defined by (20), and the lower limit of integration has been extended to $-\infty$ since the actual value of this limit is $t - \Delta t$, where $\Delta t \gg \tau_c$; the contribution from the extended range of integration is assumed to be negligible due to the smallness of $\langle g_\mu(\chi_\sigma, t) g_\nu(\chi'_\sigma, t')\rangle$ for $|t - t'| \gg \tau_c$.

Note that (33), (34), (35), and (36) give an equation for $\langle F\rangle$ completely in terms of quantities which refer to the coordinate system χ_μ. No reference to the special system used in deriving the equation remains. Thus, we now have a prescription for finding the equation governing $\langle F\rangle$ in any coordinate system χ_μ. All that is necessary is to write down (33)-(36) for that system and to find χ_μ as a function of χ'_σ, where χ_μ and χ'_σ are points on the same unperturbed orbit corresponding to t and $t + t'$, respectively.

We should also note that the quantities Γ_μ and $\Delta_{\mu\nu}$ cannot in general be interpreted as Fokker-Planck coefficients in the usual sense; e.g., Γ_μ is not $\lim_{\text{“}\Delta t \to 0\text{”}} \langle \Delta\chi_\mu / \Delta t\rangle$, with $\Delta\chi_\mu$ being the displacement a particle at χ_μ at time t undergoes in the time interval $(t, t + \Delta t)$ due to the presence of g_μ. Hence we call these quantities "generalized Fokker-Planck coefficients."

If $G_\mu^Z = 0$, then the special system Y_μ and the system Z_μ are identical and (33) takes on the form of the usual Fokker-Planck equation in the Z system derived elsewhere; in this case, we see from (34), (35), and (36) (with $\chi_\mu = Y_\mu = Z_\mu$ for this special case) that the quantities Γ_μ^Z and $\Delta_{\mu\nu}^Z$ may be interpreted as $\langle \Delta Z_\mu / \Delta t\rangle$ and $\langle \Delta Z_\mu \Delta Z_\nu / \Delta t\rangle$. The spatially homogeneous plasma with no external forces is such a case since there the only coordinates of interest are momenta, for which $G_\mu^Z = 0$. In Manheimer's extension of this case to include a uniform external magnetic field he proceeded via the $G_\mu = 0$ coordinate system for that case—i.e., he used a special case of our more general treatment.

From this discussion we see that we have successfully extended the Fokker-Planck treatment to turbulent plasma in an arbitrary external field configuration. Our choice for phase-space coordinates to describe the problem is unrestricted. We have made one assumption in addition to those usually made in deriving the Fokker-Planck equation, and that is that the turbulence is weak; this assumption is necessitated by the increased time-scale of the scattering process in a turbulence-dominated plasma as opposed to a collision dominated plasma.

We now turn our attention to an alternative approach to this general problem.

IV. QUASILINEAR APPROACH

In this section we treat the same general problem but start from the continuity equation. This approach is nearly identical to that devised by Hall and Sturrock[10]; however, they failed to keep terms which were unimportant in the problems they treated but which can make significant contributions in the general case.[17]

Our choice of coordinate systems in the generalized set χ_μ defined in (15). The continuity equation (5) transformed to this coordinate system is just

$$\frac{\partial F}{\partial t} + G_\mu \frac{\partial F}{\partial \chi_\mu} = -\frac{1}{h} \frac{\partial}{\partial \chi_\mu} (h g_\mu F). \tag{37}$$

We consider first an ensemble of possible realizations of the fluctuating fields g_μ. Each member of the ensemble will have a distribution function F obeying Eq. (37). We look for an equation for $\langle F\rangle$, the distribution function averaged over the ensemble, in terms of ensemble averages of functionals of g_μ. Taking the ensemble average of (37), we find that $\langle F\rangle$ satisfies

$$\frac{\partial \langle F\rangle}{\partial t} + G_\mu \frac{\partial \langle F\rangle}{\partial \chi_\mu} = -\frac{1}{h} \frac{\partial}{\partial \chi_\mu} (h \langle g_\mu F\rangle). \tag{38}$$

Now we subtract (38) from (37), letting $\delta F = F - \langle F\rangle$. This procedure yields the following equation for δF:

$$\frac{\partial \delta F}{\partial t} + G_\mu \frac{\partial \delta F}{\partial \chi_\mu} = -\frac{1}{h} \frac{\partial}{\partial \chi_\mu} [h(g_\mu F - \langle g_\mu F\rangle)]. \tag{39}$$

For $\langle g_\mu\rangle = 0$, (38) and (39) become

$$\frac{\partial \langle F\rangle}{\partial t} + G_\mu \frac{\partial \langle F\rangle}{\partial \chi_\mu} = -\frac{1}{h} \frac{\partial}{\partial \chi_\mu} (h \langle g_\mu \delta F\rangle), \tag{40}$$

$$\frac{\partial \delta F}{\partial t} + G_\mu \frac{\partial \delta F}{\partial \chi_\mu} = -\frac{1}{h} \frac{\partial}{\partial \chi_\mu} (h g_\mu \langle F\rangle) - \frac{1}{h} \frac{\partial}{\partial \chi_\mu} [h(g_\mu \delta F - \langle g_\mu \delta F\rangle)]. \tag{41}$$

Equations (40) and (41) may be formally iterated to give δF as a perturbation expansion in orders of g; however, we are interested only in the lowest order terms. If we know the characteristics of the operator $\frac{\partial}{\partial t} + G_\mu \frac{\partial}{\partial \chi_\mu}$ (i.e., the unperturbed orbits), the solution of (41) for δF may then be written, to lowest order, as

$$\delta F(\chi_\sigma, t) = -\int_{t_0}^{t} dt' \frac{1}{h'} \frac{\partial}{\partial \chi_\mu'} \times \{h' g_\mu[\chi'_\sigma(t'), t'] \langle F[\chi'_\sigma(t'), t']\rangle\} + O(g^2), \tag{42}$$

where $\chi'_\sigma(t')$ is the position at time t' on the unperturbed orbit passing through χ_σ at time t and t_0 is an initial time when $F = \langle F\rangle$ for all members of the ensemble. The existence of such a time t_0 is assured if we again assume that we are considering a Markov process, as in Secs. II and III.

We now suppose that g_μ is sufficiently small and that there exists a time interval $t - t_0 \gg \tau_c$ for which the perturbation expansion (42) converges rapidly. In this event (42) may be inserted into (41), producing the following equation for $\langle F\rangle$:

$$\frac{\partial \langle F\rangle}{\partial t} + G_\mu \frac{\partial \langle F\rangle}{\partial \chi_\mu} = \frac{1}{h} \frac{\partial}{\partial \chi_\mu} \left(h \int_{t_0}^{t} dt' \left\langle g_\mu(\chi_\sigma, t) \frac{1}{h'} \frac{\partial}{\partial \chi_\mu'} \{h' g_\nu[\chi'_\sigma(t'), t'] \langle F[\chi'_\sigma(t'), t']\rangle\} \right\rangle + O(g^3) \right). \tag{43}$$

But (43) states that $\langle F[\chi'_\sigma(t'), t']\rangle - \langle F[\chi_\sigma, t]\rangle = O(g^2)$, since the only change in $\langle F\rangle$ along the unperturbed orbits is due to the right-hand side of that equation; thus, we may replace $\langle F[\chi'_\sigma(t'), t']\rangle$ by $\langle F(\chi_\sigma, t)\rangle$ in (43), making an error only of order g^4. Performing this operation and using the fact that $(1/h')(\partial/\partial\chi_\nu')(h' A'_\nu) = (1/h)(\partial/\partial\chi_\mu)[h A'_\nu (\partial\chi_\mu/\partial\chi_\nu')]$ for any A'_ν, we can put the equation for F into the more suggestive form:

$$\frac{\partial\langle F\rangle}{\partial t} + G_\mu \frac{\partial\langle F\rangle}{\partial \chi_\mu} = -\frac{1}{h}\frac{\partial}{\partial \chi_\mu}(h\Gamma_\mu\langle F\rangle) + \frac{1}{2h}\frac{\partial^2}{\partial\chi_\mu\partial\chi_\nu}(h\Lambda_{\mu\nu}\langle F\rangle), \quad (44)$$

where

$$\Gamma_\mu = \int_{t_0}^{t} dt' \left\langle g_\lambda[\chi'_\sigma(t'), t'] \frac{\partial}{\partial \chi'_\lambda} g_\mu(\chi_\sigma, t) \right\rangle \quad (45)$$

and

$$\Lambda_{\mu\nu} = 2\int_{t_0}^{t} dt' \langle g_\mu[\chi_\sigma, t] g_\lambda[\chi'_\sigma(t'), t']\rangle \frac{\partial \chi_\nu}{\partial \chi'_\lambda}. \quad (46)$$

For $t - t_0 \gg \tau_c$, $\langle g_\mu(\chi_\sigma, t) g_\nu(\chi'_\sigma, t')\rangle$ becomes negligibly small for $t' < t_0$; thus, we may extend the lower limit of integration to $-\infty$, incurring negligible error. When this is done, we see that the set of Eqs. (44)-(46) is identical to the set (33)-(36) derived from the Fokker-Planck approach. The derivation of (44) was much more straightforward than that of (33), but the same assumptions were involved in both cases. It is therefore reassuring that both produce identical results.

V. SPECIALIZATION TO ELECTROMAGNETIC FLUCTUATIONS

The equation we have derived, (33) or (44), is valid for any g_μ but undergoes significant simplification when the g_μ are due to electromagnetic fields. In such a case g_μ^Z satisfies

$$\frac{\partial g_\mu^Z}{\partial Z_\mu} = 0,$$

so that in the system χ_μ, g_μ satisfies

$$\frac{1}{h}\frac{\partial}{\partial \chi_\mu}(h g_\mu) = 0. \quad (47)$$

From (47) we may then write (45) and (46) as

$$\Gamma_\mu = \frac{1}{h}\frac{\partial}{\partial \chi_\nu}(h D_{\mu\nu}), \quad (48)$$

$$\Delta_{\mu\nu} = 2D_{\mu\nu}, \quad (49)$$

where

$$D_{\mu\nu} = \int_{t_0}^{t} dt' \langle g_\mu(\chi_\sigma, t) g_\lambda[\chi'_\sigma(t'), t']\rangle \frac{\partial \chi_\nu}{\partial \chi'_\lambda}. \quad (50)$$

Thus, (44) may be written

$$\frac{\partial\langle F\rangle}{\partial t} + G_\mu \frac{\partial\langle F\rangle}{\partial \chi_\mu} = \frac{1}{h}\frac{\partial}{\partial \chi_\mu}\left(h D_{\mu\nu}\frac{\partial\langle F\rangle}{\partial \chi_\nu}\right). \quad (51)$$

This equation resembles closely the equation derived by Hall and Sturrock; however, (50) differs from their formula for $D_{\mu\nu}$ by the factor $\partial\chi_\nu/\partial\chi'_\lambda$. Their formula is formally valid only in a coordinate system where $G_\mu = 0$, but they erroneously applied it to the general coordinate system χ_μ.

We note in passing that (51) is a generalized diffusion equation with a "diffusion tensor" $D_{\mu\nu}$. This equation is a generalization of the quasilinear diffusion equation which appears extensively in the literature.[11-15]

VI. CONCLUSIONS

We have developed a formalism capable of treating a wider class of plasma turbulence problems than was heretofore possible. This formalism can be developed from either of two approaches, both giving the same results; this fact bolsters our confidence in the validity of the formalism. This formalism may be applied to such problems of astrophysical interest as heat conduction in the solar wind[4] and particle acceleration in solar flares.[18]

ACKNOWLEDGMENTS

The author would like to express his appreciation for many helpful conversations with D. E. Hall and E. Tademaru. This work was supported in part by the Office of Naval Research through Contract N0014-67-0112-0062 and by the National Aeronautics and Space Administration through Grant NGL 05-020-272.

APPENDIX: THE EXISTENCE OF THE COORDINATE SYSTEM Y_μ

That there exists a coordinate system $\chi_\mu = Y_\mu$, related to Z_μ by a transformation of the form (15), such that $G_\mu = 0$ is easily seen from the following argument: Consider the "unperturbed" equation of motion in the Z system [i.e., (2) with $g_\mu^Z = 0$],

$$\frac{dZ_\mu}{dt} = G_\mu^Z(Z_\sigma, t). \quad (A1)$$

For any given G_μ^Z, (A1) has a solution of the form

$$Z_\mu = \xi_\mu(Y_\sigma, t), \quad (A2)$$

where Y_σ ($\sigma = 1, 2, \cdots, 6$) are six integration constants and

$$\frac{\partial \xi_\mu}{\partial t} = G_\mu^Z. \quad (A3)$$

For each particular choice of Y_μ, $\xi_\mu(t)$ traces out a particular "unperturbed orbit" in phase space. Furthermore, through each point Z_μ there passes one and only one of these unperturbed orbits; hence, there is a one-to-one correspondence between Z_μ and Y_μ. Therefore, Eq. (A2) may be interpreted as implicitly defining new variables Y_μ as functions of Z_μ and t, i.e., relating Y_μ to Z_μ by a transformation of the form (15). Thus, Y_μ is a system of coordinates belonging to the set of systems denoted by χ_μ; we may then use (18) to find G_μ^Y or

$$G_\mu^Y = \frac{\partial Y_\mu}{\partial t} + G_\sigma^Z \frac{\partial Y_\mu}{\partial Z_\sigma} = 0, \quad (A4)$$

since Y_μ, as defined by (A2) and (A3), is constant along the characteristics of the operator $(\partial/\partial t) + G_\mu^Z(\partial/\partial Z_\mu)$. The parameters Y_μ thus determined then form a suitable coordinate system in which $G_\mu = 0$.

[1] E. Fermi, Phys. Rev. **75**, 1169 (1949).
[2] V. L. Ginzburg and S. I. Syrovatskii, *The Origin of Cosmic Rays* (MacMillan, New York, 1964).
[3] J. R. Jokipii, Astrophys. J. **146**, 480 (1966).
[4] C. E. Newman and P. A. Sturrock, Phys. Fluids **12**, 2553 (1969).
[5] O. P. Manley and S. Olbert, Astrophys. J. **157**, 223 (1969).
[6] L. D. Smullin and W. D. Getty, Phys. Rev. Lett. **9**, 3 (1969).

[7]P. A. Sturrock, J. Math. Phys. **1,** 405 (1960).
[8]P. A. Sturrock, Phys. Rev. **141,** 186 (1966).
[9]S. Chandrasekhar, Rev. Mod. Phys. **15,** 31 (1943).
[10]D. E. Hall and P. A. Sturrock, Phys. Fluids **10,** 2620 (1967).
[11]C. F. Kennel and F. Engelmann, Phys. Fluids **9,** 2377 (1966).
[12]D. B. Melrose, Astrophys. Space Sci. **2,** 171 (1968).
[13]V. N. Tsytovich, *Nonlinear Effects in Plasma* (Plenum, New York, 1970).
[14]A. L. Alexeev and Yu. P. Nikitin, Zh. Eksp. Teor. Fiz. **50,** 915 (1966) [Sov. Phys.-JETP **23,** 608 (1966)].
[15]E. G. Harris, *Advances in Plasma Physics* (Interscience, New York, 1969), Vol. 3.
[16]W. M. Manheimer, Phys. Fluids **12,** 901 (1969).
[17]D. E. Hall, private communication to E. Tademaru and C. E. Newman.
[18]M. Friedman, Phys. Rev. **182,** 1408 (1969).

Theory of elastic constants of heterogeneous media

J. Korringa*

Chevron Oil Field Research Company, La Habra, California 90631
(Received 28 July 1972)

The theory of elastic equilibrium of heterogeneous, macrohomogeneous bodies is cast in the form of a linear integral equation with the following properties: (1) The equation contains a constant elastic tensor which can be chosen freely to optimize any approximation method. (2) The singularity of the kernel is such that the integral must be interpreted as the principal value, thereby providing a clear separation between structure-dependent and structure-independent contributions. (3) The mean elastic tensor equals the above-mentioned constant tensor plus the space average of the dependent variable. (4) A term proportional to this average appears separately in the equation. (5) The integral equation facilitates the rederivation of several well-known approximations and opens new possibilities, of which some are discussed, in particular, a self-consistent imbedding method and an approximation based on correlation functions.

1. INTRODUCTION

The first theoretical estimates of the velocity of acoustic waves of long wavelength in heterogeneous elastic materials did not aim at great accuracy. Recently, theorists have searched for more refined methods, in order to improve our understanding of the relation between the mean elastic constants and the composition and texture of the medium. This paper contributes to this field in terms on an integral equation for the stress-tensor of a heterogeneous, macro-homogeneous material undergoing macro-uniform strain. Considering its simplicity and its usefulness for deriving known and new approximations, it is surprising that it has until now gone unreported.[1]

In the next section this equation is derived, on the basis of the classical theory of elasticity. Subsequent sections deal with approximate solutions.

2. DERIVATION OF THE FUNDAMENTAL EQUATION

Let the medium be given in terms of the elastic tensor $C(x) \equiv C_{ijkl}(x)$. To a displacement field $u(x) \equiv u_i(x)$ belongs a strain tensor

$$\epsilon(x) \equiv \epsilon_{ij}(x) = \tfrac{1}{2}(\partial_i u_j + \partial_j u_i). \tag{1}$$

The corresponding stress tensor $\sigma(x)$ is given by Hooke's law, viz.,

$$\sigma(x) = C(x)\epsilon(x) \equiv C_{ijkl}(x)\epsilon_{kl}(x). \tag{2}$$

The static equilibrium equation is

$$\mathfrak{F}_{ij}u_j(x) = 0, \tag{3}$$

where $\mathfrak{F}$ is the tensor operator

$$\mathfrak{F}_{il} = \partial_j C_{ijkl}(x)\partial_k. \tag{4}$$

Let V be a finite volume, bounded by the surface Σ, and let $u(x)$ be defined by (2) and (3) for $x \in V$, with the boundary condition

$$u(\hat{x}) = u^0(\hat{x}), \qquad \hat{x} \in \Sigma, \tag{5}$$

where

$$u_i^0(x) = \epsilon_{ij}^0 x_j, \qquad \epsilon_{ij}^0 = \text{const.} \tag{6}$$

In terms of $\epsilon(x)$ and $\sigma(x)$ thus defined, the mean elastic tensor C^* is, as is well known, given by[2]

$$\langle\sigma\rangle \equiv \langle C\epsilon\rangle = C^*\langle\epsilon\rangle. \tag{7}$$

Here the symbol $\langle\,\rangle$ stands for

$$\langle A\rangle = \lim_{V\to\infty} V^{-1}\int_V A(x)d^3x. \tag{8}$$

This boundary value problem will now be transformed by means of Green's theorem for an auxiliary differential operator $\mathfrak{F}^0$, defined as in Eq. (4), but with a constant tensor[3] C^0:

$$\mathfrak{F}^0_{il} = \partial_j C^0_{ijkl}\partial_k. \tag{9}$$

One has, for any two fields $U_i^a(x)$, $a = 1, 2$,

$$\int_V [U_i^2(x')\mathfrak{F}^0_{ij}U_j^1(x') - U_i^1(x')\mathfrak{F}^0_{ij}U_j^2(x')]d^3x' = \int_\Sigma [U_i^2(\hat{x}')\tau_i^1(\hat{x}') - U_i^1(\hat{x}')\tau_i^2(\hat{x}')]d^2x', \tag{10}$$

where

$$\tau_i^a(\hat{x}) = C^0_{ijmn}(\hat{x})n_j(\hat{x})\epsilon^a_{mn}(\hat{x}), \qquad a = 1, 2, \tag{11}$$

$n(\hat{x})$ being the outward normal of Σ in the point $\hat{x}$. Equation (10) is used with

$$U_i^1(x') = u_i(x') - u_i^0(x'), \qquad U_i^2(x') = G_{ik}(x - x'), \quad k = 1, 2, 3 \text{ resp.}, \tag{12}$$

where G is the Green's function of $\mathfrak{F}^0$:

$$\mathfrak{F}^0_{ij}G_{jk}(x - x') = -\delta_{ik}\delta^3(x - x'), \qquad G \sim |x - x'|^{-1}, \tag{13}$$

i.e., the static displacement produced in the medium C^0 by a point force under the boundary conditions (5). This gives, using (5),

$$u_k(x) - u_k^0(x) + \int_V G_{ik}(x - x')\mathfrak{F}^0_{ij}[u_j(x') - u_j^0(x')]d^3x' = \int_\Sigma G_{ik}(x - \hat{x}')C^0_{ijmn}n_j(\hat{x}')[\epsilon_{mn}(\hat{x}') - \epsilon^0_{mn}]d^2\hat{x}'. \tag{14}$$

In the integral at left one can substitute (3) and trivially $\mathfrak{F}^0_{ij}u_j^0 = 0$ to get

$$\mathfrak{F}^0_{ij}(u_j - u_j^0) = (\mathfrak{F}^0_{ij} - \mathfrak{F}_{ij})u_j. \tag{15}$$

Introducing the symmetric tensor[4]

$$p_{ij}(x) = [C_{ijkl}(x) - C^0_{ijkl}]\epsilon_{kl}(x), \tag{16}$$

one has thus

$$\mathfrak{F}^0_{ij}(u_j - u_j^0) = \partial_j p_{ij}. \tag{17}$$

The volume integral in Eq. (14) will now be transformed by means of Gauss' theorem giving

$$\int_{V-v} G_{ik}(x-x')\mathfrak{F}^0_{ij}[u_j(x')-u^0_j(x')d^3x' = \int_{V-v} S_{ijk}(x-x')p_{ij}(x')d^3x' - \int_{\Sigma-\sigma} G_{ik}(x-\hat{x}')p_{ij}(\hat{x}')n_j(\hat{x}')d^2\hat{x}', \quad (18)$$

where the 3-index tensor S is defined by

$$S_{ijk} = \tfrac{1}{2}(\partial_j G_{ik} + \partial_i G_{jk}). \quad (19)$$

The argument of S in Eq. (18) is $(x-x')$, and in the volume integrals an infinitesimal sphere containing the singular point $x = x'$ has been excluded; σ is the surface of this sphere, n on σ and Σ is away from the singular point.

Next, the symmetrized derivative of Eq. (14) is formed. Owing to the exclusion of the singular point, this can be done under the integral sign. This gives

$$\epsilon_{kl}(x) - \epsilon^0_{kl} + \int_{V-v} T_{klij}(x-x')p_{ij}(x')d^3x' + \int_\sigma S_{kli}(x-\hat{x}')p_{ij}(\hat{x}')n_j(\hat{x}')d^2x' = \int_\Sigma S_{kli}(x-\hat{x}')[C^0_{ijmn}(\epsilon_{mn}(\hat{x}')-\epsilon^0_{mn}) + p_{ij}(\hat{x}')]n_j(\hat{x}')d^2\hat{x}', \quad (20)$$

where

$$T_{klij} = \tfrac{1}{2}(\partial_l S_{ijk} + \partial_k S_{ijl}). \quad (21)$$

In preparation of taking the limit $V \to \infty$, uniformly to all sides, one now subtracts the term $\int_\Sigma S\langle p\rangle_V n d^2\hat{x}'$ from both sides, where $\langle p\rangle_V$ is the average of p over V.[5] In the left-hand side this is transformed with Gauss' theorem into

$$\int_\Sigma S_{kli}(x-\hat{x}')\langle p_{ij}\rangle_V n_j(\hat{x}')d^2\hat{x}' = \int_{V-v} T_{klij}(x-x')\langle p_{ij}\rangle_V d^3x' + \int_\sigma S_{kli}(x-\hat{x}')\langle p_{ij}\rangle_V n_j(\hat{x}')d^2\hat{x}'. \quad (22)$$

From the so transformed Eq. (20) one verifies, by taking the average over V,

$$\langle\epsilon\rangle_V = \epsilon^0, \quad (23)$$

which is obvious from the definition of ϵ and the boundary conditions. One can now take the limit $V \to \infty$. As the medium is macrohomogeneous, one has

$$\lim_{V\to\infty}\int_\Sigma S_{kli}(x-\hat{x}')[C^0_{ijmn}(\epsilon_{mn}(\hat{x}')-\epsilon^0_{mn}) + p_{ij}(\hat{x}') - \langle p_{ij}\rangle_V]n_j(\hat{x}')d^2\hat{x}' = 0. \quad (24)$$

Next, taking the limit $v \to 0$ defines the volume integrals as the principal value and allows the integral over σ to be done. Defining a constant tensor Γ in terms of C^0 by

$$\Gamma_{ijkl} = \int_\sigma S_{ijk}(\hat{x})n_l(\hat{x})d^2\hat{x}, \quad (25)$$

one has, denoting $\langle p\rangle = \lim_{V\to\infty}\langle p\rangle_V$,

$$\lim_{\sigma\to 0}\int_\sigma S_{kli}(x-x')[p_{ij}(x') - \langle p_{ij}\rangle]n_j(x') = -\Gamma_{klij}[p_{ij}(x) - \langle p_{ij}\rangle]. \quad (26)$$

Thus one obtains, in tensor notation,

$$\epsilon(x) - \epsilon^0 + A(p(x) - \langle p\rangle) - \Gamma(p(x) - \langle p\rangle) = 0, \quad (27)$$

where the linear operator A is defined by the principal value integral

$$(Ap(x))_{kl} \equiv A_{klij}p_{ij}(x) = \int T_{klij}(x-x')p_{ij}(x')d^3x'. \quad (28)$$

From the definition of T one sees that A applied to a constant is zero, so that the first term in $\langle p\rangle$ of Eq. (27) can be dropped.

As a final transformation, one substitutes ϵ as a function of p from Eq. (16). Defining the reciprocal, B^{-1}, of a tensor B_{ijkl} by $B^{-1}_{ijkl}B_{klmn} = \tfrac{1}{2}(\delta_{im}\delta_{jn} + \delta_{in}\delta_{jm})$, one has

$$\{[C(x) - C_0]^{-1} - \Gamma\}p(x) + Ap(x) + \Gamma\langle p\rangle = \epsilon^0. \quad (29)$$

Defining

$$\varphi(x) = \{[C(x) - C^0]^{-1} - \Gamma\}^{-1} \quad (30)$$

and using the linearity of Eq. (29) to introduce a new variable $q(x)$ with

$$p(x) = q(x)\epsilon^0, \quad (31)$$

one finally has

$$q(x) + \varphi(x)Aq(x) + \varphi(x)(\Gamma\langle q\rangle - 1) = 0. \quad (32)$$

This is the desired integral equation. From its solution $q(x)$ one finds C^* by means of

$$C^* = C^0 + \langle q\rangle. \quad (33)$$

This is seen from Eqs. (7), (16), (23), and (31).

Equations (32) and (33) give a general and quite compact formulation of the problem. It is also flexible in that C^0 can still be selected freely. This choice affects the definition of q, A, φ, and Γ, but C^* does, of course, not depend on it.

A reasonable restriction is that C^0 reflects the rotational macro-symmetry of the medium. Specific choices will be discussed in connection with approximation methods.

3. A CRUDE APPROXIMATION

The simplest estimate of C^* is found from Eq. (32) by neglecting the term φAq. This gives

$$\langle q\rangle + \langle\varphi\rangle(\Gamma\langle q\rangle - 1) \approx 0, \quad (34)$$

i.e.,

$$C^* \approx C^0 + (1 + \langle\varphi\rangle\Gamma)^{-1}\langle\varphi\rangle. \quad (35)$$

This estimate depends on C^0, and two cases are known from the literature. Hashin and Shtrikman,[2] assuming micro- and macroisotropy, obtained the estimate (35) for the special cases $C^0 = \max C(x)$ and $C^0 = \min C(x)$, and showed that C^* so obtained were upper and lower bounds, respectively. Kröner,[6] in a statistical analysis of the elastic equation, obtained (35) with $C^0 = \langle C(x)\rangle$ and concluded that this approximation neglects the effects of correlation.

The term neglected in Eq. (35) is $\langle\varphi Aq\rangle$ and the best choice of C^0 in Eq. (35) would be one that minimizes this expression. Conforming to the approximation (34),

an estimate of this term is $\langle\varphi A\varphi\rangle(\Gamma\langle q\rangle - 1)$. A simple and reasonable choice is defined by

$$\langle\varphi\rangle = 0 \tag{36}$$

which, when inserted in Eq. (35), makes this approximation 'self-consistent.' In lieu of solving the implicit equation (36) directly, it is easier to use an iteration based on Eq. (35). In cases where this approximation makes sense, this process will converge. Beginning with $C^0 = \langle C(x)\rangle$, Kröner's result is obtained in the first step. The ultimate approximation of C^* is entirely expressed in terms of space-averages of functions of $C(x)$, in short, it depends on the composition of the medium, but not on its texture. In a following section, the approximation defined by (35) and (36) will be obtained as a special case of a more general class of approximations. This self-consistent, texture independent approximation will be denoted C^{*0}.

4. SELF-CONSISTENT IMBEDDING

It is well known that C^* is a proximity property.[7] That is, it depends in a sensitive way on the variation of $C(x)$ from point to point in any small neighbourhood, but does not depend on the correlation between the values of $C(x)$ in widely separated points.

This property can be used to find approximate solutions of Eq. (32). Let Ω be any finite simply connected volume, and let an elastic medium be defined by $C_\Omega(x)$ through

$$\begin{aligned} C_\Omega(x) &= C(x), \quad x \in \Omega, \\ C_\Omega(x) &= C^0, \quad x \notin \Omega, \end{aligned} \tag{37}$$

i.e., C^0 takes on the function of imbedding medium. Correspondingly,

$$\begin{aligned} \varphi_\Omega(x) &= \varphi(x), \quad x \in \Omega, \\ \varphi_\Omega(x) &= 0, \quad x \notin \Omega. \end{aligned} \tag{38}$$

Let $q_\Omega(x)$ be the solution of Eq. (32) for this medium, while retaining $\langle q\rangle$ in this equation as an unknown constant, i.e.,

$$q_\Omega + \varphi_\Omega A q_\Omega + \varphi_\Omega(\Gamma\langle q\rangle - 1) = 0. \tag{39}$$

One has

$$q_\Omega(x) = 0, \quad x \notin \Omega \tag{40}$$

and one expects the approximate relation

$$q_\Omega(x) \approx q(x), \quad x \in \Omega, \tag{41}$$

the error being dependent on C^0. In order to find $\langle q\rangle$, consider Eq. (39) to be solved for an ensemble of neighborhoods. Then define

$$\langle q\rangle = \langle\langle q_\Omega\rangle_\Omega\rangle_{\mathrm{av}}, \tag{42}$$

where $\langle\ \rangle_\Omega$ is the average over Ω, and $\langle\ \rangle_{\mathrm{av}}$ is the ensemble average.

Equations (39) and (42) define the problem, but C^0 is still arbitrary. A good choice is to imbed in the mean medium itself, aiming at $\langle q\rangle = 0$. In the present approximation this means

$$\langle\langle q_\Omega\rangle_\Omega\rangle_{\mathrm{av}} = 0. \tag{43}$$

Equations (33), (39), and (42) suggest a simple iteration scheme to achieve this: one begins with, e.g., $C^0 = C^{*0}$, solves (39) and (42), and calculates the first approximation of C^* with Eq. (33). Using this value as new C^0, one repeats the calculation, etc.

For $\Omega \to \infty$, the above results become exact. The actual solution of Eq. (39) in finite, small neighborhoods Ω can, e.g., be obtained with numerical methods.

5. ELLIPSOIDAL MODEL

A case of the self-consistent imbedding approximation which can be traced analytically is obtained by considering only neighborhoods of ellipsoidal shape E inside of which the $C(x)$ are constants. How, for a given material, such an ensemble can be constructed and whether it is adequate, will not be discussed. The special thing about this case is that Eq. (39) has the solution

$$q_E(x) = \text{const}, \quad x \in E, \tag{44}$$

and that

$$Aq_E(x) = (\Lambda_E - \Gamma)q_E, \tag{45}$$

where Λ_E is a tensor whose components are known functions of the value of C inside the ellipsoids, the ratio of the values of the ellipsoidal axis (a, b, c) and its orientation in space. Eshelby[8] has solved this problem. From the definition (28) one sees that A, where applied to a function which is constant inside an ellipsoid and zero outside, can be transformed to

$$Aq_E = \left(\int_{\Sigma-\sigma} S n d^2\hat{x}\right) q_E, \tag{46}$$

where Σ is the surface of the ellipsoid, σ the surface of a small sphere around the singular point. Both integrals are indeed constants and size-independent. Defining

$$\Lambda_E = \int_\Sigma S n d^2\hat{x} \tag{47}$$

and remembering Eq. (25), one obtains Eq. (45). In the special case that all ellipsoids are spheres, one has $\Lambda_E = \Gamma$ and therefore $Aq_E = 0$, which reduces to the approximation in Sec. 3.

Substitution of (45) in Eq. (39) gives

$$q_E + \Psi_E(\Gamma\langle q\rangle - 1) = 0, \tag{48}$$

where

$$\Psi_E = [(C_E - C^0)^{-1} - \Lambda_E]^{-1}. \tag{49}$$

Taking the ensemble average and using Eq. (42) gives

$$\langle q\rangle = (1 + \langle\Psi\rangle\Gamma)^{-1}\langle\Psi\rangle, \tag{50}$$

where

$$\langle\Psi\rangle = \langle\Psi_E\rangle_{\mathrm{av}}. \tag{51}$$

The calculation of $\langle\Psi\rangle$ requires the specification of the ensemble over shapes, orientations and elastic constants of the ellipsoids.

Walsh[9] has used Eshelby's results to estimate the effect of small concentrations of ellipsoidal inclusions in an otherwise homogeneous medium. When used as a special case of a self-consistent imbedding approximation, as suggested here, one can expect greater flexibility and accuracy.

6. APPROXIMATION IN TERMS OF CORRELATION FUNCTIONS

An alternative to the approximations discussed thus far is one in terms of correlation functions. Equation (32)

lends itself to this approach. Without touching upon the merits of this method, a simple example will illustrate this point.

From Eq. (32) one obtains by a single iteration

$$q - \varphi A\varphi Aq - (\varphi A\varphi - \varphi)(\Gamma\langle q\rangle - 1) = 0. \tag{52}$$

Next, the operator A is applied, the resulting equation is multiplied with the components of the tensor $\varphi(x + y)$ (total multiplication, denoted $\otimes$) and averaged over space. This gives

$$\Pi(y) - \langle\varphi(x + y)\otimes A\varphi(x)A\varphi(x)Aq(x)\rangle + [K^{(1)}(y) - K^{(2)}(y)](\Gamma\langle q\rangle - 1) = 0, \tag{53}$$

where

$$\Pi(y) = \langle\varphi(x + y)\otimes Aq(x)\rangle, \tag{54}$$

$$K^{(n)}(y) = \langle\varphi(x + y)\otimes[A\varphi(x)]^n\rangle. \tag{55}$$

The second term of Eq. (53), involving the average of four factors in four different points is now truncated in such a way that the result is correct for all configurations of these points with a possible exception when all four points are close together. The validity of this approximation will thus depend on the relative contribution of these configurations to this term.

For a product of four random variables a, b, c and d this truncation is defined by

$$\langle abcd\rangle \to (\langle abc\rangle\langle d\rangle + \cdots) + (\langle ab\rangle\langle cd\rangle + \cdots) - 2(\langle ab\rangle\langle c\rangle\langle d\rangle + \cdots) + 6\langle a\rangle\langle b\rangle\langle c\rangle\langle d\rangle. \tag{56}$$

As applied to Eq. (53), three of the variables are φ's in different points and one is q in a fourth point. When it is assumed that there is no long-range correlation, the substitution (56) is in error only when the four points form a cluster.

Of all the terms in (56) those in which the operator A is applied to an average, i.e., to a constant, are zero. Of the other terms those of the form $\langle a\langle b\rangle cd\rangle$ and $\langle ab\langle c\rangle d\rangle$ are new unknown quantities. In order that the truncation succeeds, it is therefore necessary to choose C^0 in such a way that these terms vanish. This is achieved by

$$\langle\varphi\rangle = 0, \tag{57}$$

i.e., the C^0 have to be the texture-independent estimate C^{*0} found in Sec. 3. With this, the second term in Eq. (53) reduces to

$$\langle\varphi(x + y)\otimes A\varphi(x)A\varphi(x)Aq(x) \to \langle\varphi(x + y)\otimes A\langle\langle\varphi(x)A\varphi(x)\rangle\rangle Aq(x)\rangle + \langle\varphi(x + y)\otimes A\langle\langle\varphi(x)A\varphi(x)\rangle\rangle Aq(x)\rangle\rangle, \tag{58}$$

where double brackets are used to indicate the pairing. In both terms the factor q is combined with φ as in one of the components of the tensor Π, while the other factor is a component of the tensor

$$F(z) = \langle\varphi(x + z)\otimes\varphi(x)\rangle. \tag{59}$$

In order to write the resulting equations explicitly, the indices have to be reintroduced. As all tensors are symmetric in pairs of (Latin) indices, each pair can be denoted by one (Greek) index, e.g., $(i, j) = \lambda$.

Interchanging the order of the averaging with the integration in the A operator gives

$$\Pi_{\rho\sigma\tau\eta}(y) - \int dx' T_{\kappa\mu}(x')\int dx'' T_{\tau\lambda}(x'') \times[F_{\rho\sigma\mu\nu}(x' + x'' + y)\Pi_{\lambda\kappa\nu\eta}(x') + \Pi_{\rho\sigma\mu\eta}(x' + x'' + y)F_{\lambda\kappa\mu\nu}(x')] + [K^{(1)}_{\rho\sigma\mu\nu}(y) - K^{(2)}_{\rho\sigma\tau\nu}(y)](\Gamma_{\nu\mu}\langle q_{\mu\eta}\rangle - \delta_{\nu\eta}) = 0. \tag{60}$$

Here, according to Eqs. (54) and (55),

$$F_{\rho\sigma\mu\nu}(y) = \langle\varphi_{\rho\sigma}(x + y)\varphi_{\mu\nu}(x)\rangle,$$
$$K^{(1)}_{\rho\sigma\tau\nu}(y) = \int dx' T_{\tau\lambda}(x')F_{\rho\sigma\lambda\nu}(x' + y),$$
$$K^{(2)}_{\rho\sigma\tau\nu}(y) = \int dx'' T_{\tau\lambda}(x'')\int dx' T_{\kappa\mu}(x') \times\langle\varphi_{\rho\sigma}(x + x'' + y)\varphi_{\lambda\kappa}(x)\varphi_{\mu\nu}(x - x')\rangle. \tag{61}$$

The quantity $\langle q\rangle$ in the last term of Eq. (60) can be expressed in terms of $\Pi(0)$. By taking the average of Eq. (32) and using $\langle\varphi\rangle = 0$, one finds, from the definition (54) of Π,

$$\langle q_{\rho\mu}\rangle = -\Pi_{\rho\lambda\lambda\mu}(0). \tag{62}$$

From the solution $\Pi(y)$ of Eq. (60) the mean elastic constants are found with Eq. (33) and Eq. (62). Thus,

$$C^*_{\rho\eta} = C^0_{\rho\eta} - \Pi_{\rho\sigma\sigma\eta}(0), \tag{63}$$

where C^0 have the values C^{*0} as indicated above.

The two-point correlation function $F(z)$ defined by Eq. (59) plays a very special role in the kernal of Eq. (60) for $\Pi(y)$. This illustrates the difference between a cluster expansion, as used above, and a straightforward iteration method[1,10]. For example, from Eq. (32) one easily derives

$$q + (-1)^n(\varphi A)^{n+1}q + \sum_{k=0}^{n}(-1)^k(\varphi A)^k\varphi(\Gamma\langle q\rangle - 1) = 0. \tag{64}$$

Assuming that the expansion converges in the average, i.e., that

$$\lim_{n\to\infty}\langle(\varphi A)^n q\rangle = 0, \tag{65}$$

one is led to

$$\langle q\rangle + \sum_{k=0}^{\infty}(-1)^k\langle(\varphi A)^k\varphi\rangle(\Gamma\langle q\rangle - 1) = 0. \tag{66}$$

In this series the first term is zero on account of Eq. (57). The second term vanishes,

$$\langle\varphi A\varphi\rangle \equiv \langle\varphi(x)\int d^3x' T(x - x')\varphi(x')\rangle = \int d^3x'\langle\varphi(x)T(x - x')\varphi(x')\rangle = 0 \tag{67}$$

because of the relation between $T(x - x')$ and the symmetry of the crystal. (For example, if the medium is macro-isotropic, $\langle\varphi(x)\varphi(x - x')\rangle$ depends only on $|x'|$, while the average of T over all directions is zero.) This leaves only third- and higher-order correlations in φ, in apparent contrast to Eq. (60). However, one can show that $K^{(1)}$ in Eq. (60) is zero for macroisotropic bodies, thus removing the two-point correlation from the driving term.

ACKNOWLEDGMENT

The author is indebted to Dr. R. J. Runge for valuable discussions and a critical reading of the manuscript.

*Permanent address: Physics Dept., The Ohio State University, Columbus, Ohio 43210.

[1]For up-to-date references see the review article by M. J. Beran, Phys. Status Solidi A **6,** 365 (1971).

[2]This definition is equivalent to $U \equiv <\epsilon C \epsilon> = <\epsilon> C^* <\epsilon>$, as used in, e.g., Ref. 3. This follows from (3), (5), and (23) with use of Gauss' theorem. Here and in the following a tensor notation is used, where possible, that avoids trivial use of indices. A vector or tensor is indicated by the same letter that serves for the full identification of its components, e.g., $x \equiv x_i$, $u \equiv u_i$, $\epsilon \equiv \epsilon_{ij}$, $\sigma \equiv \sigma_{ij}$, $C \equiv C_{ijkl}$, etc. Tensor products are written as the (noncommutative) products of these symbols, but this convention is restricted to the case that summation is over a pair of indices in which both tensors are symmetric, e.g., Eq. (2). All other cases are written out explicitly.

[3]Z. Hashin and S. Shtrikman, J. Mech. Phys. Solids **11,** 127 (1963).

[4]This tensor was first introduced in J. D. Eshelby, Proc. R. Soc. A **241,** 376 (1957).

[5]The explicit appearance of $<p>$ in the final equations was first noted in Ref. 3.

[6]E. Kröner, J. Mech. Phys. Solids **15,** 319 (1967).

[7]This is analogous to the case of electron band structure of random alloys; see W. H. Butler and W. Kohn in "Electronic Density of States," in *N.B.S. Special Publication,* edited by L. H. Bennet (Natl. Bur. Stds., Washington, D. C., 1971), Vol. 323, p. 465.

[8]J. D. Eshelby, Ref. 4.

[9]J. B. Walsh, J. Geophys. Res. **74,** 4333 (1969).

[10]M. N. Miller, J. Math. Phys. **10,** 1988 (1969); J. Math. Phys. **10,** 2005 (1969).

Statistical theory of effective electrical, thermal, and magnetic properties of random heterogeneous materials. I. Perturbation expansions for the effective permittivity of cell materials

Motoo Hori

Department of Applied Physics, Tokyo Institute of Technology, Meguroku, Tokyo 152, Japan
(Received 2 August 1972; revised manuscript received 16 November 1972)

A perturbation formulation is developed for the effective permittivity of random heterogeneous materials that are statistically homogeneous but not necessarily statistically isotropic. The formal perturbation solution is expressed explicitly by means of the many-point correlation functions of the permittivity field. For a broad class of random multiphase materials called cell materials, the second- and third-order perturbation effects are determined as functions of the depolarizing factor tensors of cells. In the special case where the medium is statistically isotropic, the formulas giving the effective permittivity to third order are shown to be in substantial accord with the results obtained by previous investigators.

1. INTRODUCTION

The present work is concerned with the effective or overall physical property of a heterogeneous material whose local property may be regarded as a random function of position. It is of practical importance to predict the effective physical constants of heterogeneous materials such as suspensions, mixtures, polycrystals, and composite materials. We shall restrict ourselves to electrical, thermal, and magnetic properties that are described by a proportionality factor between a solenoidal vector and an irrotational vector. The terminology used hereafter will be that of the dielectric problem. For reasons of mathematical analogy the results obtained also hold for magnetic permeability, electrical and thermal conductivity, and diffusion constant.

The problem of determining the effective permittivity of a random heterogeneous material has received repeated attention in recent years. Brown,[1] Beran and Molyneux,[2] and other authors[3-8] employed perturbation methods to analyze the static dielectric behavior of a medium with small random fluctuations in permittivity. The perturbation theories were extended to treat the dynamic dielectric behavior of a time-dependent heterogeneous medium.[9-14] Brown,[15] Beran,[16] and Beran and Molyneux[17] derived bounds on the effective permittivity by using perturbation series as trial functions. Miller[18,19] and Beran and Silnutzer[20] showed that for a wide class of random multiphase materials these bounds can be expressed in terms of volume fractions and shape factors.

Most of previous investigations deal with random media which are considered to be statistically homogeneous and isotropic. However, in some cases the assumption of statistical isotropy is not appropriate. For example, suspensions of oriented particles or fiber-reinforced composites are anisotropic on a macroscopic scale. The purpose of this paper is to obtain perturbation expansions for the effective permittivity of such inhomogeneous anisotropic media. In Sec. 2 we shall develop a general perturbation formulation to cover the case where the material is not necessarily statistically isotropic. In Sec. 3 explicit calculations for the second-order and third-order perturbation effects will be carried out on the basis of the cell model introduced by Miller.[18,19]

2. OUTLINE OF THE PERTURBATION FORMULATION

A. Basic concepts and governing equations

We are interested in a heterogeneous material with random variations in permittivity. The volume V of the material is taken to be infinite. Assume that the medium is locally isotropic and let $\epsilon(\mathbf{r})$ be the permittivity at a point $\mathbf{r}$. Then $\{\epsilon(\mathbf{r}), \mathbf{r}\in V\}$ forms a stochastic process with continuous space parameters. The n-point moment of $\epsilon(\mathbf{r})$ is defined as $\langle\epsilon(\mathbf{r}_1)\epsilon(\mathbf{r}_2)\cdots\epsilon(\mathbf{r}_n)\rangle$, where the brackets $\langle\ \rangle$ denote the mathematical expectation or the ensemble average. Assume further that the process $\{\epsilon(\mathbf{r})\}$ is almost surely differentiable as many times as required. As a matter of fact, the process representing a random multiphase material possesses derivatives of all orders with probability one. In such a case it may be shown that under mild conditions there exists the nth partial derivative of $\langle\epsilon(\mathbf{r}_1)\epsilon(\mathbf{r}_2)\cdots\epsilon(\mathbf{r}_n)\rangle$ with respect to $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$.

In the following we shall suppose that the medium is statistically homogeneous, or equivalently, that the process $\{\epsilon(\mathbf{r})\}$ is strictly stationary in space. Thus, the n-point moment $\langle\epsilon(\mathbf{r}_1)\epsilon(\mathbf{r}_2)\cdots\epsilon(\mathbf{r}_n)\rangle$ does not depend upon the absolute positions $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$, but instead depends only upon the relative positions $\mathbf{r}_{12} = \mathbf{r}_2 - \mathbf{r}_1, \mathbf{r}_{13} = \mathbf{r}_3 - \mathbf{r}_1, \ldots, \mathbf{r}_{1n} = \mathbf{r}_n - \mathbf{r}_1$; that is,

$$\langle\epsilon(\mathbf{r}_1)\epsilon(\mathbf{r}_2)\cdots\epsilon(\mathbf{r}_n)\rangle = \langle\epsilon(0)\epsilon(\mathbf{r}_{12})\cdots\epsilon(\mathbf{r}_{1n})\rangle. \tag{2.1}$$

In particular, the expectation $\langle\epsilon(\mathbf{r})\rangle$ becomes a constant within the material volume V, so that we can write $\langle\epsilon(\mathbf{r})\rangle = \langle\epsilon\rangle$. Note that statistical isotropy is not always assumed. For a statistically homogeneous medium it is common to make an ergodic-type hypothesis that the ensemble average can be replaced by the volume average. In this sense, we postulate that

$$\langle\epsilon(\mathbf{r})\rangle = \lim_{V\to\infty}(1/V)\int_V \epsilon(\mathbf{r})d\omega \tag{2.2}$$

holds with probability one, $d\omega$ indicating a volume element of V at $\mathbf{r}$.

The basic equations governing the static electric field in a material with variable permittivity $\epsilon(\mathbf{r})$ are

$$\frac{\partial}{\partial x_i}[\epsilon(\mathbf{r})E_i(\mathbf{r})] = 0, \tag{2.3}$$

$$E_i(\mathbf{r}) = -\frac{\partial\Phi(\mathbf{r})}{\partial x_i}. \tag{2.4}$$

Here $E_i(\mathbf{r})$ is the ith component of the electric field, $\Phi(\mathbf{r})$ is the electrostatic potential, and the summation convention has been employed. For simplicity we shall

consider the average electric field to be independent of $\mathbf{r}$. Fluctuations about the mean values of $E_i(\mathbf{r})$ and $\epsilon(\mathbf{r})$ are therefore given by

$$E_i'(\mathbf{r}) = E_i(\mathbf{r}) - \langle E_i \rangle, \quad \langle E_i'(\mathbf{r}) \rangle = 0, \tag{2.5}$$

$$\epsilon'(\mathbf{r}) = \epsilon(\mathbf{r}) - \langle \epsilon \rangle, \quad \langle \epsilon'(\mathbf{r}) \rangle = 0. \tag{2.6}$$

In terms of $E_i'(\mathbf{r})$ and $\epsilon'(\mathbf{r})$ the governing equation (2.3) can be written as

$$\frac{\partial E_i'(\mathbf{r})}{\partial x_i} = -\frac{\langle E_i \rangle}{\langle \epsilon \rangle} \frac{\partial \epsilon'(\mathbf{r})}{\partial x_i} - \frac{1}{\langle \epsilon \rangle} \frac{\partial}{\partial x_i} [\epsilon'(\mathbf{r}) E_i'(\mathbf{r})]. \tag{2.7}$$

Using the free-space Green's function for the Laplacian operator, one may formally solve Eq. (2.7) as follows[2]:

$$E_i'(\mathbf{r}_1) = \frac{1}{4\pi\langle \epsilon \rangle} \left(\langle E_j \rangle \int_V d\omega_2 \frac{x_{12,i}}{r_{12}^3} \frac{\partial \epsilon'(\mathbf{r}_2)}{\partial x_{2,j}} + \int_V d\omega_2 \frac{x_{12,i}}{r_{12}^3} \frac{\partial}{\partial x_{2,j}} [\epsilon'(\mathbf{r}_2) E_j'(\mathbf{r}_2)] \right), \tag{2.8}$$

where $d\omega_2$ designates a volume element in the space of $\mathbf{r}_2$. It should be noticed that the comma in a subscript does not mean partial differentiation. Equation (2.8) is a random integral equation for $E_i'(\mathbf{r}_1)$ which we wish to solve.

The effective permittivity tensor ϵ_{ij}^* of a heterogeneous medium having statistical anisotropy is defined by the linear relation

$$\langle \epsilon(\mathbf{r}) E_i(\mathbf{r}) \rangle = \epsilon_{ij}^* \langle E_j \rangle, \tag{2.9}$$

or alternatively by the energy relation

$$\tfrac{1}{2} \langle \epsilon(\mathbf{r}) E_i(\mathbf{r}) E_i(\mathbf{r}) \rangle = \tfrac{1}{2} \epsilon_{ij}^* \langle E_i \rangle \langle E_j \rangle. \tag{2.10}$$

The equivalence of the above two definitions can be established exactly as in the case of statistically isotropic materials.[21] Substitution of Eqs. (2.5) and (2.6) into Eq. (2.9) gives

$$\epsilon_{ij}^* \langle E_j \rangle = \langle \epsilon \rangle \langle E_i \rangle + \langle \epsilon'(\mathbf{r}) E_i'(\mathbf{r}) \rangle. \tag{2.11}$$

If we set

$$\langle \epsilon'(\mathbf{r}) E_i'(\mathbf{r}) \rangle = -A_{ij} \langle \epsilon \rangle \langle E_j \rangle, \tag{2.12}$$

the factor A_{ij} is a dimensionless quantity determined by the statistical properties of $\epsilon(\mathbf{r})$. It follows from Eq. (2.11) that

$$\epsilon_{ij}^* = \langle \epsilon \rangle (\delta_{ij} - A_{ij}), \tag{2.13}$$

where δ_{ij} is the Kronecker delta. For a statistically isotropic medium the effective permittivity tensor ϵ_{ij}^* reduces to a scalar ϵ^* such that

$$\epsilon^* = \langle \epsilon \rangle (1 - A). \tag{2.14}$$

B. Formal perturbation solution

In order to use a perturbation procedure we should have a small quantity that is of interest. Here the fluctuations in permittivity are considered to be very small compared with its mean value. In statistical terms we assume that with probability one

$$|\epsilon'(\mathbf{r})| / \langle \epsilon \rangle \ll 1, \tag{2.15}$$

and consequently that for any positive integer n

$$\langle |\epsilon'^n| \rangle / \langle \epsilon \rangle^n \ll 1, \quad \langle |\epsilon'^{n+1}| \rangle / \langle \epsilon \rangle^{n+1} = o(\langle |\epsilon'^n| \rangle / \langle \epsilon \rangle^n). \tag{2.16}$$

Moreover, it is supposed that the solution of Eq. (2.7) can be expanded in a power series of the parameter $\epsilon'(\mathbf{r})/\langle \epsilon \rangle$ as follows:

$$E_i'(\mathbf{r}) = E_i^{(1)}(\mathbf{r}) + E_i^{(2)}(\mathbf{r}) + \cdots = \sum_{n=1}^{\infty} E_i^{(n)}(\mathbf{r}), \tag{2.17}$$

where $E_i^{(n)}(\mathbf{r})$ is of the same order of smallness as $\langle E_i \rangle [\epsilon'(\mathbf{r})/\langle \epsilon \rangle]^n$ and satisfies

$$\langle E_i^{(n)}(\mathbf{r}) \rangle = 0. \tag{2.18}$$

In this paper we shall not discuss the problem of convergence of such perturbation expansions.

The integral equation (2.8) may be solved formally by iteration. Substituting Eq. (2.17) into Eq. (2.8) and equating terms of like order of magnitude, we find

$$E_i^{(1)}(\mathbf{r}_1) = \frac{\langle E_j \rangle}{4\pi\langle \epsilon \rangle} \int_V d\omega_2 \frac{x_{12,i}}{r_{12}^3} \frac{\partial \epsilon'(\mathbf{r}_2)}{\partial x_{2,j}}, \tag{2.19}$$

$$E_i^{(2)}(\mathbf{r}_1) = \frac{\langle E_j \rangle}{(4\pi\langle \epsilon \rangle)^2} \int_V d\omega_2 \frac{x_{12,i}}{r_{12}^3} \frac{\partial}{\partial x_{2,k}} \int_V d\omega_3 \frac{x_{23,k}}{r_{23}^3} \times \frac{\partial \epsilon'(\mathbf{r}_2) \epsilon'(\mathbf{r}_3)}{\partial x_{3,j}}, \tag{2.20}$$

and so on. For later convenience, we rewrite these equations in the form

$$E_i^{(1)}(\mathbf{r}_1) = \frac{\langle E_j \rangle}{4\pi\langle \epsilon \rangle} \int_V d\omega_{12} \frac{x_{12,i}}{r_{12}^3} \frac{\partial \epsilon'(\mathbf{r}_2)}{\partial x_{12,j}}, \tag{2.21}$$

$$\begin{aligned} E_i^{(2)}(\mathbf{r}_1) &= \frac{\langle E_j \rangle}{(4\pi\langle \epsilon \rangle)^2} \int_V d\omega_2 \frac{x_{12,i}}{r_{12}^3} \frac{\partial}{\partial x_{2,k}} \int_V d\omega_{23} \times \frac{x_{23,k}}{r_{23}^3} \frac{\partial \epsilon'(\mathbf{r}_2) \epsilon'(\mathbf{r}_3)}{\partial x_{23,j}} \\ &= \frac{\langle E_j \rangle}{(4\pi\langle \epsilon \rangle)^2} \int_V d\omega_{12} \frac{x_{12,i}}{r_{12}^3} \frac{\partial}{\partial x_{12,k}} \int_V d\omega_{23} \frac{x_{23,k}}{r_{23}^3} \times \frac{\partial \epsilon'(\mathbf{r}_2) \epsilon'(\mathbf{r}_3)}{\partial x_{23,j}} \\ &= \frac{\langle E_j \rangle}{(4\pi\langle \epsilon \rangle)^2} \int_V d\omega_{12} \int_V d\omega_{23} \frac{x_{12,i}}{r_{12}^3} \frac{x_{23,k}}{r_{23}^3} \frac{\partial^2 \epsilon'(\mathbf{r}_2) \epsilon'(\mathbf{r}_3)}{\partial x_{12,k} \partial x_{23,j}}. \end{aligned} \tag{2.22}$$

Similarly,

$$E_i^{(n)}(\mathbf{r}_1) = \frac{\langle E_j \rangle}{(4\pi\langle \epsilon \rangle)^n} \int_V d\omega_{12} \int_V d\omega_{23} \cdots \int_V d\omega_{n,n+1} \times \frac{x_{12,i}}{r_{12}^3} \frac{x_{23,k}}{r_{23}^3} \cdots \frac{\partial^n \epsilon'(\mathbf{r}_2) \epsilon'(\mathbf{r}_3) \cdots \epsilon'(\mathbf{r}_{n+1})}{\partial x_{12,k} \partial x_{23,h} \cdots \partial x_{n,n+1;j}}. \tag{2.23}$$

The independent variables in the integrals have been changed from $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{n+1}$ to $\mathbf{r}_1, \mathbf{r}_{12} = \mathbf{r}_2 - \mathbf{r}_1, \ldots, \mathbf{r}_{n,n+1} = \mathbf{r}_{n+1} - \mathbf{r}_n$. If averaging commutes with differentiation and integration,

$$\langle E_i^{(n)}(\mathbf{r}_1)\rangle = \frac{\langle E_j\rangle}{(4\pi\langle\epsilon\rangle)^n}\int_V d\omega_{12}\int_V d\omega_{23}\cdots\int_V d\omega_{n,n+1}$$

$$\times\frac{x_{12,i}}{r_{12}^3}\frac{x_{23,k}}{r_{23}^3}\cdots\frac{\partial^n\langle\epsilon'(\mathbf{r}_2)\epsilon'(\mathbf{r}_3)\cdots\epsilon'(\mathbf{r}_{n+1})\rangle}{\partial x_{12,k}\partial x_{23,h}\cdots\partial x_{n,n+1;j}}. \quad (2.24)$$

The nth-order perturbation field $E_i^{(n)}(\mathbf{r}_1)$ in Eq. (2.23) obeys automatically the requirement (2.18), because $\langle\epsilon'(\mathbf{r}_2)\epsilon'(\mathbf{r}_3)\cdots\epsilon'(\mathbf{r}_{n+1})\rangle$ is independent of $\mathbf{r}_{12}$.

Let us seek the perturbation series of the effective permittivity tensor ϵ_{ij}^*. By introducing Eq. (2.17) in Eq. (2.11), we see that

$$\epsilon_{ij}^*\langle E_j\rangle = \langle\epsilon\rangle\langle E_i\rangle + \sum_{n=1}^{\infty}\langle\epsilon'(\mathbf{r}_1)E_i^{(n)}(\mathbf{r}_1)\rangle. \quad (2.25)$$

By virtue of Eq. (2.23), it follows that

$$\langle\epsilon'(\mathbf{r}_1)E_i^{(n)}(\mathbf{r}_1)\rangle = \frac{\langle E_j\rangle}{(4\pi\langle\epsilon\rangle)^n}\int_V d\omega_{12}\int_V d\omega_{23}\cdots\int_V d\omega_{n,n+1}$$

$$\times\frac{x_{12,i}}{r_{12}^3}\frac{x_{23,k}}{r_{23}^3}\cdots\frac{\partial^n\langle\epsilon'(\mathbf{r}_1)\epsilon'(\mathbf{r}_2)\cdots\epsilon'(\mathbf{r}_{n+1})\rangle}{\partial x_{12,k}\partial x_{23,h}\cdots\partial x_{n,n+1;j}}. \quad (2.26)$$

Since the medium is statistically homogeneous, the $(n+1)$th moment $\langle\epsilon'(\mathbf{r}_1)\epsilon'(\mathbf{r}_2)\cdots\epsilon'(\mathbf{r}_{n+1})\rangle$ can be represented as

$$\langle\epsilon'(\mathbf{r}_1)\epsilon'(\mathbf{r}_2)\cdots\epsilon'(\mathbf{r}_{n+1})\rangle = \langle\epsilon'^{n+1}\rangle f(\mathbf{r}_{12},\mathbf{r}_{13},\ldots,\mathbf{r}_{1,n+1}). \quad (2.27)$$

The dimensionless quantity $f(\mathbf{r}_{12},\mathbf{r}_{13},\ldots,\mathbf{r}_{1,n+1})$ signifies the normalized $(n+1)$-point correlation function of $\epsilon(\mathbf{r})$. Combination of Eq. (2.13) with Eqs. (2.26) and (2.27) yields

$$\epsilon_{ij}^* = \langle\epsilon\rangle\left(\delta_{ij} - \sum_{n=2}^{\infty}(-1)^n A_{ij}^{(n)}\frac{\langle\epsilon'^n\rangle}{\langle\epsilon\rangle^n}\right), \quad (2.28)$$

where

$$A_{ij}^{(n)} = \left(\frac{-1}{4\pi}\right)^{n-1}\int_V d\omega_{12}\int_V d\omega_{23}\cdots\int_V d\omega_{n-1,n}$$

$$\times\frac{x_{12,i}}{r_{12}^3}\frac{x_{23,k}}{r_{23}^3}\cdots\frac{\partial^{n-1}f(\mathbf{r}_{12},\mathbf{r}_{13},..,\mathbf{r}_{1n})}{\partial x_{12,k}\partial x_{23,h}\cdots\partial x_{n-1,n;j}}. \quad (2.29)$$

The nth-order perturbation coefficient $A_{ij}^{(n)}$ is thus expressed explicitly in terms of the normalized n-point correlation function $f(\mathbf{r}_{12},\mathbf{r}_{13},\ldots,\mathbf{r}_{1n})$.

C. Second-order perturbation term

The second-order perturbation coefficient $A_{ij}^{(2)}$ is a dimensionless second-order tensor defined as

$$A_{ij}^{(2)} = -\frac{1}{4\pi}\int_V d\omega_{12}\frac{x_{12,i}}{r_{12}^3}\frac{\partial f(\mathbf{r}_{12})}{\partial x_{12,j}}, \quad (2.30)$$

with

$$f(\mathbf{r}_{12}) = \langle\epsilon'(\mathbf{r}_1)\epsilon'(\mathbf{r}_2)\rangle/\langle\epsilon'^2\rangle. \quad (2.31)$$

Henceforth, we shall call $A_{ij}^{(2)}$ the average depolarizing factor tensor or the effective magnetometric demagnetizaton tensor. As regards the reason of this naming we refer to Sec. 3B. The normalized two-point correlation function $f(\mathbf{r}_{12})$ satisfies the well-known inequality

$$|f(\mathbf{r}_{12})| < 1. \quad (2.32)$$

Boundary conditions imposed upon $f(\mathbf{r}_{12})$ are

$$f(\mathbf{r}_{12}) \to \begin{cases} 1 & \text{as } |\mathbf{r}_{12}| \to 0, \\ 0 & \text{as } |\mathbf{r}_{12}| \to \infty, \end{cases} \quad (2.33)$$

for it is natural to expect that the correlation between $\epsilon'(\mathbf{r}_1)$ and $\epsilon'(\mathbf{r}_1+\mathbf{r}_{12})$ must vanish as $|\mathbf{r}_{12}|\to\infty$. In a spherical coordinate system $(r_{12},\theta_{12},\phi_{12})$ the two-point correlation function $f(r_{12},\theta_{12},\phi_{12})$ meets the conditions

$$f(0,\theta_{12},\phi_{12}) = 1, \quad f(\infty,\theta_{12},\phi_{12}) = 0. \quad (2.34)$$

For the partial derivatives of $f(r_{12},\theta_{12},\phi_{12})$ we assume that as $r_{12}\to\infty$

$$\frac{\partial f(r_{12},\theta_{12},\phi_{12})}{\partial r_{12}} = o\left(\frac{1}{r_{12}}\right), \quad (2.35a)$$

$$\frac{\partial f(r_{12},\theta_{12},\phi_{12})}{\partial\theta_{12}} \to 0, \quad (2.35b)$$

$$\frac{\partial f(r_{12},\theta_{12},\phi_{12})}{\partial\phi_{12}} \to 0. \quad (2.35c)$$

The expression (2.30) for $A_{ij}^{(2)}$ may be recast as

$$A_{ij}^{(2)} = -\frac{1}{4\pi}\int_V d\omega_{12}\frac{1}{r_{12}}\frac{\partial^2 f(\mathbf{r}_{12})}{\partial x_{12,i}\partial x_{12,j}}. \quad (2.36)$$

In fact, after integrating by parts and applying Gauss' theorem, we obtain from Eq. (2.30)

$$A_{ij}^{(2)} = \frac{1}{4\pi}\int_V d\omega_{12}\frac{\partial}{\partial x_{12,i}}\left(\frac{1}{r_{12}}\right)\frac{\partial f(\mathbf{r}_{12})}{\partial x_{12,j}}$$

$$= \frac{1}{4\pi}\int_S\frac{d\sigma_{12}}{r_{12}^2}n_i r_{12}\frac{\partial f(\mathbf{r}_{12})}{\partial x_{12,j}} - \frac{1}{4\pi}\int_V d\omega_{12}$$

$$\times\frac{1}{r_{12}}\frac{\partial^2 f(\mathbf{r}_{12})}{\partial x_{12,i}\partial x_{12,j}}, \quad (2.37)$$

where S is the bounding surface of the medium V, $d\sigma_{12}$ an area element of the surface S, and n_i the x_i component of an outward unit normal to the surface element $d\sigma_{12}$. Since Eqs. (2.35a)–(2.35c) imply that $r_{12}\partial f(\mathbf{r}_{12})/\partial x_{12,j}\to 0$ as $r_{12}\to\infty$, the surface integral on the right-hand side of Eq. (2.37) approaches zero as $V\to\infty$, so that the proof of Eq. (2.36) is accomplished.

Another important formula expressing $A_{ij}^{(2)}$ is

$$A_{ij}^{(2)} = \frac{1}{(4\pi)^2}\int_V d\omega_1\int_V d\omega_2\frac{x_{1,k}}{r_1^3}\frac{x_{2,k}}{r_2^3}\frac{\partial^2 f(\mathbf{r}_{12})}{\partial x_{1,i}\partial x_{2,j}}. \quad (2.38)$$

Through the transformation of variables $(\mathbf{r}_1=\mathbf{r}_1, \mathbf{r}_{12}=\mathbf{r}_2-\mathbf{r}_1)$ the right-hand side in Eq. (2.38) becomes

$$\frac{1}{(4\pi)^2}\int_V d\omega_1\int_V d\omega_2\frac{x_{1,k}}{r_1^3}\frac{x_{2,k}}{r_2^3}\frac{\partial^2 f(\mathbf{r}_{12})}{\partial x_{1,i}\partial x_{2,j}}$$

$$= -\frac{1}{(4\pi)^2}\int_V d\omega_{12}\frac{\partial^2 f(\mathbf{r}_{12})}{\partial x_{12,i}\,\partial x_{12,j}}\int_V d\omega_1 \frac{x_{1,k}}{r_1{}^3}\frac{x_{2,k}}{r_2{}^3}. \tag{2.39}$$

In spherical coordinates we have

$$-\int_V d\omega_1 \frac{x_{1,k}}{r_1{}^3}\frac{x_{2,k}}{r_2{}^3} = \int_V d\omega_1 \frac{x_{1,k}}{r_1{}^3}\frac{\partial}{\partial x_{1,k}}\left(\frac{1}{r_2}\right)$$

$$= \int_0^{2\pi} d\phi_1 \int_0^{\pi} d\theta_1 \sin\theta_1 \int_0^{\infty} dr_1 \frac{x_{1,k}}{r_1}\frac{\partial}{\partial x_{1,k}}\left(\frac{1}{r_2}\right)$$

$$= \int_0^{2\pi} d\phi_1 \int_0^{\pi} d\theta_1 \sin\theta_1 \int_0^{\infty} dr_1 \frac{\partial}{\partial r_1}\left(\frac{1}{r_2}\right) = -\frac{4\pi}{r_{12}}, \tag{2.40}$$

considering that $\partial x_{1,k}/\partial r_1 = x_{1,k}/r_1$. Comparison of Eqs. (2.39) and (2.40) with Eq. (2.36) verifies that Eq. (2.38) is valid as asserted.

We now observe that the average depolarizing factor tensor $A_{ij}^{(2)}$ has the following fundamental properties:

$$\text{(i)}\quad A_{ij}^{(2)} = A_{ji}^{(2)}, \tag{2.41}$$

$$\text{(ii)}\quad A_{ii}^{(2)} = 1, \tag{2.42}$$

$$\text{(iii)}\quad A_{i(i)}^{(2)} \geq 0; \tag{2.43}$$

the subscript (i) indicates no summation on i. In words, $A_{ij}^{(2)}$ is a symmetric second-order tensor with nonnegative diagonal elements whose trace equals unity. Property (i) is obvious from Eq. (2.36) or (2.38). To deduce property (ii), set $j = i$ in Eq. (2.30) and sum up over the index i. Then,

$$A_{ii}^{(2)} = -\frac{1}{4\pi}\int_V \frac{d\omega_{12}}{r_{12}{}^2}\frac{x_{12,i}}{r_{12}}\frac{\partial f(\mathbf{r}_{12})}{\partial x_{12,i}}$$

$$= -\frac{1}{4\pi}\int_0^{2\pi} d\phi_{12}\int_0^{\pi} d\theta_{12}\sin\theta_{12} \times \int_0^{\infty} dr_{12}\frac{\partial f(r_{12},\theta_{12},\phi_{12})}{\partial r_{12}} = 1. \tag{2.44}$$

For the proof of property (iii) we use Eq. (2.38); that is,

$$A_{i(i)}^{(2)}$$

$$= \frac{1}{(4\pi)^2}\int_V d\omega_1\int_V d\omega_2 \frac{x_{1,k}}{r_1{}^3}\frac{x_{2,k}}{r_2{}^3}\frac{\partial^2}{\partial x_{1,i}\,\partial x_{2,(i)}}\frac{\langle\epsilon'(\mathbf{r}_1)\epsilon'(\mathbf{r}_2)\rangle}{\langle\epsilon'^2\rangle}$$

$$= \frac{1}{(4\pi)^2\langle\epsilon'^2\rangle}\left\langle\int_V d\omega_1 \frac{x_{1,k}}{r_1{}^3}\frac{\partial\epsilon'(\mathbf{r}_1)}{\partial x_{1,i}}\int_V d\omega_2 \frac{x_{2,k}}{r_2{}^3}\frac{\partial\epsilon'(\mathbf{r}_2)}{\partial x_{2,(i)}}\right\rangle$$

$$= \frac{1}{(4\pi)^2\langle\epsilon'^2\rangle}\left\langle\left(\int_V d\omega \frac{x_k}{r^3}\frac{\partial\epsilon'(\mathbf{r})}{\partial x_{(i)}}\right)^2\right\rangle \geq 0. \tag{2.45}$$

On account of the symmetric character of the average depolarizing factor tensor $A_{ij}^{(2)}$, there exist three principal values $A_1^{(2)}, A_2^{(2)}, A_3^{(2)}$, which satisfy

$$A_1^{(2)} + A_2^{(2)} + A_3^{(2)} = 1, \tag{2.46}$$

and

$$A_1^{(2)}, A_2^{(2)}, A_3^{(2)} \geq 0. \tag{2.47}$$

When the principal axes are parallel to the coordinate axes, we can write

$$A_{ij}^{(2)} = A_{(i)}^{(2)}\delta_{ij}. \tag{2.48}$$

The transformation law of $A_{ij}^{(2)}$ shows that

$$A_{ij}^{(2)} = a_{ik}a_{jk}A_{(k)}^{(2)}, \tag{2.49}$$

where

$$a_{ik}\,a_{jk} = a_{ki}a_{kj} = \delta_{ij}. \tag{2.50}$$

The coefficient a_{ij} stands for the direction cosine of the ith coordinate axis with respect to the jth principal axis.

As a special case we consider a statistically isotropic material for which $f(\mathbf{r}_{12})$ is a function of $r_{12} = |\mathbf{r}_{12}|$ alone. In this case, $A_{ij}^{(2)}$ must be an isotropic tensor such that

$$A_{ij}^{(2)} = A^{(2)}\delta_{ij}. \tag{2.51}$$

Equations (2.42) and (2.51) permit us to determine the value of $A^{(2)}$ as

$$A^{(2)} = \tfrac{1}{3}A_{ii}^{(2)} = \tfrac{1}{3}, \tag{2.52}$$

which is equivalent to

$$A_1^{(2)} = A_2^{(2)} = A_3^{(2)} = \tfrac{1}{3}. \tag{2.53}$$

The expression for the effective permittivity ϵ^* of a statistically isotropic material to second order reduces to

$$\epsilon^* = \langle\epsilon\rangle(1 - \tfrac{1}{3}\langle\epsilon'^2\rangle/\langle\epsilon\rangle^2), \tag{2.54}$$

in agreement with the results of Brown,[1] Nedoluha,[3] Landau and Lifshitz,[4] Herring,[5] Beran and Molyneux,[2] and Beran.[8,16]

D. Third-order perturbation term

In this section we shall be concerned with the behavior of the third-order perturbation coefficient $A_{ij}^{(3)}$. Since the normalized three-point correlation function $f(\mathbf{r}_{12},\mathbf{r}_{13})$ may also be regarded as a function of $\mathbf{r}_{21}$ and $\mathbf{r}_{23}$, we have

$$A_{ij}^{(3)} = \frac{1}{(4\pi)^2}\int_V d\omega_{12}\int_V d\omega_{23}\frac{x_{12,i}}{r_{12}{}^3}\frac{x_{23,k}}{r_{23}{}^3}\frac{\partial^2 f(\mathbf{r}_{12},\mathbf{r}_{13})}{\partial x_{12,k}\partial x_{23,j}},$$

$$= \frac{1}{(4\pi)^2}\int_V d\omega_{21}\int_V d\omega_{23}\frac{x_{21,i}}{r_{21}{}^3}\frac{x_{23,k}}{r_{23}{}^3}\frac{\partial^2 g(\mathbf{r}_{21},\mathbf{r}_{23})}{\partial x_{21,k}\partial x_{23,j}}, \tag{2.55}$$

where

$$f(\mathbf{r}_{12},\mathbf{r}_{13}) = g(\mathbf{r}_{21},\mathbf{r}_{23}) = \langle\epsilon'(\mathbf{r}_1)\epsilon'(\mathbf{r}_2)\epsilon'(\mathbf{r}_3)\rangle/\langle\epsilon'^3\rangle. \tag{2.56}$$

In spherical coordinates boundary conditions for $g(r_{21},\theta_{21},\phi_{21};r_{23},\theta_{23},\phi_{23})$ are

$$g(0,\theta_{21},\phi_{21};0,\theta_{23},\phi_{23}) = 1, \tag{2.57a}$$

$$g(0,\theta_{21},\phi_{21};r_{23},\theta_{23},\phi_{23}) = \text{a function of } r_{23},\theta_{23},\phi_{23} \text{ alone}, \tag{2.57b}$$

$$g(r_{21},\theta_{21},\phi_{21};0,\theta_{23},\phi_{23}) = \text{a function of } r_{21},\theta_{21},\phi_{21} \text{ alone}, \tag{2.57c}$$

$$g(\infty,\theta_{21},\phi_{21};r_{23},\theta_{23},\phi_{23}) = g(r_{21},\theta_{21},\phi_{21};\infty,\theta_{23},\phi_{23}) = g(\infty,\theta_{21},\phi_{21};\infty,\theta_{23},\phi_{23}) = 0. \quad (2.57d)$$

Subsidiary assumptions corresponding to Eqs. (2.35a)–(2.35c) are

$$\frac{\partial g(r_{21},\theta_{21},\phi_{21};r_{23},\theta_{23},\phi_{23})}{\partial r_{21}} = o\left(\frac{1}{r_{21}}\right) \quad \text{as } r_{21}\to\infty, \quad (2.58a)$$

$$\frac{\partial g(r_{21},\theta_{21},\phi_{21};r_{23},\theta_{23},\phi_{23})}{\partial r_{23}} = o\left(\frac{1}{r_{23}}\right) \quad \text{as } r_{23}\to\infty, \quad (2.58b)$$

$$\frac{\partial g(\infty,\theta_{21},\phi_{21};r_{23},\theta_{23},\phi_{23})}{\partial\theta_{21}} = \frac{\partial g(\infty,\theta_{21},\phi_{21};r_{23},\theta_{23},\phi_{23})}{\partial\phi_{21}} = 0, \quad (2.58c)$$

$$\frac{\partial g(r_{21},\theta_{21},\phi_{21};\infty,\theta_{23},\phi_{23})}{\partial\theta_{23}} = \frac{\partial g(r_{21},\theta_{21},\phi_{21};\infty,\theta_{23},\phi_{23})}{\partial\phi_{23}} = 0. \quad (2.58d)$$

Proceeding in the same manner as we did in the previous section, we recast Eq. (2.55) as

$$A^{(3)}_{ij} = \frac{1}{(4\pi)^2}\int_V d\omega_{21}\int_V d\omega_{23}\frac{1}{r_{21}}\frac{1}{r_{23}} \times \frac{\partial^4 g(\mathbf{r}_{21},\mathbf{r}_{23})}{\partial x_{21,i}\,\partial x_{21,k}\partial x_{23,k}\partial x_{23,j}}, \quad (2.59)$$

which is obtained by use of partial integration and Gauss' theorem. The above formula demonstrates that

$$A^{(3)}_{ij} = A^{(3)}_{ji}, \quad (2.60)$$

whence $A^{(3)}_{ij}$ has three principal values $A^{(3)}_1, A^{(3)}_2, A^{(3)}_3$, and obeys the transformation law

$$A^{(3)}_{ij} = a_{ik}a_{jk}A^{(3)}_{(k)}. \quad (2.61)$$

Similarly, the relations

$$A^{(n)}_{ij} = A^{(n)}_{ji} \quad (2.62)$$

hold for all n. This result is compatible with the well-established fact that the permittivity tensor of an anisotropic material is symmetric.

When the medium is statistically isotropic, the third-order perturbation coefficient $A^{(3)}_{ij}$ must be of the form

$$A^{(3)}_{ij} = A^{(3)}\delta_{ij} = \tfrac{1}{3}A^{(3)}_{kk}\delta_{ij}. \quad (2.63)$$

The two-point correlation function $g(\mathbf{r}_{21},\mathbf{r}_{23})$ depends only on $|\mathbf{r}_{21}|$, $|\mathbf{r}_{23}|$, and $\mathbf{r}_{21}\cdot\mathbf{r}_{23}$, so that

$$g(\mathbf{r}_{21},\mathbf{r}_{23}) = g(r_{21},r_{23},u), \quad (2.64)$$

where

$$u = \frac{x_{21,i}\,x_{23,i}}{r_{21}r_{23}} = \cos\theta_{21}\cos\theta_{23} + \sin\theta_{21}\sin\theta_{23}\cos(\phi_{21}-\phi_{23}). \quad (2.65)$$

Putting

$$g(r_{21},r_{23},u) = f(r_{21})f(r_{23}) + \Delta g(r_{21},r_{23},u), \quad (2.66)$$

we get

$$A^{(3)} = \frac{1}{9} + \frac{1}{3(4\pi)^2}\int_V d\omega_{21}\int_V d\omega_{23}\frac{x_{21,i}}{r_{21}^3}\frac{x_{23,j}}{r_{23}^3} \times \frac{\partial^2\Delta g(r_{21},r_{23},u)}{\partial x_{21,j}\,\partial x_{23,i}}. \quad (2.67)$$

The quantity $\Delta g(r_{21},r_{23},u)$ satisfies the conditions

$$\Delta g(0,0,u) = \Delta g(\infty,r_{23},u) = \Delta g(r_{21},\infty,u) = \Delta g(\infty,\infty,u) = 0, \quad (2.68a)$$

$$\Delta g(0,r_{23},u) = \Delta g(0,r_{23},0), \quad (2.68b)$$

$$\Delta g(r_{21},0,u) = \Delta g(r_{21},0,0), \quad (2.68c)$$

$$\frac{\partial\Delta g(r_{21},r_{23},u)}{\partial r_{21}} = o\left(\frac{1}{r_{21}}\right) \quad \text{as } r_{21}\to\infty, \quad (2.69a)$$

$$\frac{\partial\Delta g(r_{21},r_{23},u)}{\partial r_{23}} = o\left(\frac{1}{r_{23}}\right) \quad \text{as } r_{23}\to\infty, \quad (2.69b)$$

$$\frac{\partial\Delta g(\infty,r_{23},u)}{\partial u} = \frac{\partial\Delta g(r_{21},\infty,u)}{\partial u} = 0. \quad (2.69c)$$

We want to express $A^{(3)}$ in terms of spherical coordinates. The integral

$$I_{ji}(\mathbf{r}_{21}) = -\frac{1}{4\pi}\int_V d\omega_{23}\frac{x_{23,j}}{r_{23}^3}\frac{\partial\Delta g(r_{21},r_{23},u)}{\partial x_{23,i}} \quad (2.70)$$

appearing in Eq. (2.67) is an isotropic tensor and has the form[22]

$$I_{ji}(\mathbf{r}_{21}) = B(r_{21})\frac{x_{21,j}}{r_{21}}\frac{x_{21,i}}{r_{21}} + C(r_{21})\delta_{ji}. \quad (2.71)$$

Accordingly,

$$I_{ii}(\mathbf{r}_{21}) = B(r_{21}) + 3C(r_{21}) = -\frac{1}{4\pi}\int_0^{2\pi}d\phi_{23}\int_0^{\pi}d\theta_{23}\sin\theta_{23}\int_0^{\infty}dr_{23}\frac{\partial\Delta g}{\partial r_{23}} = \Delta g(r_{21},0,0). \quad (2.72)$$

For the evaluation of $B(r_{21})$ and $C(r_{21})$ choose $\mathbf{r}_{21}$ to lie along the x_3 axis; then

$$u = \cos\theta_{23}, \quad \frac{\partial r_{23}}{\partial x_{23,3}} = \frac{x_{23,3}}{r_{23}} = \cos\theta_{23}, \quad (2.73a)$$

$$\frac{\partial u}{\partial x_{23,3}} = \frac{1}{r_{23}}\left(\frac{x_{21,3}}{r_{21}} - u\frac{x_{23,3}}{r_{23}}\right) = \frac{1-\cos^2\theta_{23}}{r_{23}}. \quad (2.73b)$$

From Eqs. (2.70), (2.71), and (2.73) we obtain

$$I_{33}(\mathbf{r}_{21}) = B(r_{21}) + C(r_{21}) = -\frac{1}{4\pi}\int_0^{2\pi}d\phi_{23}\int_{-1}^{1}du\int_0^{\infty}dr_{23}u\left(u\frac{\partial\Delta g}{\partial r_{23}} + \frac{1-u^2}{r_{23}}\frac{\partial\Delta g}{\partial u}\right) = \frac{1}{3}\Delta g(r_{21},0,0) - \frac{1}{2}\int_0^{\infty}\frac{dr_{23}}{r_{23}}\int_{-1}^{1}du\,u(1-u^2)\frac{\partial g}{\partial u}. \quad (2.74)$$

The solutions of Eqs. (2.72) and (2.74) are

$$B(r_{21}) = -\frac{3}{4}\int_0^\infty \frac{dr_{23}}{r_{23}}\int_{-1}^{1} du\, u(1-u^2)\,\frac{\partial g}{\partial u}, \tag{2.75}$$

$$C(r_{21}) = \frac{1}{3}\Delta g(r_{21},0,0) + \frac{1}{4}\int_0^\infty \frac{dr_{23}}{r_{23}}\int_{-1}^{1} du\, u(1-u^2)\,\frac{\partial g}{\partial u}. \tag{2.76}$$

It is easy to calculate $A^{(3)}$ by employing the values of $B(r_{21})$ and $C(r_{21})$ given above. Differentiation of $I_{ji}(\mathbf{r}_{21})$ with respect to $x_{21,j}$ yields

$$\frac{\partial I_{ji}(\mathbf{r}_{21})}{\partial x_{21,j}} = \left(\frac{dB(r_{21})}{dr_{21}} + \frac{2B(r_{21})}{r_{21}} + \frac{dC(r_{21})}{dr_{21}}\right)\frac{x_{21,i}}{r_{21}}. \tag{2.77}$$

Inserting Eqs. (2.75)–(2.77) into Eq. (2.67), we have, finally,

$$\begin{aligned} A^{(3)} &= \frac{1}{9} - \frac{1}{3\cdot 4\pi}\int_V d\omega_{21}\,\frac{x_{21,i}}{{r_{21}}^3}\,\frac{\partial I_{ji}(\mathbf{r}_{21})}{\partial x_{21,j}} \\ &= \frac{1}{9} - \frac{1}{3}\int_0^\infty dr_{21}\left(\frac{dB(r_{21})}{dr_{21}} + \frac{2B(r_{21})}{r_{21}} + \frac{dC(r_{21})}{dr_{21}}\right) \\ &= \frac{1}{9} + \frac{1}{2}\int_0^\infty \frac{dr_{21}}{r_{21}}\int_0^\infty \frac{dr_{23}}{r_{23}}\int_{-1}^{1} du\, u(1-u^2) \\ &\quad\times \frac{\partial g(r_{21},r_{23},u)}{\partial u}, \end{aligned} \tag{2.78}$$

a result that has already been found by Molyneux.[7] The effective constant ϵ^* to third order is therefore written as

$$\begin{aligned} \epsilon^* &= \langle\epsilon\rangle\left[1 - \frac{1}{3}\frac{\langle\epsilon'^2\rangle}{\langle\epsilon\rangle^2} + \frac{1}{9}\frac{\langle\epsilon'^3\rangle}{\langle\epsilon^3\rangle} + \frac{1}{2}\int_0^\infty \frac{dr_{21}}{r_{21}}\int_0^\infty \frac{dr_{23}}{r_{23}}\right. \\ &\quad\left.\times\int_{-1}^{1} du\, u(1-u^2)\,\frac{\partial g(r_{21},r_{23},u)}{\partial u}\,\frac{\langle\epsilon'^3\rangle}{\langle\epsilon\rangle^3}\right] \\ &= \langle\epsilon\rangle\left[1 - \frac{1}{3}\frac{\langle\epsilon'^2\rangle}{\langle\epsilon\rangle^2} + \frac{1}{9}\frac{\langle\epsilon'^3\rangle}{\langle\epsilon\rangle^3} + \frac{1}{2}\int_0^\infty \frac{dr_{21}}{r_{21}}\int_0^\infty \frac{dr_{23}}{r_{23}}\right. \\ &\quad\left.\times\int_{-1}^{1} du(3u^2-1)g(r_{21},r_{23},u)\,\frac{\langle\epsilon'^3\rangle}{\langle\epsilon\rangle^3}\right]. \end{aligned} \tag{2.79}$$

3. CALCULATIONS BASED ON THE CELL MODEL

A. Definition of cell materials

Miller[18,19] has proposed a cell model to represent a broad class of random multiphase materials. The material volume is assumed to be subdivided into a large number of closed cells within which the material property ϵ is constant. The subdivision of the material space is arbitrary except for fulfilling the following requirements:

(i) The space is completely covered by nonoverlapping cells;
(ii) cells are distributed in a manner such that the material is statistically homogeneous and isotropic;
(iii) the material property ϵ of a cell is statistically independent of the material property of any other cell.

In order to generalize Miller's cell model to statistically anisotropic materials, we must replace the assumption (ii) by

(ii') cells are distributed in a manner such that the material is statistically homogeneous.

A cell material can be defined as a random heterogeneous material that satisfies the above requirements (i), (ii), (iii) or (i), (ii'), (iii). Furthermore, the cell material is called symmetric or asymmetric, according as the following additional requirement is or is not satisfied:

(iv) The conditional probabilities of n points being and n' points not being in the same cell of a particular material, given that one point is in a cell of that material, are the same for each material.

In other words, the term symmetric cell material means a cell material in which the statistical properties of the geometry of all phases are identical. For the asymmetric cell materials, however, the geometry of the cells of different permittivities is dissimilar.

We consider an asymmetric cell material composed of N phases. Let v_l be the volume fraction of the lth phase with property ϵ_l, or equivalently, the probability of the event $\{\epsilon(\mathbf{r}) = \epsilon_l\}$. Then we have

$$\langle\epsilon'(\mathbf{r}_1)\epsilon'(\mathbf{r}_2)\rangle = \sum_{l=1}^{N} v_l\langle\epsilon'(\mathbf{r}_1)\epsilon'(\mathbf{r}_2)|\epsilon(\mathbf{r}_1) = \epsilon_l\rangle, \tag{3.1}$$

where $\langle\epsilon'(\mathbf{r}_1)\epsilon'(\mathbf{r}_2)|\epsilon(\mathbf{r}_1) = \epsilon_l\rangle$ indicates the expectation of $\epsilon'(\mathbf{r}_1)\epsilon'(\mathbf{r}_2)$ under the condition that $\epsilon(\mathbf{r}_1) = \epsilon_l$. It is obvious from the independence assumption (iii) that this conditional expectation takes the values $\epsilon_l'^2 = (\epsilon_l - \langle\epsilon\rangle)^2$ when both points $\mathbf{r}_1$ and $\mathbf{r}_2$ are in the same cell and 0 otherwise. The two-point correlation function is therefore

$$f(\mathbf{r}_{12}) = \frac{\langle\epsilon'(\mathbf{r}_1)\epsilon'(\mathbf{r}_2)\rangle}{\langle\epsilon'^2\rangle} = \frac{1}{\langle\epsilon'^2\rangle}\sum_{l=1}^{N} v_l\epsilon_l'^2 P_l(\mathbf{r}_1,\mathbf{r}_2). \tag{3.2}$$

Here $P_l(\mathbf{r}_1,\mathbf{r}_2)$ denotes the conditional probability that two points $\mathbf{r}_1$ and $\mathbf{r}_2$ lie in the same cell, given that one of the points lies in a cell with property ϵ_l. Substitution of Eq. (3.3) into Eq. (2.30) gives

$$A_{ij}^{(2)} = \frac{1}{\langle\epsilon'^2\rangle}\sum_{l=1}^{N} v_l\epsilon_l'^2 A_{ij}^{(2)}(l), \tag{3.3}$$

where $A_{ij}^{(2)}(l)$ is the average depolarizing factor tensor of the lth phase, namely,

$$A_{ij}^{(2)}(l) = -\frac{1}{4\pi}\int_V d\omega_{12}\,\frac{x_{12,i}}{{r_{12}}^3}\,\frac{\partial P_l(\mathbf{r}_1,\mathbf{r}_2)}{\partial x_{12,j}}. \tag{3.4}$$

The symmetric cell material is a special case of the asymmetric cell material in which the probability $P_l(\mathbf{r}_1,\mathbf{r}_2)$ does not depend on l, so that

$$f(\mathbf{r}_{12}) = P_l(\mathbf{r}_1,\mathbf{r}_2) = P(\mathbf{r}_1,\mathbf{r}_2). \tag{3.5}$$

In this case the two-point correlation function $f(\mathbf{r}_{12})$ may be interpreted as the absolute probability that two points $\mathbf{r}_1$ and $\mathbf{r}_2$ are contained in the same cell. The expression (2.30) for $A_{ij}^{(2)}$ becomes

$$A_{ij}^{(2)} = -\frac{1}{4\pi}\int_V d\omega_{12}\,\frac{x_{12,i}}{{r_{12}}^3}\,\frac{\partial P(\mathbf{r}_1,\mathbf{r}_2)}{\partial x_{12,j}}. \tag{3.6}$$

Comparison of Eqs. (3.4) and (3.6) leads to the following conclusion: Using an averaging formula as described in (3.3), we can estimate the values of $A_{ij}^{(2)}$ for an asymmetric cell material from those for symmetric cell materials. Similar arguments would apply also to higher-order perturbation coefficients. Hence we shall confine ourselves to the symmetric-cell case.

In addition, Eq. (3.6) shows that $A_{ij}^{(2)}$ for a symmetric cell material depends only on the geometry of cells. Important parameters characterizing the cell geometry are shape, size, orientation, and arrangement. For instance, $A_{ij}^{(2)}$ for a symmetric material comprising cells of different shapes is determined by the volume average of $A_{ij}^{(2)}$ over all shapes. If w_m and $A_{ij}^{(2)}[m]$ are the volume fraction and the average depolarizing factor tensor of cells of the mth shape, respectively, we obtain a simple averaging formula

$$A_{ij}^{(2)} = \sum_m w_m A_{ij}^{(2)}[m]. \tag{3.7}$$

In Sec. 3B it will be verified that $A_{ij}^{(2)}$ is independent of the cell size. Owing to the independence assumption (iii), the influence of relative positions of cells upon $A_{ij}^{(2)}$ need not be taken into account.

B. Second-order perturbation term

We first treat a symmetric cell material composed of cells of uniform shape, size, and orientation. In the preceding section we have pointed out that the two-point correlation function $f(\mathbf{r}_{12})$ is equal to the probability $P(\mathbf{r}_1, \mathbf{r}_2)$ that two points $\mathbf{r}_1$ and $\mathbf{r}_2$ are in the same cell. Denote by M_1 the cell whose center (of mass) is located at $\mathbf{r}_1$. Then centers of cells including the point $\mathbf{r}_1$ fall into a region $\bar{M}_1$ that results from the inversion of M_1 about $\mathbf{r}_1$. Similarly, centers of cells including the point $\mathbf{r}_2$ form another region $\bar{M}_2$ which can be transformed into $\bar{M}_1$ by means of a translation. Because of the statistical homogeneity of the medium, $P(\mathbf{r}_1, \mathbf{r}_2)$ may be regarded as the ratio of the volume of the intersection $\bar{M}_1 \cap \bar{M}_2$ to that of $\bar{M}_1$ or $\bar{M}_2$. If the cell volume is v,

$$P(\mathbf{r}_1, \mathbf{r}_2) = \int_{\bar{M}_1 \cap \bar{M}_2} d\omega_{20} \Big/ \int_{\bar{M}_2} d\omega_{20} = \frac{1}{v} \int_{M_1 \cap M_2} d\omega_{20}. \tag{3.8}$$

Especially, for spherical cells of radius ρ,

$$P(\mathbf{r}_1, \mathbf{r}_2) = \begin{cases} 1 - 3r_{12}/4\rho + r_{12}^3/16\rho^3 & \text{for } r_{12} \le 2\rho \\ 0 \quad \text{for } r_{12} \ge 2\rho. \end{cases} \tag{3.9}$$

We introduce a characteristic function $I(\mathbf{r}_{10})$ such that $I(\mathbf{r}_{10}) = 1$ or 0 according to whether $\mathbf{r}_0 \in M_1$ or not. In terms of $I(\mathbf{r}_{10})$, we rewrite Eq. (3.8) as

$$P(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{v} \int_v d\omega_{20} I(\mathbf{r}_{10}). \tag{3.10}$$

Thus, Eq. (3.6) leads to

$$\begin{aligned} A_{ij}^{(2)} &= -\frac{1}{4\pi v} \int_V d\omega_{12} \int_v d\omega_{20} \frac{x_{12,i}}{r_{12}^3} \frac{\partial I(\mathbf{r}_{10})}{\partial x_{12,j}} \\ &= \frac{1}{4\pi v} \int_V d\omega_{10} \int_v d\omega_{20} \frac{\partial}{\partial x_{10,i}} \left(\frac{1}{r_{12}}\right) \frac{\partial I(\mathbf{r}_{10})}{\partial x_{10,j}}. \end{aligned} \tag{3.11}$$

Integration by parts yields

$$\begin{aligned} A_{ij}^{(2)} &= -\frac{1}{v} \int_v d\omega_{10} \frac{\partial^2}{\partial x_{10,i}\, \partial x_{10,j}} \int_v \frac{d\omega_{20}}{4\pi r_{12}} \\ &= -\frac{1}{v} \int_v d\omega_{10} \int_v d\omega_{20} \frac{\partial^2}{\partial x_{10,i}\, \partial x_{10,j}} \left(\frac{1}{4\pi r_{12}}\right) \\ &= \frac{1}{v} \int_v d\omega_{10} \int_v d\omega_{20} \frac{\partial^2}{\partial x_{10,i}\, \partial x_{20,j}} \left(\frac{1}{4\pi r_{12}}\right). \end{aligned} \tag{3.12}$$

This result shows that the second-order perturbation coefficient $A_{ij}^{(2)}$ coincides with the so-called magnetometric demagnetization tensor.[23,24]

The point-function demagnetization tensor $L_{ij}(\mathbf{r}_{10})$ is defined as follows:[24,25]

$$\begin{aligned} L_{ij}(\mathbf{r}_{10}) &= -\frac{\partial^2 \Phi(\mathbf{r}_{10})}{\partial x_{10,i}\, \partial x_{10,j}} \\ &= \int_v d\omega_{20} \frac{\partial^2}{\partial x_{10,i}\, \partial x_{20,j}} \left(\frac{1}{4\pi r_{12}}\right), \end{aligned} \tag{3.13}$$

where

$$\Phi(\mathbf{r}_{10}) = \int_v \frac{d\omega_{20}}{4\pi r_{12}}. \tag{3.14}$$

Clearly, $\Phi(\mathbf{r}_{10})$ given above is the potential of a uniformly charged particle of volume v with unit density, satisfying Poisson's equation in the interior and Laplace's equation in the exterior. Therefore, we obtain

$$\Delta\Phi(\mathbf{r}_{10}) = \frac{\partial^2 \Phi(\mathbf{r}_{10})}{\partial x_{10,i}^2} = \begin{cases} -1 & \text{if } \mathbf{r}_0 \in M_1, \\ 0 & \text{otherwise,} \end{cases} \tag{3.15}$$

whence

$$L_{ii}(\mathbf{r}_{10}) = 1. \tag{3.16}$$

The magnetometric demagnetization tensor is considered to be the volume average of the point-function demagnetization tensor; that is,

$$A_{ij}^{(2)} = \frac{1}{v} \int_v d\omega_{10} L_{ij}(\mathbf{r}_{10}) = -\frac{1}{v} \int_v d\omega_{10} \frac{\partial^2 \Phi(\mathbf{r}_{10})}{\partial x_{10,i}\, \partial x_{10,j}}. \tag{3.17}$$

Now we can confirm that the average depolarizing factor tensor $A_{ij}^{(2)}$ depends only on the shape of cells and not on their size. Assume that a cell M_1' with volume v' is geometrically similar and similarly oriented to M_1. Applying the formula (3.12) and changing coordinates, we see that

$$\begin{aligned} A_{ij}^{(2)'} &= \frac{1}{v'} \int_{v'} d\omega_{10} \int_{v'} d\omega_{20} \frac{\partial^2}{\partial x_{10,i}\, \partial x_{20,j}} \left(\frac{1}{4\pi r_{12}}\right) \\ &= \frac{1}{v} \int_v d\omega_{10}' \int_v d\omega_{20}' \frac{\partial^2}{\partial x_{10,i}'\, \partial x_{20,j}'} \left(\frac{1}{4\pi r_{12}'}\right) \\ &= A_{ij}^{(2)}, \end{aligned} \tag{3.18}$$

which completes the proof. Generally speaking, $A_{ij}^{(2)}$ is not affected by the size distribution of cells, even when the medium consists of cells of varying size.

Next, let us discuss the case where identical cells are randomly oriented. In this case the cell material is statistically isotropic, so that $A_{ij}^{(2)}$ must be $\frac{1}{3}\delta_{ij}$ as shown in Eq. (2.52). In fact, Eq. (2.52) can be derived from the governing equations of $A_{ij}^{(2)}$ for uniformly

oriented cells. It goes without saying that Eq. (2.49) describes the orientation dependence of $A_{ij}^{(2)}$. Averaging Eq. (2.49) over all possible orientations, we have

$$A_{ij}^{(2)} = \langle a_{ik} a_{jk} \rangle A_{(k)}^{(2)}. \tag{3.19}$$

For randomly oriented cells,

$$\langle a_{i1} a_{j1} \rangle = \langle a_{i2} a_{j2} \rangle = \langle a_{i3} a_{j3} \rangle, \tag{3.20}$$

which, together with the relation $\langle a_{ik} a_{jk} \rangle = \delta_{ij}$, gives

$$A_{ij}^{(2)} = \tfrac{1}{3}(A_1^{(2)} + A_2^{(2)} + A_3^{(2)})\delta_{ij} = \tfrac{1}{3}\delta_{ij}, \tag{3.21}$$

as was to be proved.

C. Third-order perturbation term

To seek $A_{ij}^{(3)}$ for a symmetric cell material in which identical cells are uniformly oriented, we investigated the behavior of the three-point correlation function $f(\mathbf{r}_{12}, \mathbf{r}_{13}) = g(\mathbf{r}_{21}, \mathbf{r}_{23})$. The independence hypothesis asserts that $\langle \epsilon'(\mathbf{r}_1)\epsilon'(\mathbf{r}_2)\epsilon'(\mathbf{r}_3) \rangle$ vanishes whenever all the points $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ are not in the same cell. Accordingly,

$$f(\mathbf{r}_{12}, \mathbf{r}_{13}) = g(\mathbf{r}_{21}, \mathbf{r}_{23}) = P(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3), \tag{3.22}$$

where $P(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$ stands for the probability that three points $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ fall into the same cell. Analogous to the derivation of Eq. (3.10), we have

$$P(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) = \frac{1}{v}\int_v d\omega_{20} I(\mathbf{r}_{10}) I(\mathbf{r}_{30}). \tag{3.23}$$

The product of the two characteristic functions $I(\mathbf{r}_{10}) \times I(\mathbf{r}_{30})$ assumes the values 1 when $\mathbf{r}_0 \in M_1 \cap M_3$ and 0 otherwise.

Combining Eqs. (2.55), (3.22), and (3.23), we can write

$$\begin{aligned} A_{ij}^{(3)} &= \frac{1}{(4\pi)^2 v}\int_V d\omega_{21}\int_V d\omega_{23}\int_v d\omega_{20} \\ &\quad \times \frac{x_{21,i}}{r_{21}^3}\frac{x_{23,k}}{r_{23}^3}\frac{\partial^2 I(\mathbf{r}_{10}) I(\mathbf{r}_{30})}{\partial x_{21,k}\partial x_{23,j}} \\ &= \frac{1}{(4\pi)^2 v}\int_V d\omega_{10}\int_V d\omega_{30}\int_v d\omega_{20} \\ &\quad \times \frac{\partial(1/r_{21})}{\partial x_{10,i}}\frac{\partial(1/r_{23})}{\partial x_{30,k}}\frac{\partial I(\mathbf{r}_{10})}{\partial x_{10,k}}\frac{\partial I(\mathbf{r}_{30})}{\partial x_{30,j}}. \end{aligned} \tag{3.24}$$

By partial integration with respect to the variables $x_{10,k}$ and $x_{30,j}$, Eq. (3.24) becomes

$$\begin{aligned} A_{ij}^{(3)} &= \frac{1}{v}\int_v d\omega_{10}\int_v d\omega_{20}\int_v d\omega_{30} \\ &\quad \times \frac{\partial^2}{\partial x_{10,i}\partial x_{10,k}}\left(\frac{1}{4\pi r_{21}}\right)\frac{\partial^2}{\partial x_{30,k}\partial x_{30,j}}\left(\frac{1}{4\pi r_{23}}\right) \\ &= \frac{1}{v}\int_v d\omega_{10}\int_v d\omega_{20}\int_v d\omega_{30} \\ &\quad \times \frac{\partial^2}{\partial x_{20,i}\partial x_{20,k}}\left(\frac{1}{4\pi r_{21}}\right)\frac{\partial^2}{\partial x_{20,k}\partial x_{20,j}}\left(\frac{1}{4\pi r_{23}}\right). \end{aligned} \tag{3.25}$$

It turns out that the third-order perturbation coefficient $A_{ij}^{(3)}$ is related to the point-function demagnetization tensor $L_{ij}(\mathbf{r}_{20})$ by

$$\begin{aligned} A_{ij}^{(3)} &= \frac{1}{v}\int_v d\omega_{20}\frac{\partial^2\Phi(\mathbf{r}_{20})}{\partial x_{20,i}\partial x_{20,k}}\frac{\partial^2\Phi(\mathbf{r}_{20})}{\partial x_{20,k}\partial x_{20,j}} \\ &= \frac{1}{v}\int_v d\omega_{20} L_{ik}(\mathbf{r}_{20}) L_{kj}(\mathbf{r}_{20}). \end{aligned} \tag{3.26}$$

We shall deal further with the statistically isotropic case where congruent cells are oriented at random. In exactly the same way as we derived Eq. (3.21),

$$\begin{aligned} A_{ij}^{(3)} &= A^{(3)}\delta_{ij} = \tfrac{1}{3}(A_1^{(3)} + A_2^{(3)} + A_3^{(3)})\delta_{ij} \\ &= \frac{1}{3v}\delta_{ij}\int_v d\omega_{20} L_{kk}^2(\mathbf{r}_{20}). \end{aligned} \tag{3.27}$$

Using the algebraic identity

$$\begin{aligned} 3(L_{11}^2 + L_{22}^2 + L_{33}^2) &= (L_{11} + L_{22} + L_{33})^2 \\ &\quad + (L_{11} - L_{22})^2 + (L_{22} - L_{33})^2 + (L_{33} - L_{11})^2, \end{aligned} \tag{3.28}$$

we get from Eqs. (3.16) and (3.27)

$$A^{(3)} > \frac{1}{3v}\int_v d\omega_{20}[L_{11}^2(\mathbf{r}_{20}) + L_{22}^2(\mathbf{r}_{20}) + L_{33}^2(\mathbf{r}_{20})] \geq \tfrac{1}{9}. \tag{3.29}$$

Miller[18] has already conjectured that for a quantity G corresponding to $A^{(3)}$

$$\tfrac{1}{9} \leq G \leq \tfrac{1}{3}. \tag{3.30}$$

In a subsequent paper it will be proved that $A^{(3)}$ is equivalent to G introduced by Miller.

D. Ellipsoidal cells

The depolarizing factor tensors of an ellipsoid and its degenerate shapes (spheroid, sphere, thin plate, long rod, etc.) are well defined and unique. For these geometries the point-function demagnetization tensor is constant throughout the interior of the body, so that it need not be distinguished from the magnetometric demagnetization tensor. The eigenvalues of this tensor are the depolarizing or demagnetizing factors that have been extensively discussed and tabulated.[26,27] It follows from Eqs. (3.17) and (3.26) that for uniformly oriented ellipsoidal cells

$$A_{ij}^{(2)} = L_{ij}, \tag{3.31}$$

$$A_{ij}^{(3)} = L_{ik} L_{kj}. \tag{3.32}$$

In words, the second-order and the third-order perturbation coefficients equal the depolarizing factor tensor and its square, respectively.

When axes of ellipsoidal cells are aligned parallel with the coordinate axes, Eqs. (3.31) and (3.32) lead to

$$A_{ij}^{(2)} = A_{(i)}^{(2)}\delta_{ij} = L_{(i)}\delta_{ij}, \tag{3.33}$$

$$A_{ij}^{(3)} = A_{(i)}^{(3)}\delta_{ij} = L_{(i)}^2\delta_{ij}, \tag{3.34}$$

where L_i designate the depolarizing factors of the ellipsoid. According to potential theory,[28] the potential at an internal point $\mathbf{r}_0$ of an ellipsoid of unit uniform density,

with center $\mathbf{r}_1$ and semiaxes a_i in the x_i directions, may be expressed in the form

$$\Phi(\mathbf{r}_{10}) = \frac{a_1a_2a_3}{4}\int_0^\infty \left(1 - \sum_{i=1}^{3}\frac{x_{10,i}^2}{s + a_i^2}\right) \times \frac{ds}{[(s + a_1^2)(s + a_2^2)(s + a_3^2)]^{1/2}}. \quad (3.35)$$

Consequently,

$$A_i^{(2)} = L_i = -\frac{\partial\Phi(\mathbf{r}_{10})}{\partial x_{10,(i)}^2} = \frac{a_1a_2a_3}{2} \times \int_0^\infty \frac{ds}{(s + a_i^2)[(s + a_1^2)(s + a_2^2)(s + a_3^2)]^{1/2}}. \quad (3.36)$$

As is well known, the integrals in the formulas (3.36) for the depolarizing factors of an ellipsoid are in general elliptic integrals. For a spheroid these integrals reduce to inverse circular or logarithmic functions. Let $a_3 = a$ be the polar semiaxis and $a_1 = a_2 = c$ be the equatorial semiaxis, and put

$$L_3 = L, \quad L_1 = L_2 = \tfrac{1}{2}(1 - L). \quad (3.37)$$

Then, for an oblate spheroid with the axial ratio $\alpha = a/c \leq 1$,[26]

$$L = \frac{1}{1 - \alpha^2}\left[1 - \frac{\alpha}{(1 - \alpha^2)^{1/2}}\arccos\alpha\right] \geq \tfrac{1}{3}. \quad (3.38)$$

For a prolate spheroid with $\alpha = a/c \geq 1$,[26]

$$L = \frac{1}{\alpha^2 - 1}\left\{\frac{\alpha}{(\alpha^2 - 1)^{1/2}}\log[\alpha + (\alpha^2 - 1)^{1/2}] - 1\right\} \leq \tfrac{1}{3}. \quad (3.39)$$

As the axial ratio α increases from 0 to ∞, the depolarizing factor L decreases monotonically from 1 to 0.

Several special cases are of interest. When $\alpha = 0$, the spheroid degenerates into a plate which is of infinite extent in two dimensions. In this case, the expression (3.38) yields

$$L = 1, \quad A_1^{(2)} = A_2^{(2)} = 0, \quad A_3^{(2)} = 1. \quad (3.40)$$

Considering that any component of $A_i^{(2)}$ along the axis of infinite length must vanish under usual conditions, one can derive the above equations immediately from the fundamental relation (2.46). If $\alpha = 1$, the cell becomes a sphere and

$$L = \tfrac{1}{3}, \quad A_1^{(2)} = A_2^{(2)} = A_3^{(2)} = \tfrac{1}{3}. \quad (3.41)$$

This is a natural consequence of the isotropy of spheres. As $\alpha \to \infty$, the spheroid degenerates into an infinitely long rod of circular cross section, for which

$$L = 0, \quad A_1^{(2)} = A_2^{(2)} = \tfrac{1}{2}, \quad A_3^{(2)} = 0. \quad (3.42)$$

Unless the rod is axially symmetric, however, $A_1^{(2)}$ and $A_2^{(2)}$ are not necessarily equal to each other. Actually, for a long rod of elliptic cross section with $a_1 \neq a_2$ and $a_3 \to \infty$,

$$A_1^{(2)} = \frac{a_2}{a_1 + a_2}, \quad A_2^{(2)} = \frac{a_1}{a_1 + a_2}, \quad A_3^{(2)} = 0. \quad (3.43)$$

Now we can evaluate $A_{ij}^{(3)}$ for randomly oriented spheroids. Equations (3.27) and (3.37) lead to

$$A^{(3)} = \tfrac{1}{3}\left[2\left(\frac{1 - L}{2}\right)^2 + L^2\right] = \frac{3L^2 - 2L + 1}{6}, \quad (3.44)$$

from which one may easily deduce the bounds

$$\tfrac{1}{9} \leq A^{(3)} \leq \tfrac{1}{3}. \quad (3.45)$$

Even when the spheroid is assumed to be oblate, namely, $\frac{1}{3} \leq L \leq 1$, the range of $A^{(3)}$ is given by the inequalities (3.45). On the other hand, when the spheroid is prolate, we have $0 \leq L \leq \frac{1}{3}$, so that

$$\tfrac{1}{9} \leq A^{(3)} \leq \tfrac{1}{6}. \quad (3.46)$$

These results do not agree with the bounds derived by Miller,[18] who has obtained $\frac{1}{9} \leq A^{(3)} \leq \frac{1}{6}$ for oblate spheroids and $\frac{1}{6} \leq A^{(3)} \leq \frac{1}{3}$ for prolate spheroids.

In particular, $A^{(3)} = \frac{1}{3}$ corresponds to the shape of plates ($L = 1$), $A^{(3)} = \frac{1}{9}$ to the spherical shape ($L = \frac{1}{3}$), and $A^{(3)} = \frac{1}{6}$ to the shape of rods of circular cross section ($L = 0$) or the shape of oblate spheroids with $L = \frac{2}{3}$. Note that the cell shape is not uniquely determined from the value of $A^{(3)}$. If the long rods with $a_3 \to \infty$ have elliptic cross section, $A^{(3)}$ becomes

$$A^{(3)} = \frac{1}{3}\left[\left(\frac{a_2}{a_1 + a_2}\right)^2 + \left(\frac{a_1}{a_1 + a_2}\right)^2\right] = \frac{1}{3}\left[1 - \frac{2a_1a_2}{(a_1 + a_2)^2}\right], \quad (3.47)$$

whence we find

$$\tfrac{1}{6} \leq A^{(3)} \leq \tfrac{1}{3}. \quad (3.48)$$

It may be shown that the above inequalities hold also for randomly oriented rods of arbitrary cross section.

E. Nonellipsoidal cells

In nonellipsoidal bodies, the point-function demagnetization tensor varies with position, and it is necessary to define the magnetometric demagnetization tensor as the volume average. Hence the perturbation coefficients of nonellipsoidal cells are extremely difficult to compute in contrast to the case of ellipsoids. For circular cylinders or rectangular parallelepipeds, however, both the point-function demagnetization tensor and magnetometric demagnetization tensor have so far been calculated.[29-31] By way of example we shall present the eigenvalues of $A_{ij}^{(2)}$ for cells of rectangular parallelepipeds.

Let us consider a cell material composed of rectangular parallelepipeds with edge lengths $2b_i$ in the x_i directions. Then the principal axes of $A_{ij}^{(2)}$ are parallel to the edges of the ractangular parallelepipeds. Direct calculation shows[29]

$$L_3(\mathbf{r}_{10}) = -\frac{\partial^2\Phi(\mathbf{r}_{10})}{\partial x_{10,3}^2} = \frac{1}{4\pi} \times \sum_{e_1,e_2,e_3=\pm 1} \operatorname{arccot}\left\{\frac{(b_3 + e_3x_{10,3})\left[\sum_{i=1}^{3}(b_i + e_ix_{10,i})^2\right]^{1/2}}{(b_1 + e_1x_{10,1})(b_2 + e_2x_{10,2})}\right\}. \quad (3.49)$$

The principal value of the inverse cotangent should be taken. From Eq. (3.17) we obtain

$$A_3^{(2)} = \frac{1}{v}\int_v d\omega_{10} L_3(\mathbf{r}_{10})$$
$$= \frac{1}{4\pi b_1 b_2 b_3}\int_0^{2b_1} d\xi_1 \int_0^{2b_2} d\xi_2 \int_0^{2b_3} d\xi_3$$
$$\times \operatorname{arccot} \frac{\xi_3(\xi_1^2 + \xi_2^2 + \xi_3^2)^{1/2}}{\xi_1\xi_2}. \tag{3.50}$$

The result of integration is as follows[31]:

$$A_3^{(2)} = \frac{1}{\pi}\left\{2 \operatorname{arccot} \frac{[1 + k^2(\beta_1,\beta_2)]^{1/2}}{\beta_1\beta_2} - \frac{\beta_1^2 - 1}{2\beta_1}\right.$$
$$\times K([\beta_1^2 + 1]^{1/2}, \beta_2) - \frac{\beta_2^2 - 1}{2\beta_2} K([\beta_2^2 + 1]^{1/2}, \beta_1)$$
$$+ \frac{\beta_1 K(\beta_1,\beta_2)}{2} + \frac{\beta_2 K(\beta_2,\beta_1)}{2} - \frac{K(1,\beta_2)}{2\beta_1}$$
$$- \frac{K(1,\beta_1)}{2\beta_2} - \frac{\beta_2 k(1,\beta_2)}{\beta_1} + \frac{\beta_2^2}{3\beta_1} + \frac{2k^3(1,\beta_2)}{3\beta_1\beta_2}$$
$$- \frac{\beta_1 k(1,\beta_1)}{\beta_2} + \frac{\beta_1^2}{3\beta_2} + \frac{2k^3(1,\beta_1)}{3\beta_1\beta_2} - \frac{2}{3\beta_1\beta_2}$$
$$- \frac{k^3(\beta_1,\beta_2)}{3\beta_1\beta_2} + \frac{k^2(\beta_1,\beta_2)}{\beta_1\beta_2}[1 + k^2(\beta_1,\beta_2)]^{1/2}$$
$$\left. - \frac{2}{3\beta_1\beta_2}[1 + k^2(\beta_1,\beta_2)]^{3/2}\right\}, \tag{3.51}$$

where

$$\beta_1 = b_1/b_3, \quad \beta_2 = b_2/b_3, \tag{3.52a}$$

$$k^2(\xi,\eta) = \xi^2 + \eta^2, \tag{3.52b}$$

$$K(\xi,\eta) = \log\frac{k(\xi,\eta) + \eta}{k(\xi,\eta) - \eta}. \tag{3.52c}$$

[1]W. F. Brown Jr., J. Chem. Phys. 23, 1514 (1955).
[2]M. Beran and J. Molyneux, Nuovo Cimento **30,** 1406 (1963).
[3]A. Nedoluha, Z. Phys. **148,** 248 (1957).
[4]L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media* (Pergamon, New York, 1960), p. 45.
[5]C. Herring, J. Appl. Phys. **31,** 1939 (1960).
[6]S. Prager, J. Chem. Phys. **33,** 122 (1960).
[7]J. Molyneux, Ph. D. thesis, University of Pennsylvania, 1964; see also M. J. Beran, *Statistical Continuum Theories* (Interscience, New York, 1968), p. 232.
[8]M. Beran, Nuovo Cimento Suppl. **3,** 448 (1965).
[9]V. M. Finkel'berg, Zh. Eksp. Teor. Fiz. **46,** 725 (1964) [Sov. Phys.-JETP **19,** 494 (1964)].
[10]V. M. Finkel'berg, Zh. Tekh. Fiz. **34,** 509 (1964) [Sov. Phys.-Tech. Phys. **9,** 396 (1964)].
[11]Yu. A. Ryzhov, V. V. Tamoĭkin, and V. I. Tatarskiĭ, Zh. Eksp. Teor. Fiz. **48,** 656 (1965) [Sov. Phys.-JETP **21,** 433 (1965)].
[12]J. B. Keller and F. C. Karal Jr., J. Math. Phys. **7,** 661 (1966).
[13]Yu. A. Ryzhov, Zh. Eksp. Teor. Fiz. **55,** 567 (1968) [Sov. Phys.-JETP **28,** 296 (1969)].
[14]F. G. Bass, G. A. Begiashvili, and Yu. S. Monin, Fiz. Tverd. Tela **10,** 3553 (1968) [Sov. Phys.-Solid State **10,** 2822 (1969)].
[15]W. F. Brown Jr., Trans. Soc. Rheol. **9,** 357 (1965).
[16]M. Beran, Nuovo Cimento **38,** 771 (1965).
[17]M. Beran and J. Molyneux, Q. Appl. Math. **24,** 107 (1966).
[18]M. N. Miller, J. Math. Phys. **10,** 1988 (1969).
[19]M. N. Miller, J. Math. Phys. **10,** 2005 (1969); J. Math. Phys. **12,** 1057 (1971).
[20]M. J. Beran and R. Silnutzer, J. Compos. Mater. **5,** 246 (1971).
[21]M. J. Beran, *Statistical Continuum Theories* (Interscience, New York, 1968), p. 210.
[22]See, for instance, G. Batchelor, *The Theory of Homogeneous Turbulence* (Cambridge U. P., London, 1953), p. 40.
[23]W. F. Brown Jr., *Magnetostatic Principles in Ferromagnetism* (North-Holland, Amsterdam, 1962), p. 49.
[24]R. Moskowitz and E. Della Torre, IEEE Trans. Mag. **2,** 739 (1966).
[25]E. Schlömann, J. Appl. Phys. **33,** 2825 (1962).
[26]E. C. Stoner, Philos. Mag. **36,** 803 (1945).
[27]J. A. Osborn, Phys. Rev. **67,** 351 (1945).
[28]See, for instance, O. D. Kellogg, *Foundations of Potential Theory* (Springer, Berlin, 1967), p. 192.
[29]R. I. Joseph and E. Schlömann, J. Appl. Phys. **36,** 1579 (1965).
[30]R. I. Joseph, J. Appl. Phys. **37,** 4639 (1966).
[31]R. I. Joseph, J. Appl. Phys. **38,** 2405 (1967).

A computable expansion for multiparticle propagators*

D. J. Galas

Department of Applied Science, University of California, and U. S. Air Force, Livermore, California

C. A. Jensen and H. L. Sahlin

Lawrence Livermore Laboratory, University of California, Livermore, California 94550
(Received 26 September 1972)

A formulation of the propagator for a quantum system is constructed to simplify computation for multiparticle systems. The propagator is manifested in terms of any complete set on the Hilbert space of physical states of the system. In the short-time limit a particularly simple matrix expression is obtained using an energy eigenstate basis. This formulation is shown to contain all orders of perturbation theory in an approximate form. Identical particle exchange is handled in a straightforward manner in which only the number of terms, but not the complexity, of expressions is affected. The number of terms is held to manageable proportions by approximating multiple-particle exchange with repeated operations containing only two-particle exchange. Three-particle exchange is represented exactly by this approximation in the limit of infinitesimal time steps. An argument is presented that indicates that this is also true for an arbitrary particle number.

1. INTRODUCTION

The question of computing the properties of a quantum system is posed most generally as the problem of determining the propagator, or Green's function of the equation of motion. This determination may be approached from various points of view, but the most direct method is based on the Feynman functional integral formulation of quantum mechanics.[1] The relation between the path integral expression and the corresponding differential equation has been extensively explored.[2,3]

The Wiener integral may be regarded as the "imaginary time" analog of the Feynman integral, and the two integrals stem from differential equations with identical eigenstate solutions. Wiener and Wiener-like functional integrals have been directly evaluated numerically for a number of simple cases.[4–6] However, for the Feynman integral, its elegant simplicity and intuitive appeal are not reflected in ease of computation for any systems but the free particle and harmonic oscillator. In these two cases the Feynman integral can be evaluated analytically. For essentially all other problems, it has so far resisted both analytic and numerical evaluation.

The propagator formulation proposed here provides an alternative to the functional integral expression, while retaining much of its direct simplicity. We were led to this formulation by the fact that the propagator may be manifested in terms of any complete set on the Hilbert space of physical states of a system. By manipulating the limiting forms of the propagator for small times and applying the semigroup composition law for the propagator, we arrive at a particularly simple expression when we choose an energy eigenstate basis.

This formulation appears to be intermediate between the Feynman and Schrödinger points of view. Our intention is to provide a framework in which it may be possible to perform the *ab initio* calculation of quantum systems in the fashion promised, but yet unfulfilled, by direct functional integration.

For the purpose of practical calculation, two principal advantages are promised by the proposed formalism. It retains those characteristics of the Feynman integral that make time evolution and identical particle exchange easy to handle.[6] Also, the ploy of computation in "imaginary time" (or the equivalent, reciprocal temperature) allows the ground state within the symmetry class of a given propagator to be directly extracted.[5,6] For example, if the symmetries of the first n eigenstates are known and are distinct, we may obtain these n eigenstates simply by this method. An additional advantage is that the formalism allows one to use available physical insight by employing basis states that closely approximate the system. In this regard, the formalism is akin to approaches stemming directly from the Schrödinger equation, but it differs strongly in that this insight is not necessary.

An analysis of this formalism for the time evolution operator in terms of its relation to standard perturbation theory expressions is useful. We find that this formalism contains all orders of approximation in some form. As a consequence, it has the potential for handling large "perturbing" Hamiltonians.

Because exchange may be introduced into a calculation at each infinitesimal time step, the time evolution operators may be approximated by operators involving only two particle exchange. In this approximation, multiple particle exchange occurs naturally over several infinitesimal time steps.

The major points alluded to above and the development of the formalism are taken up in the sections below in the following order: development and discussion of the fundamental relations including the energy representation; relationship of this formalism to perturbation theory; relationship to the interaction picture and introduction of identical particle exchange. In Appendix A we discuss the generation of full permutation symmetries by successive two particle exchange.

2. GENERAL FORMALISM

A quantum system may be described in terms of its unitary time evolution operator such that

$$|\psi(t)\rangle = \hat{U}(t, t_0)|\psi(t_0)\rangle \tag{2.1}$$

and, therefore, from the Schrödinger equation,

$$\hat{U}(t, t_0) = 1 - i/\hbar \int_{t_0}^{t} \hat{H}(\tau)\hat{U}(\tau, t_0)d\tau \,. \tag{2.2}$$

To realize a manifestation of $\hat{U}$ as an integral operator in a function space, we consider that the physical Hilbert space, which has all physically permissible $|\psi(t)\rangle$ as

elements, is spanned by a complete set of elements $\{|\alpha\rangle\}$. This set may be parametrized by a discrete or continuous set of complex numbers $\{\alpha\}$. Integrals defined over α are henceforth to be understood in a general sense: as either sums or integrals as is appropriate to the particular set $\{\alpha\}$. The integral representation of $\hat{U}$ in terms of α is then

$$\langle\alpha'|\psi(t)\rangle = \int d\alpha\,\langle\alpha'|\hat{U}(t,t_0)|\alpha\rangle\langle\alpha|\psi(t_0)\rangle, \tag{2.3}$$

where the integral operator is now over the new Hilbert space of functions on the parameter set $\{\alpha\}$. We will refer here to the kernel of Eq. (2.3) as the propagator:

$$K(\alpha,\alpha',t,t_0) \equiv \langle\alpha'|\hat{U}(t,t_0)|\alpha\rangle. \tag{2.4}$$

When the complete set employed is $\{|x\rangle\}$, the configuration space states, we recover the familiar form

$$\begin{aligned} K(\mathbf{x}_1,\mathbf{x}_2,t,t_0) &= \left\langle \mathbf{x}_2 \middle| \exp\left(-\frac{i}{\hbar}\int_{t_0}^{t}\hat{H}(\tau)\,d\tau\right)\middle|\mathbf{x}_1\right\rangle \\ &= \exp\left(-\frac{i}{\hbar}\int_{t_0}^{t}\hat{H}(\mathbf{x}_2)\,d\tau\right)\delta(\mathbf{x}_2-\mathbf{x}_1) \end{aligned} \tag{2.5}$$

(where $\hat{H}(\mathbf{x}_2)$ acts only on the $\mathbf{x}_2$ coordinate) that may easily be cast into the usual functional integral form[1]

$$K(\mathbf{x}_1,\mathbf{x}_2,t,t_0) = A\int \exp\left(\frac{i}{\hbar}\int_{t_0}^{t}\mathcal{L}\,d\tau\right)\mathfrak{D}(\mathbf{x}), \tag{2.6}$$

with $\mathcal{L}$ the classical Lagrangian, A a constant, and $\int\mathfrak{D}(\mathbf{x})$ indicating functional integration over all "paths" from $\mathbf{x}_1$ to $\mathbf{x}_2$. Another choice, leading to a functional integral expression also having a straightforward relation to classical mechanics, is the continuous set of overcomplete coherent states $\{|z\rangle\}$, studied in their application to the propagator by Klauder[7] and Schweber,[8] and in their generalization to arbitrary "continuous representations" by Klauder.[9]

If the interval $(t-t_0)$ is divided into N equal small intervals τ, the composition law

$$\hat{U}(t,t_0) = \hat{U}(t,t_1)\,\hat{U}(t_1,t_0)$$

permits us to write

$$K(\alpha_0,\alpha_N,t,t_0) = \int d\alpha_1\int d\alpha_2\cdots\int d\alpha_{N-1}\prod_{j=0}^{N-1} K(\alpha_j,\alpha_{j+1},\tau). \tag{2.7}$$

But $K(\alpha_j,\alpha_{j+1},\tau)$, for small τ, may be replaced by a simplified approximate form that effectively transforms Eq. (2.7) into a limiting multiple Riemann integral form for the functional integral:

$$K(\alpha_0,\alpha_N,t,t_0) = \lim_{\substack{\tau\to0\\N\to\infty\\(N\tau=t-t_0)}} \int d\alpha_1\int d\alpha_2\cdots\int d\alpha_{N-1}\prod_{j=0}^{N-1}\tilde{K}(\alpha_j,\alpha_{j+1},\tau). \tag{2.8}$$

The possibility of replacing $\tilde{K}(\alpha_j,\alpha_{j+1},\tau)$ by $\exp[(i/\hbar)\tau S(\alpha_j,\alpha_{j+1},\tau)]$ where this (generalized function) is defined by

$$S(\alpha_j,\alpha_{j+1},\tau) = -(i\hbar/\tau)\operatorname{Ln}\tilde{K}(\alpha_j,\alpha_{j+1},\tau) + \text{const} \tag{2.9}$$

would permit the formulation of the propagator for a given set $\{|\alpha\rangle\}$ in a manner analogous to Eq. (2.6), but not necessarily with the inherent classical significance.

Using Eq. (2.9), we have the limiting expression

$$\begin{aligned} K(\alpha_0,\alpha_N,t,t_0) &= \lim_{\substack{\tau\to0\\N\to\infty\\(N\tau=t-t_0)}} A\int d\alpha_1\int d\alpha_2\cdots\int d\alpha_{N-1} \\ &\quad\times\exp\left(\frac{i}{\hbar}\sum_{j=0}^{N-1}\tau S(\alpha_j,\alpha_{j+1},\tau)\right) \\ &= A\int\mathfrak{D}(\alpha)\exp\left(\frac{i}{\hbar}\int_{t_0}^{t}S(\alpha_0,\alpha,\alpha_N,\tau)\,d\tau\right). \end{aligned} \tag{2.10}$$

Assuming, for simplicity, that the Hamiltonian is not time dependent, we adopt a complete set of orthogonal states corresponding to some Hermitian operator (which we will represent as an arbitrary part of the system Hamiltonian). Let $\hat{H} = \hat{H}_1 + \hat{H}_2$, where

$$\hat{H}_1|\mathbf{n}\rangle = E_n|\mathbf{n}\rangle \quad\text{and}\quad \langle\mathbf{n}'|\mathbf{n}\rangle = \delta'_{\mathbf{n},\mathbf{n}} \quad\text{for all } \mathbf{n} \text{ and } \mathbf{n}'.$$

We emphasize that the choice of states $\{|\mathbf{n}\rangle\}$ (and therefore of $\hat{H}_1$) is arbitrary and need not represent a large part of the Hamiltonian $\hat{H}$. We will further assume that the parameter set $\{\mathbf{n}\}$ is discrete, since this simplification does not substantially alter our results. The propagator may be written in this representation as

$$K(\mathbf{n},\mathbf{n}',t-t_0) = \langle\mathbf{n}'|\exp[-(i/\hbar)(\hat{H}_1+\hat{H}_2)(t-t_0)]|\mathbf{n}\rangle. \tag{2.11}$$

In the limit of small $(t-t_0)$ the exponential operator may be factored into two parts, since the first correction to this factoring approximation is quadratic in $(t-t_0)$:

$$K(\mathbf{n},\mathbf{n}',\tau) \cong \langle\mathbf{n}'|\exp[-(i/\hbar)\hat{H}_1\tau]\exp[-(i/\hbar)\hat{H}_2\tau]|\mathbf{n}\rangle, \tag{2.12a}$$

$$\tilde{K}(\mathbf{n},\mathbf{n}',\tau) \equiv \exp[-(i/\hbar)E_{1,\mathbf{n}'}\tau]\langle\mathbf{n}'|\exp[(i/\hbar)\hat{H}_2\tau]|\mathbf{n}\rangle. \tag{2.12b}$$

The ordering of the operators in Eq. (2.12) is clearly arbitrary. The limiting form of K is independent of the choice of ordering, of course, but the approximate form for finite τ differs with the ordering of the operators. For instance, the approximation in Eq. (2.12) can be made exact to second order in τ, rather than first, by symmetrically splitting the $\hat{H}_1$ factor.

Define the matrix elements

$$[\underline{a}(\tau)]_{\mathbf{n}',\mathbf{n}} \equiv \langle\mathbf{n}'|\exp[-(i/\hbar)\hat{H}_2\tau]|\mathbf{n}\rangle, \tag{2.13a}$$

$$[\underline{A}(\tau)]_{\mathbf{n}',\mathbf{n}} \equiv \exp[-(i/\hbar)E_{1,\mathbf{n}'}\tau][\underline{a}(\tau)]_{\mathbf{n}',\mathbf{n}}. \tag{2.13b}$$

We may write Eq. (2.8), in this representation, as a matrix element of the Nth power of $\underline{A}$:

$$K(\mathbf{n}_0,\mathbf{n}_N,t-t_0) = \lim_{\substack{\tau\to0\\N\to\infty\\(N\tau=t-t_0)}} \sum_{\mathbf{n}_1}\sum_{\mathbf{n}_2}\cdots\sum_{\mathbf{n}_{N-1}}\prod_{j=0}^{N-1}\tilde{K}(\mathbf{n}_j,\mathbf{n}_{j+1},\tau), \tag{2.14a}$$

$$= \lim_{\substack{\tau\to0\\N\to\infty\\(N\tau=t-t_0)}} [\underline{A}^N(\tau)]_{n_N,n_0}. \tag{2.14b}$$

In general, A is a matrix of infinite order, and as such the matrix elements of the Nth power are understood as the limit of the Nth power of a finite order matrix as the order increases without bound. The τ limit is taken subsequent to this process. However, as we have

our attention focussed on possible applications to numerical computations, the precise definitions of the limiting processes are of secondary importance here.

It is quite illustrative of the inherent simplicity of Eq. (2.14) to write the product in terms of the diagonalizing transformation for $\underline{A}(t)$:

$$\underline{A}^N = \underline{S}(\underline{S}^{-1}\underline{A}\underline{S})^N\underline{S}^{-1} = \underline{S}\,\underline{\Lambda}_N\underline{S}^{-1}, \tag{2.15}$$

where

$$\underline{\Lambda}_N = \begin{bmatrix} \lambda_1^N & & \\ & \lambda_2^N & 0 \\ 0 & & \ddots \end{bmatrix}$$

for $\{\lambda_i\}$ the eigenvalues of $\underline{A}$.

The matrix $\underline{S}$ may be represented as[10]

$$(\underline{S})_{\mathbf{n},\mathbf{m}} = \langle \mathbf{n}|\,\tilde{\mathbf{m}}\rangle, \tag{2.16}$$

where $\{|\tilde{\mathbf{m}}\rangle\}$ are the eigenstates of the operator $\exp[-(i/\hbar)\hat{H}_1\tau]\exp[-(i/\hbar)\hat{H}_2\tau]$. We may use this to rewrite Eq. (2.15) as

$$(\underline{A}^N)_{\mathbf{n},\mathbf{m}} = \sum_{\mathbf{l},\mathbf{k}} \langle \mathbf{n}|\,\tilde{\mathbf{l}}\rangle\,(\underline{\Lambda}_N)_{\mathbf{l},\mathbf{k}}\langle\tilde{\mathbf{k}}|\,\mathbf{m}\rangle.$$

The limiting expression (in which $N\to\infty, N\tau = t - t_0$ are understood)

$$\lim_{\tau\to 0}(\underline{A}^N)_{\mathbf{n},\mathbf{m}} = \lim_{\tau\to 0}\sum_{\mathbf{l}}\langle \mathbf{n}|\,\tilde{\mathbf{l}}\rangle\langle\bar{\mathbf{l}}\,|\,\mathbf{n}\rangle\lambda_1^N \tag{2.17}$$

is determined by the fact that

$$\lim_{\tau\to 0}|\,\tilde{\mathbf{m}}\rangle = |\,\bar{\mathbf{m}}\rangle, \tag{2.18}$$

where the $\{|\,\bar{\mathbf{m}}\rangle\}$ are the eigenstates of $(\hat{H}_1 + \hat{H}_2)$.

The final result is that

$$\lim_{\tau\to 0}(\underline{A}^N)_{\mathbf{n},\mathbf{m}}(t) = \left\langle \mathbf{n}\middle|\left(\sum_{\mathbf{l}}|\,\bar{\mathbf{l}}\rangle\langle\bar{\mathbf{l}}\,|\exp[-(i/\hbar)\bar{E}_1 t]\right|\mathbf{m}\right\rangle, \tag{2.19}$$

where $t = N\tau$ and the $\{\bar{E}_1\}$ are the eigenvalues of $(\hat{H}_1 + \hat{H}_2)$. Therefore, Eq. (2.14b) does lead us back to the simple propagator expression in terms of the eigenstates of the full Hamiltonian (let $t_0 = 0$).

For finite values of τ the matrix $\underline{A}$ is still unitary, a property which yields a set of simple sum rules for powers of $\underline{A}$. Referring to Eq. (2.12), we define

$$\sum_{\mathbf{m}}\tilde{K}^N(\mathbf{m},\mathbf{n},N\tau)\langle\mathbf{m}|\,\psi(t_0)\rangle \equiv \langle\mathbf{n}|\,\tilde{\psi}(N\tau + t_0)\rangle,$$

such that

$$\lim_{\tau\to 0}|\,\tilde{\psi}(N\tau + t_0)\rangle = |\,\psi(t)\rangle.$$

By the unitarity of $\underline{A}$, $\langle\tilde{\psi}\,|\,\tilde{\psi}\rangle = 1$ and choosing $|\,\psi(\tau_0)\rangle = |\,1\rangle$, we obtain

$$\sum_{\mathbf{n}}|\,\tilde{K}^N(\mathbf{n},\mathbf{1},N\tau)\,|^2 = \sum_{\mathbf{n}}|\,(\underline{A}^N(\tau))_{1,\mathbf{n}}\,|^2 = 1. \tag{2.20a}$$

Referring to the definitions in Eq. (2.13), it is obvious that

$$\sum_{\mathbf{n}}|\,(\underline{a}(\tau))_{1,\mathbf{n}}\,|^2 = 1. \tag{2.20b}$$

A particularly useful form of the equations of motion of a quantum system where the Hamiltonian is partitioned (as it is here) is the interaction representation. From our point of view of taking vanishingly small time steps to generate the full propagator, we go to an interaction representation that involves only a trivial modification of the equations. In the representation defined by the Hamiltonian $\hat{H}_1$

$$|\,\psi_I(t)\rangle \equiv \exp[(i/\hbar)\hat{H}_1 t]\exp[-(i/\hbar)(\hat{H}_1 + \hat{H}_2)t]|\,\psi(t=0)\rangle. \tag{2.21}$$

Therefore,

$$\tilde{K}_I(\mathbf{n}',\mathbf{n},\tau) = \langle\mathbf{n}|\exp[(i/\hbar)\hat{H}_1\tau]\exp[-(i/\hbar)\hat{H}_1\tau] \times \exp[-(i/\hbar)\hat{H}_2\tau]|\,\mathbf{n}'\rangle.$$

Thus, the appropriate matrix replacing $\underline{A}(t)$ is simply $\underline{a}(t)$, and we have

$$K_I(\mathbf{n}',\mathbf{n},t-t_0) = \lim_{\tau\to 0}(a^N(\tau))_{\mathbf{n},\mathbf{n}'}. \tag{2.22}$$

The effect is simply to eliminate the phase factors associated with the Hamiltonian $\hat{H}_1$. The sum rule in Eq. (2.20b) is clearly independent of the representation.

3. RELATION TO PERTURBATION THEORY

The approximate solution for a complex quantum system is often obtained by some variant of perturbation theory based on the solution to a simpler problem. In this light, it is of interest to analyze the approximate solutions contained in Eq. (2.14) for finite values of τ and N. We address the problem of extracting the groups of terms from Eq. (2.14) corresponding to different orders of standard perturbation theory for the case where $\hat{H}_2$ is such that the perturbation series converges. Following this program, we find that for finite N and τ the approximate equivalence of the two sides of Eq. (2.14) leads to approximate integration formulas for the multiple time integrals of the first N orders of standard perturbation theory. In the limit where Eq. (2.14) becomes exact, the expression extends to all orders of perturbation theory as the integral approximations become exact.

We wish to group the terms in Eq. (2.14) in powers of the "perturbing" Hamiltonian $\hat{H}_2$. The expression for our comparison is the series form of the time evolution operator[11]

$$\hat{U}(t,t_0) = \hat{U}_0(t,t_0) + \sum_{n=1}^{\infty}\hat{U}^{(n)}(t,t_0), \tag{3.1a}$$

where

$$\hat{U}_0(t,t_0) = \exp[-(i/\hbar)(t-t_0)\hat{H}_1]$$

$$\hat{U}^{(n)}(t,t_0) = (i\hbar)^{-n}\int\cdots\int d\tau_n d\tau_{n-1}\cdots d\tau_1\hat{U}_0(t,t_n) \times \hat{H}_2\hat{U}_0(\tau_n,\tau_{n-1})\cdots\hat{H}_2\hat{U}_0(\tau_2,t_0), \quad t>\tau_n>\tau_{n-1}>\cdots>\tau_1>t_0. \tag{3.1b}$$

The integration is over that part of n-dimensional space satisfying the indicated time-ordering constraint. To obtain all the mth order terms in $\hat{H}_2$ from Eq. (2.14), we need only consider the first m terms in the $\hat{H}_2$ expansion of the exponential form $\underline{a}(t)$. We have assumed here that $\hat{H}_2$ is free of explicit time dependence. Then, for a jth order expression, we use the terms

$$(\underline{a}(\tau))_{\mathbf{n},\mathbf{m}} \cong \sum_{\alpha=0}^{j}\left(-\frac{i}{\hbar}\right)\frac{\tau^\alpha}{\alpha!}(\underline{H}_2^\alpha)_{\mathbf{n},\mathbf{m}}, \tag{3.2}$$

with

$$(\underline{H}_2^\alpha)_{\mathbf{n},\mathbf{m}} = \sum_{\mathbf{k}_1,\mathbf{k}_2,\ldots,\mathbf{k}_{\alpha-1}}\langle\mathbf{n}|\hat{H}_2|\mathbf{k}_1\rangle \times\langle\mathbf{k}_1|\hat{H}_2|\mathbf{k}_2\rangle\cdots\langle\mathbf{k}_{\alpha-1}|\hat{H}_2|\mathbf{m}\rangle.$$

The jth order terms of A^N are then obtained by extracting them from the Nth power of A obtained from substituting the expression from Eq. (3.2) for $(a(t))_{\mathbf{n}',\mathbf{n}}$ in Eq. (2.13b).

To illustrate this procedure, we will obtain the first few orders while keeping in mind that this specific ordering of the components of the evolution operator need not be retained. The first order terms arise from the expression

$$\{\exp[-(i/\hbar)E_{\mathbf{n}'}\tau][\delta_{\mathbf{n}',\mathbf{n}}-(i/\hbar)\tau H_{\mathbf{n}',\mathbf{n}}]\}^N$$

(where the operation of raising to the Nth power is understood to apply to the matrix having the $(\mathbf{n}',\mathbf{n})$ elements indicated in the brackets). The first order terms, indicated by $\langle\mathbf{n}'|\tilde{u}^{(1)}(t_1t_0)|\mathbf{n}\rangle$, are given by

$$\begin{aligned}\langle\mathbf{n}'|\tilde{u}^{(1)}(t_1t_0)|\mathbf{n}\rangle &= -i/\hbar\,\tau H_{\mathbf{n}',\mathbf{n}}\\ &\times\{\exp[-(i/\hbar)E_{\mathbf{n}}N\tau]+\exp[-(i/\hbar)E_{\mathbf{n}'}(N-1)\tau]\\ &\times\exp[-(i/\hbar)E_{\mathbf{n}}\tau]+\cdots+\exp[-(i/\hbar)E_{\mathbf{n}}\tau]\\ &\times\exp[-(i/\hbar)E_{\mathbf{n}}(N-1)\tau]\},\end{aligned}\tag{3.3}$$

where t_0 has been set equal to zero for simplicity.

In the limit as $\tau\to 0(N\to\infty)$, the above expression is indeed equivalent to Eq. (3.1b) for $n=1$:

$$\begin{aligned}\lim_{\substack{\tau\to 0\\ N\tau=t}}\langle\mathbf{n}'|\tilde{u}^{(1)}(t)|\mathbf{n}\rangle &= -(i/\hbar)H_{\mathbf{n}',\mathbf{n}}\int_0^t d\tau_1\\ &\times\exp[-(i/\hbar)E_{\mathbf{n}'}(t-\tau_1)]\exp[-(i/\hbar)E_{\mathbf{n}}\tau_1].\end{aligned}\tag{3.4}$$

Equation (3.3) clearly represents a trapezoidal rule integral approximation to the limiting expression. Furthermore, the abscissas defining the formula are determined by the ordering in Eq. (2.12). For instance, splitting the $\hat{H}_1$ term yields abscissas centered in each time increment τ.

The expression in Eq. (3.3) may also be directly summed. Thus

$$\begin{aligned}\langle\mathbf{n}'|\tilde{u}^{(1)}(t,0)|\mathbf{n}\rangle &= -\frac{i}{\hbar}H_{\mathbf{n}',\mathbf{n}}\tau\sum_{j=0}^{N-1}\\ &\times\exp\left(-\frac{i}{\hbar}\tau[E_{\mathbf{n}'}(N-j)+E_{\mathbf{n}}j]\right)\\ &=-\frac{i}{\hbar}H_{\mathbf{n}',\mathbf{n}}\tau\exp\left(-\frac{i}{\hbar}E_{\mathbf{n}}t\right)\sum_{j=0}^{N-1}\\ &\times\left[\exp\left(-\frac{i}{\hbar}\tau(E_{\mathbf{n}}-E_{\mathbf{n}'})\right)\right]^j\\ &=-\frac{i}{\hbar}\tau H_{\mathbf{n}',\mathbf{n}}\exp\left(-\frac{i}{\hbar}E_{\mathbf{n}'}t\right)\left\{\frac{1-\exp(-i\omega_{\mathbf{n},\mathbf{n}'}t)}{1-\exp(-i\omega_{\mathbf{n},\mathbf{n}'}\tau)}\right\},\end{aligned}\tag{3.5}$$

where $\omega_{\mathbf{n},\mathbf{n}'}=(E_{\mathbf{n}}-E_{\mathbf{n}'})/\hbar$. The equivalence of this expression with the integral of Eq. (3.4) in the limit $\tau\to 0$ is thus established quite easily:

$$\begin{aligned}\lim_{\substack{\tau\to 0\\ N\tau=t}}\langle\mathbf{n}'|\tilde{u}^{(1)}(t)|\mathbf{n}\rangle &= -\frac{i}{\hbar}H_{\mathbf{n}',\mathbf{n}}\exp\left(-\frac{i}{\hbar}E_{\mathbf{n}'}t\right)\\ &\times\left\{\frac{1-\exp(-i\omega_{\mathbf{n},\mathbf{n}'}t)}{i\omega_{\mathbf{n},\mathbf{n}'}}\right\}.\end{aligned}\tag{3.6}$$

The j sums in the general jth order expression may be similarly calculated and shown to reduce to the appropriate integral expressions. In Appendix B we calculate the second order expressions of Eq. (3.8). The second order terms are obtained from

$$\left[\exp\left(-\frac{i}{\hbar}E_{\mathbf{n}'}\tau\right)\left(\delta_{\mathbf{n}',\mathbf{n}}-\frac{i}{\hbar}\tau H_{\mathbf{n}',\mathbf{n}}+\frac{1}{2}\left(\frac{i}{\hbar}\right)^2\tau^2\sum_1 H_{\mathbf{n}',1}H_{1,\mathbf{n}}\right)\right]^N.\tag{3.7}$$

In this case there are two classes of terms contributing to the entire second order—those stemming from the second and third terms of Eq. (3.7), respectively. Thus,

$$\begin{aligned}\langle\mathbf{n}|\tilde{u}^{(2)}(N\tau)|\mathbf{n}\rangle &= \left(-\frac{i}{\hbar}\right)^2\tau^2\sum_1 H_{\mathbf{n}',1}H_{1,\mathbf{n}}\\ &\times\sum_{k=0}^{N-2}\sum_{m=0}\exp\left(-\frac{i}{\hbar}E_{\mathbf{n}'}(N-1-k)\tau\right)\\ &\times\exp\left(-\frac{i}{\hbar}E_1(k-m+1)\tau\right)\exp\left(-\frac{i}{\hbar}E_{\mathbf{n}}m\tau\right)\\ &+\left(-\frac{i}{\hbar}\right)^2\frac{\tau^2}{2}\sum_1 H_{\mathbf{n}',1}H_{1,\mathbf{n}}\\ &\times\sum_{k'=0}^{N-1}\exp\left(-\frac{i}{\hbar}E_{\mathbf{n}'}(N-k')\tau\right)\exp\left(-\frac{i}{\hbar}E_{\mathbf{n}}k'\tau\right).\end{aligned}\tag{3.8}$$

The weighting of these sets of terms by τ^2 and $\tau^2/2$ reflects their geometric significance as components of a two-dimensional integral approximation. In Fig. 1 the appropriate square grid for $N=4$ is shown with the abscissas for the second set of terms indicated by open circles, and their integration regions are shaded.

The situation is much the same for the third order terms, with a single exception. The extremal terms, those satisfying the time ordering constraint in the limit of simultaneity ($\tau_1=\tau_2=\tau_3\cdots\tau_{n-1}=\tau_n$), are weighted by a factor that is less than that expected from direct application of a trapezoidal rule: the geometric value of the integration region is $\tau^3/4$ instead of $\tau^3/6$. Inspection of Eq. (3.2) indicates that all the extremal terms for third order and above will be weighted in a similar fashion. The ratio of the geometric to the actual weighting factors is seen to be $n!/2^{n-1}$ for the nth order extremal terms. As N becomes large, however, the approximate expressions approach the exact limit as a trapezoidal rule since the number of extremal terms is less

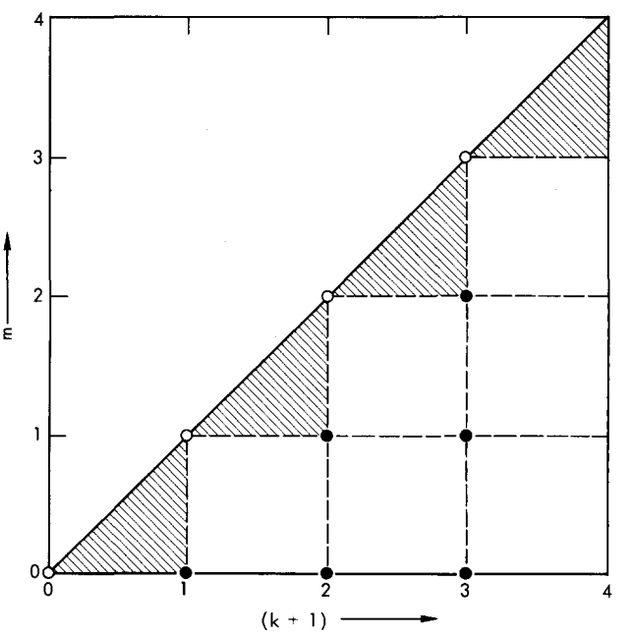


FIG. 1. $N=4$ grid for integration domain of second order terms. The dots and circles indicate the abscissas of the first and second sets of terms of Eq. (3.8), respectively.

than the total number of terms by a factor that is proportional to $1/N$.

Thus for finite τ and N the procedure defined by Eq. (2.14) is related to the perturbation series in an interesting way. It sums approximately all orders of perturbation theory where we may characterize the approximation by a trapezoidal-like integration rule.

For a given value of N, those orders of perturbation theory greater than N are represented by a single term each. These are the terms remaining after we have identified the above mentioned integral approximations for the first N orders of perturbation theory and are therefore of little significance to the approximation for reasonably large values of N.

4. EXCHANGE SYMMETRIES

The system states considered so far are completely general and therefore implicitly include proper permutation symmetries for many-particle systems. However, when faced with the problem of reducing the above considerations to a form amenable to actual computation, the retention of fully permuted states can be quite ponderous. The full consideration of exchange requires that we use a total propagator that is effectively the sum of propagators in single particle quantum numbers (or coordinates); each of the terms in the sum being characteristic of one possible permutation of the single particles. The advantage of the present approach in dealing with this complexity is that for small τ we need only consider *two*-particle exchange, because the ensuing sequence of two-particle exchanges will generate any possible exchange.[6] This contention is quite intuitive but demands a more detailed analysis that is partially provided in Appendix A.

If $K(\mathbf{n}, \mathbf{n}', t)$ is a many-particle propagator, written without exchange symmetry, we may express the full propagator (using the results of Appendix A) by the relation

$$K_E(\mathbf{n}_0, \mathbf{n}_N, N\tau) = \lim_{\tau\to 0} \sum_{\mathbf{n}_1} \cdots \sum_{\mathbf{n}_{N-1}} \prod_{j=1}^{N} A\left(1 \pm \sum^{M} \hat{P}\right) \times \tilde{K}(\mathbf{n}_j, \mathbf{n}_{j+1}, \tau), \quad (4.1)$$

where A is a normalization constant, $\hat{P}$ is the two-particle permutation operator acting over all M exchanges [$M = L(L-1)/2$ if all L particles are identical], and the $\pm$ sign indicates Bose or Fermi statistics.

In terms of the matrices in Eq. (2.13), Eq. (4.1) becomes, for L particles,

$$\begin{aligned} K_E(\mathbf{n}, \mathbf{n}', \tau) &= A\left(1 \pm \sum^{M} \hat{P}\right) \exp\left(-\frac{i}{\hbar} \sum_{j=1}^{L} E_{\mathbf{n}_j'} \tau\right) \\ &\quad \times \left\langle \{\mathbf{n}_j'\} \middle| \exp\left(-\frac{i}{\hbar} \hat{H}_2 \tau\right) \middle| \{\mathbf{n}_j\} \right\rangle \\ &= A \exp\left(-\frac{i}{\hbar} \sum_{j=1}^{L} E_{\mathbf{n}_j'}\right) \left\langle \{\mathbf{n}_j'\} \middle| \exp\left(-\frac{i}{\hbar} \hat{H}_2 \tau\right) \right. \\ &\quad \times \left. \left(1 \pm \sum^{M} \hat{P}\right) \middle| \{\mathbf{n}_j\} \right\rangle. \end{aligned} \quad (4.2)$$

Define the new exchanged matrix

$$[\underline{a}_E(\tau)]_{\{\mathbf{n}_j'\},\{\mathbf{n}_j\}} \equiv \left\langle \{\mathbf{n}_j'\} \middle| \exp\left(-\frac{i}{\hbar} \hat{H}_2 \tau\right) A\left(1 \pm \sum^{M} \hat{P}\right) \middle| \{\mathbf{n}_j\} \right\rangle, \quad (4.3)$$

and we may cast the exchanged propagator into a form identical to Eq. (2.14).

Thus,

$$K_E(\{\mathbf{n}_j\}, \{\mathbf{n}_j'\}, t) = \lim_{\tau\to 0} (\underline{A}_E^N(\tau))_{\{\mathbf{n}_j'\},\{\mathbf{n}_j\}}, \quad (4.4a)$$

where

$$(\underline{A}_E(\tau))_{\{\mathbf{n}_j'\},\{\mathbf{n}_j\}} \equiv \exp\left(-\frac{i}{\hbar} \sum_{j=1}^{L} E_{\mathbf{n}_j'} \tau\right) (\underline{a}_E(\tau))_{\{\mathbf{n}_j'\},\{\mathbf{n}_j\}}. \quad (4.4b)$$

The crux of the calculation of a fully exchanged propagator is therefore the determination of $\underline{a}_E(t)$, a small-time, two-particle exchanged matrix.

5. IMAGINARY TIME

From the form of the equation

$$\hat{H}|\psi(t)\rangle = i\hbar \frac{\partial}{\partial t}|\psi(t)\rangle \quad (5.1)$$

in the Schrödinger picture, we can replace the quantity (it) by the imaginary time coordinate ϵ to obtain a diffusion type equation.

By using the eigenstates of $\hat{H}$, $\{|\tilde{\mathbf{l}}\rangle\}$, the evolution operator for Eq. (5.1) is

$$\hat{U}(t, t_0) = \sum_{\mathbf{l}} \exp\left(-\frac{i}{\hbar} E_{\mathbf{l}}(t - t_0)\right) |\tilde{\mathbf{l}}\rangle\langle\tilde{\mathbf{l}}|. \quad (5.2)$$

Then the propagator, in the representation of eigenstates of $\hat{H}_1$, along the imaginary time coordinate becomes

$$K_I(\mathbf{n}, \mathbf{n}', \epsilon) = \sum_{l} \exp[-(\epsilon/\hbar) E_{\mathbf{l}}] \langle \mathbf{n}'|\tilde{\mathbf{l}}\rangle\langle\tilde{\mathbf{l}}|\mathbf{n}\rangle. \quad (5.3)$$

As $(t - t_0)$ becomes sufficiently large, we recognize that the ground state of the system emerges as a diagonal element of K_I:

$$\lim_{\epsilon\to\infty} K_I(\mathbf{n}, \mathbf{n}, \epsilon) = \exp[-(\epsilon/\hbar)E_0] \sum_{l_0} |\langle \mathbf{n}|\tilde{\mathbf{l}}_0\rangle|^2. \quad (5.4)$$

Therefore, the ground state energy of the system is given by

$$E_0 = -(\hbar/\epsilon) \lim_{t\to\infty} \ln[K_I(\mathbf{n}, \mathbf{n}, \epsilon + t)/K_I(\mathbf{n}, \mathbf{n}, t)]. \quad (5.5)$$

It has been pointed out before[5,6] that, by restricting the symmetry of the propagator, the results of Eqs. (5.4) and (5.5) may be obtained within each such restricted symmetry class.

APPENDIX A

The proposed method could be used to great advantage if we need consider only *two*-particle exchange in an N-particle propagator. To do this, we require that two-particle exchange generates, after a large number of time steps, not only all possible permutations, but generates them in *equal numbers* so that the final propagator is correctly symmetrized. If $\hat{P}_{ij}$ indicates the two-particle exchange operator, we may state the theorem to be proved in the form

$$\lim_{L\to\infty} \left(1 \pm \sum_{i<j}^{N} \hat{P}_{ij}\right)^L \sim \sum_{n} \delta_n^{(\pm)} P_n, \quad (A1)$$

where $\{P_n\}$ is the full symmetric group of order N and $\delta_n^{(\pm)}$ carries the appropriate algebraic sign. For $N = 2$, Eq. (A1) is true since

$$(1 \pm \hat{P}_{12})^2 = 2(1 \pm \hat{P}_{12}).$$

For $N = 3$, the problem is already much more compli-

cated. We will carry out a direct proof for $N = 3$ (since this is an important case) and then merely indicate a proof for the general case.

There are six possible permutations for $N = 3$ and three permutation classes:

$$I:\begin{pmatrix}1&2&3\\1&2&3\end{pmatrix},\qquad P_{12}:\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix},\qquad P_{13}:\begin{pmatrix}1&2&3\\3&2&1\end{pmatrix},$$

$$P_{23}:\begin{pmatrix}1&2&3\\1&3&2\end{pmatrix},\qquad C:\begin{pmatrix}1&2&3\\3&1&2\end{pmatrix},\qquad C^2:\begin{pmatrix}1&2&3\\2&3&1\end{pmatrix}.$$

Using the decomposition of the operators C and C^2,

$$\hat{C} = \hat{P}_{12}\hat{P}_{23} = \hat{P}_{13}\hat{P}_{12} = \hat{P}_{23}\hat{P}_{13},$$
$$\hat{C}^2 = \hat{P}_{23}\hat{P}_{12} = \hat{P}_{12}\hat{P}_{13} = \hat{P}_{13}\hat{P}_{23}, \tag{A2}$$

we take the powers of $[1 + (\hat{P}_{12} + \hat{P}_{13} + \hat{P}_{23})]$ and designate the coefficients of the three permutation classes as follows: A for the identity, B for $\{\hat{P}_{ij}\}$, and C for $\{\hat{C}, \hat{C}^2\}$. At this point we may simplify subsequent manipulations by absorbing the $\pm$ sign into the definition of the exchange operator $\hat{P}_{ij}$. Since the parity of a permutation is determined by the number of exchange factors into which it decomposes, this measure eliminates the need for the $\delta_n^{(\pm)}$ in Eq. (A1). The recursion relations for these coefficients (where n indicates the power involved) are

$$A_n = A_{n-1} + 3B_{n-1}, \tag{A3a}$$

$$B_n = B_{n-1} + A_{n-1} + 2C_{n-1}, \tag{A3b}$$

$$C_n = C_{n-1} + 3B_{n-1}, \tag{A3c}$$

with $A_1 = B_1 = 1$, $C_1 = 0$ as initial conditions. The equation may be simplified by eliminating C_{n-1}, using the first and third equations to obtain $C_n = A_n - 1$. Then we have

$$B_n = B_{n-1} + 3A_{n-1} - 2. \tag{A4}$$

In matrix form {with $\mathbf{A}_n = (A_n, B_n)$, $\underline{M} = \left[\begin{smallmatrix}1&3\\3&1\end{smallmatrix}\right]$ and $\gamma = (0, 2)$}, Eqs. (A3a) and (A4) become

$$\mathbf{A}_n = \underline{M}\cdot\mathbf{A}_{n-1} - \gamma. \tag{A5}$$

Iterating this relation gives

$$\mathbf{A}_n = \underline{M}^{n-1}\cdot\mathbf{A}_1 - \sum_{i=2}^{n}\underline{M}^{n-i}\cdot\gamma. \tag{A6}$$

Our theorem is proved if $\lim_{n\to\infty}\mathbf{A}_n \sim (1, 1)$, which is the case [by Eq. (A6)] if $\lim_{n\to\infty}\underline{M}^n \sim \left[\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right]$. That this is true can be shown as follows:

$$\begin{bmatrix}1&3\\3&1\end{bmatrix}^n = \left\{\begin{bmatrix}1&0\\0&1\end{bmatrix} + 3\begin{bmatrix}0&1\\1&0\end{bmatrix}\right\}^n = \sum_{i=0}^{n}\binom{n}{i}3^i\begin{bmatrix}0&1\\1&0\end{bmatrix}^i,$$

and since

$$\begin{bmatrix}0&1\\1&0\end{bmatrix}^i = \begin{cases}\begin{bmatrix}1&0\\0&1\end{bmatrix}, & i \text{ even}\\ \begin{bmatrix}0&1\\1&0\end{bmatrix}, & i \text{ odd}\end{cases},$$

$$\begin{bmatrix}1&3\\3&1\end{bmatrix}^n = \begin{bmatrix}\alpha_n&\beta_n\\\beta_n&\alpha_n\end{bmatrix},$$

where

$$\alpha_n = \sum_{2i=0}^{[n]\text{even}}\binom{n}{2i}3^{2i},\qquad \beta_n = \sum_{2i+1=1}^{[n]\text{odd}}\binom{n}{2i+1}3^{2i+1}. \tag{A7}$$

Now since we have the relations

$$\alpha_n + \beta_n = (1 + 3)^n = 4^n \tag{A8a}$$

and

$$\alpha_n - \beta_n = (1 - 3)^n = -2^n, \tag{A8b}$$

we combine them and take the limit:

$$\lim_{n\to\infty}[(\alpha_n - \beta_n)/(\alpha_n + \beta_n)] = \lim_{n\to\infty}(-1/2)^n = 0. \tag{A9}$$

Thus, the theorem is seen to be true for $N = 3$.

For the general case, we find it useful to label the permutation operations $\{P_n\}$ and the coefficients for these $\{a_n\}$. Now consider the operator

$$\left(1 \pm \sum_{i<j}^{N}\hat{P}_{ij}\right)^L \sim \sum_n a_n\delta_n^{(\pm)}P_n, \tag{A10}$$

and further restrict the $\{a_n\}$ so that $\sum_n a_n = \text{constant}$. This is simply a choice of normalization that somewhat simplifies the following. Consider the product

$$\left(1 \pm \sum_{i<j}^{N}\hat{P}_{ij}\right)\left(\sum_n a_n\hat{P}_n\delta_n^{(\pm)}\right) = \sum_n a_n'\delta_n^{(\pm)}P_n \tag{A11}$$

which results at each time step. Since the coefficients are assumed positive [the signs of the terms in Eq. (A10) determined by the parity of the permutations], we may write the change in $\{a_n\}$ as follows:

$$\Delta a_n \equiv a_n' - a_n = \sum_{j}^{a_n} a_j - \tfrac{1}{2}N(N-1)a_n, \tag{A12}$$

where the sum is over the set of operators connected to P_n by single pair exchanges. We may also decompose the above expression into the parts due to each pair exchange. If $\hat{P}_{ij}$ takes P_n into P_k, then $\hat{P}_{ij}$ also takes P_k into P_n since $(\hat{P}_{ij})^2 = 1$. Therefore, we have

$$\Delta a_n = \sum_{k}^{a_n}\Delta_k a_n,$$

where

$$\Delta_k a_n = a_k - a_n.$$

It is clear at this point that one steady state solution is in fact the desired one: where $a_k = a_l$ for all k and l ($\leq N!$). It is also clear that any difference between two coefficients results in a change that tends to eliminate the difference.

This is far from being a complete proof, but it is a strong argument for the existence of such a proof.

Note added in proof:

There exists a theorem, similar to (A1), which decomposes the sum $\sum_n \delta_n^{(\pm)}\hat{P}_n$ into a finite number of two particle exchange factors. The factorization

$$\sum_n \delta_n^{(\pm)}\hat{P}_n = \prod_{j=1}^{N-1}\left(1 \pm \sum_{i=j+1}^{N}\hat{P}_{ij}\right)$$

differs substantially from (A1), however, in that the factors are not identical and therefore may not be encorporated into the expression for the propagator as the Nth power of a matrix operator.

APPENDIX B

The second order finite time step expression for the evolution matrix, given by Eq. (3.8), may be reduced to a much simpler expression by carrying out the indicated summations. Thus,

$$\langle \mathbf{n}' | \tilde{u}^{(2)}(t) | \mathbf{n} \rangle = \left(-\frac{i}{\hbar}\right)^2 \tau^2 \sum_{\mathbf{l}} H_{\mathbf{n}',\mathbf{l}} H_{\mathbf{l},\mathbf{n}} \times \sum_{k=0}^{N-2} \sum_{m=0}^{k} \exp\left(-\frac{i}{\hbar} E_{\mathbf{n}'}(N-1-k)\tau - \frac{i}{\hbar} E_{\mathbf{l}}(k-m+1)\tau - \frac{i}{\hbar}\tau E_{\mathbf{n}} m\right) + \left(-\frac{i}{\hbar}\right)^2 \frac{\tau^2}{2} \sum_{\mathbf{l}} H_{\mathbf{n}',\mathbf{l}} H_{\mathbf{l},\mathbf{n}} \sum_{m=0}^{N-1} \times \exp\left(-\frac{i}{\hbar} E_{\mathbf{n}'}(N-m)\tau - \frac{i}{\hbar} E_{\mathbf{n}} m\tau\right). \quad \text{(B1)}$$

Performing the sum over m, we obtain

$$\langle \mathbf{n}' | \tilde{u}^{(2)}(t) | \mathbf{n} \rangle = \left(-\frac{i}{\hbar}\right)^2 \sum_{\mathbf{l}} H_{\mathbf{n}',\mathbf{l}} H_{\mathbf{l},\mathbf{n}} \exp\left(-\frac{i}{\hbar} E_{\mathbf{n}'} t\right) \times \left[\sum_{k=0}^{N-2} \tau^2 \exp[-i\omega_{\mathbf{l},\mathbf{n}'}\tau(k+1)] \times \left(\frac{1-\exp[-i\omega_{\mathbf{n},\mathbf{l}}\tau(k+1)]}{1-\exp(-i\omega_{\mathbf{n},\mathbf{l}}\tau)}\right) + \frac{\tau^2}{2}\left(\frac{1-\exp(-i\omega_{\mathbf{n},\mathbf{n}'}N\tau)}{1-\exp(-i\omega_{\mathbf{n},\mathbf{n}'}\tau)}\right)\right]. \quad \text{(B2)}$$

Summing again,

$$\sum_{k=0}^{N-2} \left(\frac{\exp[-i\omega_{\mathbf{l},\mathbf{n}'}\tau(k+1)] - \exp[-i(\omega_{\mathbf{l},\mathbf{n}'} + \omega_{\mathbf{n},\mathbf{l}})(k+1)\tau]}{1-\exp(-i\omega_{\mathbf{n},\mathbf{l}}\tau)}\right) = [1-\exp(-i\omega_{\mathbf{n},\mathbf{l}}\tau)]^{-1} \left[\left(\frac{1-\exp[-i\omega_{\mathbf{l},\mathbf{n}'}(N-1)\tau]}{1-\exp(-i\omega_{\mathbf{l},\mathbf{n}'}\tau)}\right) \times \exp(-i\omega_{\mathbf{l},\mathbf{n}'}\tau) - \left(\frac{1-\exp[-i\omega_{\mathbf{n},\mathbf{n}'}(N-1)\tau]}{1-\exp(-i\omega_{\mathbf{n},\mathbf{n}'}\tau)}\right) \exp(-i\omega_{\mathbf{n},\mathbf{n}'}\tau)\right]. \quad \text{(B3)}$$

The final closed expression becomes

$$\langle \mathbf{n}' | \tilde{u}^{(2)}(t) | \mathbf{n} \rangle = \left(-\frac{i}{\hbar}\right)^2 \sum_{\mathbf{l}} H_{\mathbf{n}',\mathbf{l}} H_{\mathbf{l},\mathbf{n}} \left[\frac{\tau^2 \exp[-(i/\hbar)E_{\mathbf{n}'}t]}{1-\exp(-i\omega_{\mathbf{n},\mathbf{l}}\tau)} \times \left(\frac{\exp(-i\omega_{\mathbf{l},\mathbf{n}'}\tau) - \exp(-i\omega_{\mathbf{l},\mathbf{n}'}t)}{1-\exp(-i\omega_{\mathbf{l},\mathbf{n}'}\tau)} - \frac{\exp(-i\omega_{\mathbf{n},\mathbf{n}'}\tau) - \exp(-i\omega_{\mathbf{n},\mathbf{n}'}t)}{1-\exp(-i\omega_{\mathbf{n},\mathbf{n}'}\tau)}\right) + \frac{\tau^2}{2}\exp\left(-\frac{i}{\hbar}E_{\mathbf{n}'}t\right)\left(\frac{1-\exp(-i\omega_{\mathbf{n},\mathbf{n}'}t)}{1-\exp(-i\omega_{\mathbf{n},\mathbf{n}'}\tau)}\right)\right]. \quad \text{(B4)}$$

If we now take the $\tau \to 0$ limit, the final term vanishes and we have

$$\lim_{\substack{\tau\to 0 \\ N\tau = t}} \langle \mathbf{n}' | \tilde{u}^{(2)}(t) | \mathbf{n} \rangle = + \left(\frac{i}{\hbar}\right)^2 \sum_{l} \frac{H_{\mathbf{n}',\mathbf{l}} H_{\mathbf{l},\mathbf{n}}}{\omega_{\mathbf{n},\mathbf{l}}} \times \left(\frac{\exp[-(i/\hbar)E_{\mathbf{n}'}t][1-\exp(-i\omega_{\mathbf{l},\mathbf{n}'}t)]}{\omega_{\mathbf{l},\mathbf{n}'}} - \frac{\exp[-(i/\hbar)E_{\mathbf{n}'}t][1-\exp(-i\omega_{\mathbf{n},\mathbf{n}'}t)]}{\omega_{\mathbf{n},\mathbf{n}'}}\right), \quad \text{(B5)}$$

which is the value of the second order integral as expected. Higher orders are quite clearly obtained by the same procedures.

---

*This work was performed under the auspices of the U. S. Atomic Energy Commission.

[1] R. P. Feynman, Rev. Mod. Phys. **20**, 367 (1948).
[2] S. G. Brush, Rev. Mod. Phys. **33**, 79 (1961).
[3] I. M. Gel'fand and A. M. Yaglom, J. Math. Phys. **1**, 48 (1960).
[4] S. V. Lawande, C. A. Jensen, and H. L. Sahlin, J. Comput. Phys. **3**, 416 (1969).
[5] S. V. Lawande, C. A. Jensen, and H. L. Sahlin, J. Comput. Phys. **4**, 451 (1969).
[6] S. V. Lawande, C. A. Jensen, and H. L. Sahlin, J. Chem. Phys. **54**, 445 (1971).
[7] J. R. Klauder, Ann. Phys. (N.Y.) **11**, 123 (1960).
[8] S. S. Schweber, J. Math. Phys. **3**, 831 (1962).
[9] J. R. Klauder, J. Math. Phys. **4**, 1055 (1963), J. Math. Phys. **4**, 1058 (1963); J. Math. Phys. **5**, 177, 878 (1964).
[10] This form is due to the particular ordering choice of Eq. (2.12a) and, as such, is not unique. The $|\tilde{\mathbf{m}}\rangle$ should be the eigenstates of whatever operator the ordering choice dictates.
[11] A. Messiah, *Quantum Mechanics* (Wiley, New York, 1961), Vol. 1.,

Green's function for Markovian systems*

J. J. Thomson and Gregory Benford

Department of Physics, University of California, Irvine, California 92664
(Received 11 July 1972)

Many perturbation series solutions to stochastic differential equations suffer from secularity; that is, individual terms in the series behave as some power of an independent variable, and thus diverge as that variable goes to infinity. We derive a method of renormalizing the "free" or "unperturbed" Green's function by summing a certain class of terms and including them in the Green's function. The terms included in this class correspond to Markovian interaction with the stochastic field. The remainder of the perturbation series corresponds to non-Markovian corrections. We give a diagrammatic interpretation of the individual terms in the perturbation series. We may solve, under certain assumptions, the explicit form of the Markovian Green's function. We apply this method to plasma turbulence, and show that it produces Dupree's theory of turbulence. We derive the condition for validity of the perturbation expansion, which is that particles are not "trapped" by the turbulent waves.

1. INTRODUCTION

The study of plasma turbulence is a subset of the general study of stochastic differential equations. Several plasma turbulence theories depend on methods of defining suitable perturbation expansions. Recently, we proposed a simple statistical model,[1] which reproduced the results of Dupree[2] and Weinstock.[3] We wish to show here the relationship between the somewhat heuristic methods we used, and classical methods of solving stochastic differential equations. In particular, we will derive a Green's function for the Vlasov equation which leads to a nonsecular perturbation series. Assuming a Markovian process, we will derive the explicit form of the Green's function.

We consider first a stochastic differential equation of the form

$$Lf = 0 \tag{1.1}$$

where L is a stochastic operator, and f, the random field quantity, is an element of an infinite dimension vector space H. The averaging operator A is a projection operator onto the space AH. L is composed of two parts: the averaged or "unperturbed" component $L_0 \equiv AL$, and the stochastic or "interaction" component $L' \equiv (I - A)L$, where I is the identity operator.

The equations for the average part of f, $\bar{f} \equiv Af$, and the stochastic part, $f' \equiv (I - A)f$, are

$$L_0\bar{f} + AL'f' = 0, \tag{1.2}$$

$$L_0 f' = (A - I)L'f' - L'\bar{f} \tag{1.3}$$

Neglecting initial value terms, the formal solution to Eq. (1.3) is

$$f' = \sum_{n=1}^{\infty} [L_0^{-1}(A - I)L']^n \bar{f}. \tag{1.4}$$

Therefore, Eq. (1.2) becomes

$$\{L_0 + AL' \sum_{n=1}^{\infty} [L_0^{-1}(A - L)L']^n\}\bar{f} = 0. \tag{1.5}$$

Eq. (1.5) is called a master equation for $\bar{f}$. Note that it is in the form

$$(L_0 - M)\bar{f} = 0. \tag{1.6}$$

M is called the mass operator. The Green's function for Eq. (1.6) is

$$G = L_0^{-1} + L_0^{-1}MG. \tag{1.7}$$

Eq. (1.7) is the Dyson equation.[4]

The difficulty with classical perturbation equations of the form of Eq. (1.5) is that the individual terms are secular—that is, if $L_0 = \partial/\partial t + (v)\partial/\partial R$, for example, they behave as some power of t and thus diverge for large t. The entire series may converge, but any finite sum of terms may diverge.

To avoid the secularity problem, one attempts partial summation of the perturbation series. Two methods of doing this are the method of smoothing,[4] and of renormalization.[5] A recent paper by Besieris[6] found stochastic master equations of the form of Eq. (1.6). The problem with his formulation, shared with many other formulations, is that the resultant Green's function is expressed as a series propagated by the "free" Green's function, L_0^{-1}. In this case, as we attempt to evaluate terms in the series, secular terms again appear.

To see this, we sketch out an argument given in more detail by Frisch.[4] The basic assumptions are that the medium is homogeneous and the process is stationary. In this case the poles of the Fourier transform of L_0^{-1} are real. When the ensemble average is taken and the transform inverted, we have contributions from terms like $(S - \omega_i)^{-n}, n > 2$, where S is the integration variable and ω_i is the ith pole of L_0^{-1}. Evaluating the contribution of this pole by residues, we obtain factors like $(1/n!)t^n \exp(i\omega_i t)$, giving rise to secularity.

To renormalize the Green's function, we introduce a damping term, i.e., an imaginary contribution to the pole of the Green's function operator G. This damping term will limit the magnitude of the terms in the perturbation series as t increases.

The object of this paper is to give a prescription for defining a "perturbed" Green's function which has a well-defined physical meaning; it applies to the case in which the random function f is undergoing a Markov process. We shall show in diagrammatic terms exactly what terms are lumped into the Green's function, and what are left in the remaining perturbation series. We can evaluate the Green's function self-consistently under certain approximations.

2. THE PERTURBED GREEN'S FUNCTION

To both sides of Eq. (1.3) we add a perturbation operator, pf':

$$(L_0 + p)f' = pf' + (A - 1)L'f' - L'\bar{f}. \tag{2.1}$$

The idea is to use pf' to subtract a certain subset of terms in the perturbation series. The Green's function we use to generate the series is defined in terms of p:

$$G = (L_0 + p)^{-1}. \tag{2.2}$$

We could immediately terminate the series by defining $pf' = -(A-1)L'f'$. Unfortunately, Eq. (2.2) becomes more difficult to solve for than the full operator L'.

Instead, we shall iterate the right hand side of Eq. (2.1) using G:

$$(L_0 + p)f' = pf' + (A-1)L'G[pf' + \underline{(A-1)L'f'} - L'\bar{f}] - L'\bar{f}. \tag{2.3}$$

We shall use pf' to cancel all or part of the underlined term in Eq. (2.3). At this point, we can take two paths. If we cancel all of the underlined term, G becomes a stochastic operator. We then have a simple perturbation series, but we will not be able to evaluate G. In the second case, we use p to cancel only the nonstochastic part of the operator $L'G(A-1)L'$. Then G is a nonstochastic operator which we may find exactly.

We anticipate that a nonstochastic G will in general be easier to handle in calculations, since the averaging, for example, does not need to include correlations between L' and G. More importantly, the nonstochastic Green's function can be considered the probability distribution function for trajectories in phase space, and thus may be used to propagate both f' and $\bar{f}$ forward in time.

A. Stochastic Green's function

We consider the first case, and denote this by the subscript 1.

$$p_1 f' \equiv -(A-1)L'G_1(A-1)L'f'. \tag{2.4}$$

Therefore, Eq. (2.1) becomes

$$(L_0 + p_1)f' = -(A-1)L'G_1(A-1)L'G_1(A-1)L'f' + (A-1)L'G_1(A-1)L'\bar{f} + (A-1)L'\bar{f}, \tag{2.5}$$

where we have used the relation $AL'\bar{f} = 0$. We have neglected initial conditions for f', but they may be kept with only a little added complexity.

Defining $K = G(A-1)L'$, the solution for f' is

$$f' = \sum_{n=0}^{\infty} (-K^3)^n (K^2 + K)\bar{f}, \tag{2.6a}$$

$$= \sum_{n=0}^{\infty} (1-K)^n K^{n+1}\bar{f}. \tag{2.6b}$$

The reduction from Eq. (2.6a) to Eq. (2.6b) is accomplished using trivial operator identities.

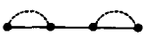
Let us compare the diagram expansions of Eqs. (1.4) and (2.6) using the following conventions. L' is a vertex, represented by a dot: "$\cdot$". L_0^{-1} is represented by a straight line. Any perturbed operator, G, is represented by a wavy line "$\sim\sim$." If two quantities are ensemble averaged, we draw a dotted line between them; e.g.,

$AL'L_0^{-1}L' =$.

The factors of $(A-1)$ mean that, when the final averaging is completed, we cannot have an unconnected diagram; that is, one that a cut can be made through between vertices without touching a dotted line. An example of a connected diagram is

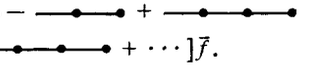
 .

An example of an unconnected diagram is

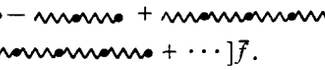
 .

We shall not retain the factors of $(A-1)$, but simply remember that, when the final averaging is done, unconnected diagrams are forbidden.

Then the diagram expansion for Eq. (1.4) is

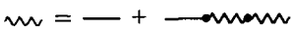
$f' = -[$  $+ \cdots]\bar{f}.$ (2.7)

While the diagram expansion for Eq. (2.6) is

$f' = -[$  $+ \cdots]\bar{f}.$ (2.8)

Note that the terms with $3n$ vertices are missing. They have been absorbed into the Green's function. The perturbation expansion shown in Eq. (2.8) appears simple, but averaging it is difficult, since G is stochastic.

The diagram form for the Green's function equation is

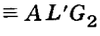
 (2.9)

B. Nonstochastic Green's function

Now we take the next approach to defining the Green's function. If we consider the stochastic differential Eq. (1.1) as defining the evolution of the function f in the phase space of independent variables, then the Green's function is a trajectory operator. The average Green's function then takes on the character of a probability distribution function for quantities defined along the statistical trajectories.

In order that the Green's function equation is the same as the Kolmogorov equation for the probability distribution function, we define

$$p_2 f' = -\langle L'G_2L'\rangle f', \tag{2.10}$$

where $\langle L'G_2L'\rangle \equiv AL'G_2L' =$  . Eq. (2.1) becomes

$$(L_0 + p)f' = [(A-1)L' - \langle L'G_2L'\rangle]f' - L'\bar{f}, \tag{2.11}$$

with the solution

$$f' = \sum_{n=0}^{\infty} \{G_2(A-1)L' - G_2\langle L'G_2L'\rangle\}^n G_2(A-1)L'\bar{f}. \tag{2.12}$$

We now substitute Eq. (2.12) into Eq. (1.2) to obtain

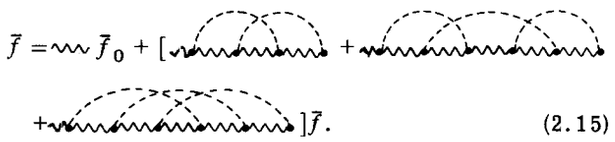
$$[L_0 - \langle L'G_2L'\rangle]\bar{f} = \sum_{n=1}^{\infty} \langle L'\{G_2(A-1)L' - G_2\langle L'G_2L'\rangle\}^n \times G_2L'\rangle\bar{f}. \tag{2.13}$$

The solution to Eq. (2.13) is

$$\bar{f} = G_2\bar{f}_0 + G_2 \sum_{n=1}^{\infty} \langle L'\{G_2(A-1)L' - G_2\langle L'G_2L'\rangle\}^n G_2L'\rangle\bar{f}, \tag{2.14}$$

where $\bar{f}_0$ is the given initial condition. Eq. (2.14) appears complex, but has a simple diagram representation.

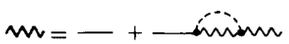
Let us neglect triple and higher order correlations; that is, we apply the random phase approximation. This is consistent with our basic assumption of Markov interactions. Eq. (2.14) becomes



(2.15)

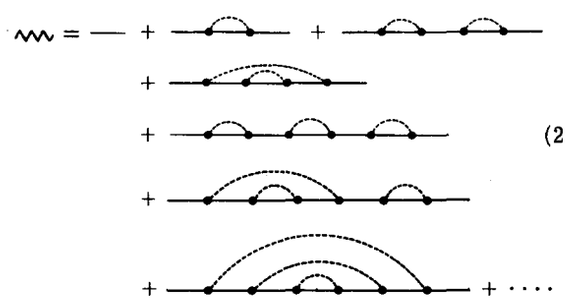
The rules for this diagram expansion are that (1) every dotted line must connect two vertices; (2) no unconnected diagrams are permitted; (3) all lines must cross.

The equation for G_2 is



(2.16)

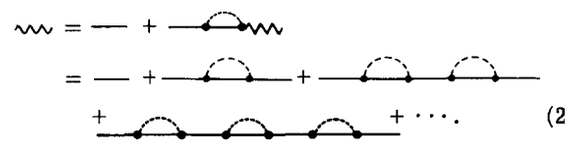
With the formal expansion,



(2.17)

Firsch[4] calls Eq. (2.16) the *Kraichnan equation*. Kraichnan derived it as a complete Green's function to a "model" problem, i.e., one whose stochastic dynamics is modeled by the theoretician in such a way as to produce an exactly solvable problem, but whose relationship to physical reality may not be clear.[7] Here we have developed the Kraichnan Green's function as an incomplete Green's function for the physical system, and have explicitly shown the resulting perturbation series.

Compare Eq. (2.17) with the equivalent expansion for the first order smoothed Green's function:



(2.18)

It is clear that Eq. (2.17) for G_2 contains many more terms than does Eq. (2.18). In particular, it contains all diagrams with two-vertex clusters and no crossing of lines. Note that the diagrams for G_2 may be unconnected. It is tempting to ascribe physical content to the individual terms of the perturbation series. As pure conjecture—but perhaps a useful one—we suppose that terms of the form

represent events occuring at different times and different places. We call these *sequentially unconnected diagrams*. In turn, a diagram such as

corresponds to interactions with the stochastic field at either different times at the same place, or different places at the same time. Both sorts of events are included because of the diffusive nature of the trajectory. We call these *nested diagrams*. Finally, interactions of a particle with the stochastic field at the same time and same place have the form

These give rise to non-Markovian corrections to the particle trajectory. These are *crossed diagrams*.

This classification points to an interesting interpretation of Eqs. (2.17) and (2.18). Both Green's functions contain all sequentially unconnected diagrams. All nested diagrams appear in Eq. (2.17), however, indicating that all Markovian interactions are present, and indeed, this Green's function does correspond to the propagator derived from the statistical view. The Green's function of Eq. (2.18) in some sense corresponds to a truncated picture of the Markovian process. It resolves the secularity difficulty, but the precise physical process it describes is not clear.

3. APPLICATION TO THE VLASOV EQUATION

The rest of this paper will apply specifically to problems of plasma turbulence, which essentially involves coupled solutions of the Vlasov equation

$$\left(\frac{\partial}{\partial t} + \mathbf{v}\cdot\frac{\partial}{\partial R} + q/m\mathbf{E}\cdot\frac{\partial}{\partial \mathbf{v}}\right)f = 0 \tag{3.1}$$

and Poisson's equation

$$\nabla\cdot\mathbf{E} = (4\pi q/c)\int dv f. \tag{3.2}$$

We shall not consider in detail the solution to Poisson's equation; rather, we assume that the turbulent spectrum is known and that we are interested in the time evolution of $\bar{f}$. For simplicity of analysis, we let $AE = 0$ and only consider one dimension.

The suggestion to use a perturbed operator was made by Dupree.[2] According to the formalism of this paper.

$$L_0 = \frac{\partial}{\partial t} + v\frac{\partial}{\partial R}, \qquad L' = q/mE'\frac{\partial}{\partial v}.$$

It the turbulence spectrum is very broad, the autocorrelation time of the fields is very short. In this case the term becomes

$$(q^2/m^2)\frac{\partial}{\partial v}\langle E'GE'\rangle\rangle\frac{\partial}{\partial v} \equiv \frac{\partial}{\partial v}D\frac{\partial}{\partial v}, \tag{3.3}$$

with D a function (the diffusion coefficient) rather than an operator. The equation for the Green's function (2.16) can be written

$$\left(\frac{\partial}{\partial t} + v\frac{\partial}{\partial R} - \frac{\partial}{\partial v}D\frac{\partial}{\partial v}\right)g\,(Rvt/R_0v_0t_0) = 0,$$
$$g(Rvt_0/R_0v_0t_0) = \delta(R - R_0)\delta(v - v_0), \tag{3.4}$$

where

$$D = (q^2/m^2)\int_{t_0}^{t} d\tau\int dR'dv'\langle E(R,t)g(Rvt/R'v'\tau)E(R',\tau)\rangle. \tag{3.5}$$

We cannot solve Eq. (3.4) for arbitrary $D(v)$, but in the special case that $D(v) \approx$ const., it can be solved. This is

again a restraint on the broad banded nature of the turbulent spectrum (see Ref. 8 for a more detailed discussion). In this case[9]

$$g(Rvt/R_0v_0t_0) = \frac{\exp\left[-(v-v_0)^2/4D\tau\right]}{\sqrt{4\pi D\tau}}$$

$$\times \frac{\exp\left[-(R-R_0-\tfrac{1}{2}(v+v_0)\tau)^2/1/3D\tau^3\right]}{\sqrt{\tfrac{1}{3}\pi D\tau^3}}, \qquad (3.6)$$

where $\tau \equiv t - t_0$. Substituting Eq. (3.6) into Eq. (3.5), we obtain Dupree's result

$$D = (q^2/m^2)\int \frac{dkd\omega}{(2\pi)^2}\langle E^*_{k\omega}E_{k\omega}\rangle \int_0^{t-t_0} d\tau \times \exp[i(\omega - kv)\tau - \tfrac{1}{3}k^2D\tau^3]. \qquad (3.7)$$

Note the damping term $-\frac{1}{3}k^2D\tau^3$ in the exponential. The unperturbed operator gives an analogous integral in the perturbation expansion:

$$\int_0^{t-t_0} d\tau \exp[i(\omega - kv)\tau].$$

One source of secularity is derived from the action of the $\partial/\partial v$ operator on this integral:

$$\frac{\partial}{\partial v}\int_0^{t-t_0} d\tau \exp[i(\omega - kv)\tau] = ik\int_0^{t-t_0} \tau d\tau \exp[i(\omega - kv)\tau].$$

However, the damping term derived here keeps the integral bounded as t goes to infinity. The other source of secularity is repeated integrations over time, such as

$$\int_0^t dt_1 \exp[i(\omega - kv)t_1]\int_0^{t_1} dt_2 \exp[i(\omega - kv)t_2],$$

which arise whenever one calculates correlated quantities. The damping term likewise removes this source of secularity.

The requirement that this perturbation series converge is seen to be[4]

$$\left\| \frac{\partial}{\partial v} GL' \right\|_2 < 1, \qquad (3.8)$$

where $\|\cdot\|_2$ is the norm in the space of square integrable functions. This requirement may be written

$$\int \frac{dkd\omega}{(2\pi)^2}\, \omega_B^2(k)\tau_{AC}^2(k) < 1, \qquad (3.9)$$

where

$$\omega_B^2(k) \equiv (qk/m)\langle E^*_{k\omega}E_{k\omega}\rangle^{\frac{1}{2}},$$

$$\tau_{AC}^2(k) \equiv \int_0^\infty \tau d\tau \exp i[(\omega - kv)\tau - \tfrac{1}{3}k^2D\tau^3].$$

τ_{AC} is a measure of the autocorrelation time of the kth mode. $\omega_B(k)$ is called the bounce frequency, since a particle trapped in the potential trough of an electric wave E_k will oscillate at this frequency.

Thus Eq. (3.9) means that no particle is trapped. The Green's function equation is diffusive, and we expect it to approximate the average nature of the particle trajectories. If a particle is trapped by a wave, obviously its trajectory will not be diffusive. Thus, the convergence criterion agrees with our heuristic expectations.

We have also derived Eqs. (3.6) and (3.7) purely on the basis of finding the joint probability distribution function in R and v for a test particle undergoing a Markovian interaction with external fields (see Refs. 1 and 8). The equivalence of the two operators justifies terming the Green's function given in Eq. (3.6) a Markovian propagator. Note that the signature of such a propagator is the absence of all diagrams with crossings.

We may use the propagator we have found to evaluate Poisson's equation and give the dielectric response. This will appear in a later paper.

4. DIFFUSION COEFFICIENT

We would like to make a final comment on the subject of evaluation of Eq. (3.7) for D, the diffusion coefficient. The usual suggestion is solution by iteration, taking as the first guess $D_0 =$; i.e., the diffusion coefficient obtained from quasilinear theory, and also that used in the first order smoothing approximation. The question we wish to resolve is whether the diffusion coefficient of Eq. (3.7) actually results from such an iterative procedure.

First we obtain a formal solution to the Kraichnan equation in the form of a continued fraction of operators.[4] Denoting by $\mathcal{L}$ the linear operator

$$\mathcal{L}[G] = \langle L'GL'\rangle.$$

We can write G as (Eq. (2.16))

$$G = L_0^{-1} + L_0^{-1}\mathcal{L}[G]G = \{1 - L_0^{-1}\mathcal{L}[G]\}^{-1}L_0^{-1}$$

and iterate to obtain

$$G = \frac{L_0^{-1}}{1 - L_0^{-1}\mathcal{L}}\left[\frac{L_0^{-1}}{1 - L_0^{-1}\mathcal{L}}\left[\frac{L_0^{-1}}{1 - \cdots}\right.\right. \qquad (4.1)$$

For the first iteration for D, we guess

$$D_0 = \mathcal{L}[L_0^{-1}].$$

The propagator corresponding to this approximation is

$$G_1 = L_0^{-1} + L_0^{-1}\mathcal{L}[L_0^{-1}]G_1 = \{1 - L_0^{-1}\mathcal{L}[L_0^{-1}]\}^{-1}L_0^{-1}.$$

The next iteration for D is

$$D_1 = \mathcal{L}[G_1],$$

with the next order propagator

$$G_2 = \{1 - L_0^{-1}\mathcal{L}[G_1]\}^{-1}G_2 = \frac{L_0^{-1}}{1 - L_0^{-1}\mathcal{L}}\left[\frac{L_0^{-1}}{1 - \mathcal{L}[L_0^{-1}]}\right. \qquad (4.2)$$

As we continue the iteration, it is seen that we are constructing the continued fraction expansion of Eq. (4.1). Thus, the iteration will converge to D, and $G_\infty = G$.

5. CONCLUSION

We have tried to indicate the connection between Dupree's theory of plasma turbulence using a perturbed

Green's function, and classic methods of solving stochastic differential equations by perturbation expansions. In particular, we have rigorously derived the perturbation series obtained with the Kraichnan Green's function and shown that the resummation of the series involves grouping all diagrams with no line crossings into the Green's function.

We then solved the self consistent equation for the Green's function under the basic assumption of sufficiently broad band turbulence. To our knowledge, this is the only exact solution so far given for such a generalized Green's function.[10] Our solution only applies to a special class of problems, but the basic results (previously derived by Dupree and his co-workers) seem to have been verified in several experimental investigations.[11,12,13,14] This particular solution may serve as a starting point to investigate more general solutions.

ACKNOWLEDGMENT

The authors would like to thank Dr. D. S. Bugnolo for suggesting that classic analyses of wave propagation in random media would complement our previous work.

APPENDIX

In this appendix, we compare our analysis with that of Weinstock.[15,16] He obtains formal results for $\bar{f}$ and the self-consistent field E through the use of a Green's function that incorporates the averaging operation. In his notation $L_0 = \partial/\partial t + \bar{L}$. Our Eq. (2.1) becomes

$$\left(\frac{\partial}{\partial t} + \bar{L} + p\right) f' = pf' + (A-1)L'f' - L'\bar{f}. \tag{A1}$$

Rather than iterating for f' at this stage, we define p so as to cancel the term on the right hand side that contains f':

$$pf' \equiv -(A-1)L'f'.$$

The Green's function satisfies

$$\left(\frac{\partial}{\partial t} + \bar{L} + (1-A)L'\right) G_A = 0, \tag{A2}$$

with the solution

$$\begin{aligned} G_A &= L_0^{-1} - L_0^{-1}(1-A)L'G_A \\ &= \sum_{n=0}^{\infty} [L_0^{-1}(A-1)L']^n L_0^{-1}. \end{aligned} \tag{A3}$$

The solution for f' is thus

$$\begin{aligned} f' &= -G_A L'\bar{f} \\ &= -\sum_{n=0}^{\infty} [L_0^{-1}(A-1)L']^n L_0^{-1} L'\bar{f}, \end{aligned} \tag{A4}$$

which is identical to Eq. (1.4). Weinstock's Green's function is not exactly the same as G_A, however. Using the fact that $AL'\bar{f} = 0$ and $A\bar{L}f' = 0$, we can write Eq. (A.1) as

$$\left(\frac{\partial}{\partial t} + (1-A)\bar{L} + P\right) f' = pf' + (A-1)L'f' + (A-1)L'\bar{f}. \tag{A5}$$

Making the same definition for p, the Green's function equation becomes

$$\left(\frac{\partial}{\partial t} + (1-A)(\bar{L} + L')\right) U_A = 0. \tag{A6}$$

This defines Weinstock's Green's function, U_A. The solution to this equation is

$$U_A = \sum_{n=0}^{\infty} [L_0^{-1}\bar{L}A + L_0^{-1}(A-1)L']^n L_0^{-1}. \tag{A7}$$

It is clear that $U_A \neq G_A$. However, its action on $L'\bar{f}$ is the same, as can be seen by direct calculation. Therefore,

$$\begin{aligned} f'(t) &= -U_A L'\bar{f} \\ &= \sum_{n=0}^{\infty} [L_0^{-1}\bar{L}A + L_0^{-1}(A-1)L']^n L_0^{-1} L'\bar{f} \\ &= -\sum_{n=0}^{\infty} [L_0^{-1}(A-1)L']^n L_0^{-1} L'\bar{f} \\ &= -G_A L'\bar{f}. \end{aligned} \tag{A7'}$$

Weinstock also includes the initial conditions in the solution for f', which is not a difficult step. We did not include them for several reasons: first, we wished to demonstrate in the clearest manner the diagram expansion, eliminating extraneous details. Second, most calculations done with this method apply to time scales long enough so that the effects of initial conditions are forgotten. For an example of the opposite situation, see Ref. 9.

We have demonstrated that Weinstock's U_A is a complete Green's function for the problem. However, it is too complicated to solve exactly, and must be expanded in terms of simpler operators. If the unperturbed operator is used, we are back to the results of ordinary perturbation theory. However, Weinstock makes an expansion in terms of $\bar{U}$, the average Vlasov operator, which, to first order in E^2, equals our Green's function (2.16). Thus, in the limit of weak turbulence, his result approaches ours (and Dupree's). The congruence of his theory and Dupree's in this limit was demonstrated in Ref. 15. A point we emphasize is that the solution using the U_A operator is a formal result. It appears that application to real systems demands a way of expressing the full U_A operator in a meaningful manner, i.e., one which makes further calculation possible. We suggest that the Green's function derived in the body of this paper is the proper vehicle for the expansion, which becomes (neglecting the term $\bar{L}A$, which produces no results),

$$U_A = \sum_{n=0}^{\infty} [G(A-1)L' - G\langle L'GL'\rangle]^n G. \tag{A8}$$

Substituting (A.8) into (A.7), we obtain our previous result, Eq. (2.12).

---

*Work supported by the National Science Foundation under Grant No. GP 28656.

[1] J. J. Thomson and Gregory Benford, Phys. Rev. Lett. **28**, 590 (1972).
[2] T. H. Dupree, Phys. Fluids **9**, 1773 (1966).
[3] J. Weinstock, Phys. Fluids **11**, 1977 (1968).
[4] U. Frisch, in *Probabilistic Methods in Applied Mathematics*, edited by A. T. Bharucha-Reid (Academic, New York, 1968), Vol. 1.
[5] V. I. Tatarski and M. E. Gersenshtein, Zh. Eksp. Teor. Fiz. **17**, 458 (1963).
[6] I. M. Besieris, J. Math. Phys. **13**, 358 (1972).
[7] S. A. Orszag and R. H. Kraichnan, Phys. Fluids **10**, 1720 (1967).
[8] Gregory Benford and J. J. Thomson, Phys. Fluids **15**, 1496 (1972).
[9] F. L. Hinton and C. Oberman, Phys. Fluids **11**, 1982 (1968). They derived the short time solution to a Fokker–Planck equation for a test particle. The three dimensional analog to Eq. (3.6) appears in this paper.

[10]Our referee points out that this statement is not quite correct, since an exact solution for a special problem may be found in U. Frisch and R. C. Bourrett, "Parastochastics," Inst. d'Astrophys., Paris, 1967.
[11]M. Lampe, *et al*., Phys. Fluids **15,** 662 (1972).
[12]T. H. Dupree, Phys. Fluids **11,** 2680 (1968).
[13]R. W. Flynn, Phys. Fluids **14,** 956 (1971).
[14]W. S. Williamson, G. Benford, J. J. Thomson, and N. Rynn (to be published), Technical Report 72-45, University of California, Irvine.
[15]J. Weinstock, Phys. Fluids **12,** 1045 (1969).
[16]J. Weinstock, Phys. Fluids **13,** 2308 (1970).

High density expansions in the Thomas–Fermi approximation

Michael Nauenberg

Division of Natural Sciences, University of California, Santa Cruz, California
(Received 15 September 1972)

Two expansions are discussed for the solution of the Thomas–Fermi equation for a degenerate electron gas at high density. In the first case, the boundary conditions appropriate to the Wigner-Seitz sphere are imposed, and terms up to fourth order are calculated explicitly to show the appearance of a logarithmic singularity in the density. This singularity accounts for the divergence difficulties in higher order encountered by previous authors, who assumed the existence of a power series expansion in the density. In the second case, we apply the boundary conditions appropriate to electron screening of an ion in a neutral plasma at high density, and an expansion is obtained which gives the nonlinear corrections to the Debye-Hückel approximation.

INTRODUCTION

The Thomas–Fermi[1,2] approximation for a degenerate electron gas has been applied extensively to obtain properties of atoms and nuclei.[3] Since the Thomas–Fermi differential equation is non-linear, solutions have generally been obtained by numerical methods. However, in the case of matter at high densities, it is possible to obtain an analytic solution as an expansion in the density. Several authors[4,5,6] have assumed that this expansion is a power series in the cube root of the electron density, but they have encountered divergence difficulties in computing higher order terms. In this paper, we discuss expansions corresponding to two different boundary conditions of physical interest. In the first case, to obtain the equation of state, the boundary condition is that the average number of electrons inside the Wigner–Seitz sphere[7] equal the charge Z of the ion at the center. In the second case, to obtain the electron screening of an ion in a neutral plasma, the boundary condition is that the electron density asymptotically approach a constant. In both cases, we find that the expansions of the solution of the Thomas–Fermi equation contain terms which are logarithmic in the density, and the lowest-order contributions are explicitly calculated.

1. WIGNER-SEITZ BOUNDARY CONDITION

The Thomas–Fermi equation[1,2] for an electron gas in the presence of an ion of charge Z is

$$\frac{d^2\psi}{dx^2} = \frac{\psi^{3/2}}{x^{1/2}}, \tag{1.1}$$

where x is the distance from the ion in units of $l = a_0(3\pi/4)^{2/3}/2Z^{1/3} = 0.885\ 34\ a_0/Z^{1/3}$, $a_0 = \hbar^2/me^2$ is the Bohr radius, and $\psi(x)$ determines the Coulomb potential $(Ze/l)[\psi(x)/x]$ and the electron density $(Z/4\pi l^3)[\psi(x)/x]^{3/2}$. The boundary conditions at the origin and at the surface of the Wigner–Seitz sphere[7] are

$$\psi(0) = 1 \quad \text{and} \quad x_0 \frac{d\psi(x_0)}{dx_0} = \psi(x_0), \tag{1.2}$$

where x_0 is then the radius of a sphere containing Z electrons or, alternatively, the radius at which the electric field of the electrons cancels the field of the ion. Hence, the mean density of n_0 of electrons is

$$n_0 = 3Z/4\pi l^3 {x_0}^3. \tag{1.3}$$

We discuss now an analytic expansion of ψ for high electron density, which is defined by the condition $x_0 \ll 1$. It is tempting to use x_0 as the small parameter for a power series expansion of ψ, as has been done by several authors,[4,5,6] but this procedure is incorrect in higher orders, as we shall show below. On physical grounds, we expect that in the high-density limit the electron density is nearly uniform except at distances very close to the ion. This suggests that the lowest-order approximation to ψ which incorporates the boundary condition at the origin, Eq. (1.2), has the form $1 + cx$, where $c \gg 1$. We therefore write ψ in the form

$$\psi(x) = 1 + cx + f(x), \tag{1.4}$$

and derive an equation for the correction function f. We require that the constant c be the slope of ψ at the origin, and introduce $y = cx$ as the independent variable instead of x. Hence, f satisfies the differential equation

$$\frac{d^2f(y)}{dy^2} = \epsilon \frac{[1 + y + f(y)]^{3/2}}{y^{1/2}}, \tag{1.5}$$

where

$$\epsilon = 1/c^{3/2}. \tag{1.6}$$

The boundary condition at $y = 0$ is

$$f = \frac{df}{dy} = 0, \tag{1.7}$$

while the boundary condition at the surface of the Wigner–Seitz sphere, Eq. (1.2), becomes

$$y_0 \frac{df}{dy_0} - f(y_0) = 1, \tag{1.8}$$

and determines the relation between ϵ and y_0. The solution of Eq. 1.5, with the boundary condition Eq. (1.7), is analytic near $\epsilon = 0$. Therefore, for small ϵ we can solve this equation by expanding f in a Taylor series in ϵ,

$$f(y) = \sum_{n=1}^{\infty} \epsilon^n f_n(y). \tag{1.9}$$

Substituting Eq. (1.9) in Eq. (1.5) and equating powers of ϵ, we obtain equations for f_n which depend only on f_m for $m < n$. To illustrate this procedure, the first four of these equations are given below:

$$\frac{d^2f_1}{dy^2} = \frac{(1+y)^{3/2}}{y^{1/2}}, \tag{1.10}$$

$$\frac{d^2f_2}{dy^2} = \frac{3}{2}\frac{(1+y)^{1/2}}{y^{1/2}} f_1, \tag{1.11}$$

$$\frac{d^2f_3}{dy^2} = \frac{3}{2}\frac{(1+y)^{1/2}}{y^{1/2}} f_2 + \frac{3}{8}\frac{f_1^2}{y^{1/2}(1+y)^{1/2}}, \tag{1.12}$$

$$\frac{d^2f_4}{dy^2} = \frac{3}{2}\frac{(1+y)^{1/2}}{y^{1/2}}f_3 + \frac{6}{8}\frac{f_1f_2}{y^{1/2}(1+y)^{1/2}} - \frac{1}{16}\frac{f_1^3}{y^{1/2}(1+y)^{3/2}}. \quad (1.13)$$

These equations are then integrated in sequence, satisfying the boundary conditions at $y = 0$,

$$f_n = \frac{df_n}{dy} = 0. \quad (1.14)$$

Finally, the solutions are then substituted in Eq. (1.8) to determine the relations between ϵ and y_0. The integration of Eq. (1.10), which satisfies the boundary condition, Eq. (1.14), is elementary and gives

$$f_1 = \frac{3}{8}(y\ln y - y) + \frac{3}{4}\frac{\ln(z+1)}{(z-1)} + \frac{1}{16}\ln\left(\frac{z+1}{z+1}\right) - \frac{3}{8(z+1)} + \frac{z}{4(z+1)} + \frac{2}{3}\frac{z}{(z^2-1)^2} + \frac{1}{6}\frac{z}{(z^2-1)^3}, \quad (1.15)$$

where $z = (1 + 1/y)^{1/2}$.

For an expansion up to fourth order in ϵ, we require only the asymptotic form of the solutions of Eqs. (1.11)–(1.13) for large y, and the boundary condition at the origin, Eq. (1.14), will actually not enter. The reason for this is that the boundary condition at the surface of the Wigner-Seitz sphere, Eq. (1.8), implies that $\epsilon y_0^3 \approx 1$. Therefore, we need not calculate terms of f_n for large y of order smaller than y^{3n-3}. The integration is therefore elementary, and we obtain [including the asymptotic expansion of Eq. (1.15)]

$$f_1 = \tfrac{1}{6}y^3 + \tfrac{3}{4}y^2 + \tfrac{3}{8}y\ln y + y(\tfrac{3}{4}\ln 2 + \tfrac{3}{16})\ln y\,\tfrac{1}{16} + \tfrac{1}{16}\ln 4 + \tfrac{5}{96}, \quad (1.16)$$

$$f_2 = \tfrac{1}{80}y^5 + \tfrac{5}{48}y^4 + \tfrac{3}{32}y^3\ln y + \tfrac{1}{16}y^3\left(\tfrac{11}{12} + 3\ln 2\right), \quad (1.17)$$

$$f_3 = y^7/1440 + \tfrac{61}{7200}y^6, \quad (1.18)$$

$$f_4 = y^9/31104. \quad (1.19)$$

Substituting Eqs. (1.16)–(1.18), evaluated at $y = y_0$, in Eq. (1.8) gives an implicit relation between ϵ and y_0. It is clear that f_1 and f_2 contribute terms which are logarithmic in y_0. After some algebra, we find the initial slope $c = 1/\epsilon^{2/3}$ in terms of x_0:

$$c = \frac{3^{2/3}}{x_0} - \frac{9}{5x_0} - \frac{3^{7/3}}{25} + \frac{3^{2/3}}{9}x_0 \times (\ln x_0 + \tfrac{1}{8}\ln 4 - \tfrac{9}{4}\ln 2 - \tfrac{2}{3}\ln 3 + \tfrac{7819}{9000}), \quad (1.20)$$

which then determines the solution ψ as a function of the physical parameter x_0. In particular, at $x = x_0$ we obtain

$$\psi(x_0) = \frac{3^{2/3}}{x_0} - \frac{3}{10} - \frac{3^{7/3}}{400}x_0 - \frac{3^{2/3}}{72} \times x_0^2(\ln x_0 - \ln 4 - \tfrac{2}{3}\ln 3 + \tfrac{8819}{4500}). \quad (1.21)$$

The first two terms in Eqs. (1.20) and (1.21) agree with the result of March.[4] However, his expansion method assumed the existence of a power series expansion in x_0, and therefore he encountered divergences in the higher orders. From Eqs. (1.20) and (1.21), we then obtain the high-density expansion for the total energy $E = (z^2e^2/l)[\frac{2}{35}x_0^{1/2}\psi^{5/2}(x_0) + \frac{3}{7}c]$ of the electron gas,

$$E = \frac{z^2e^2}{l}\left[\frac{3^{5/8}}{5x_0^2} - \frac{9}{10x_0} - \frac{3^{10/3}}{175} + \frac{3^{5/3}}{63}x_0\,(\tfrac{7}{8}\ln x_0 + \tfrac{1}{4}\ln 4 - \tfrac{7}{12}\ln 3 - \tfrac{9}{4}\ln 2 + \tfrac{23429}{36000})\right]. \quad (1.22)$$

The first term in this expansion corresponds to the kinetic energy of a uniform electron gas, while the second term is the Coulomb energy. The third term was first obtained by Salpeter[8] by applying the variational properties of the energy integral that leads to the Thomas-Fermi equation, and later on by Salpeter and Zapolski[6] from a power series expansion in x_0, but their methods also lead to divergent results in higher order. We have verified Eq. (1.22) by comparing it with the result of numerical integrations of the Thomas-Fermi equation obtained by G. Villere.

The pressure P of the electron gas can be obtained directly from the energy E, Eq. (1.22), from the relation $P = -dE/dv$, where $v = 4\pi l^3/3x_0^3$ is the volume of the Wigner-Seitz sphere, or, alternatively, by the relation $Pv = \frac{2}{15}x_0^{1/2}\psi^{5/2}(x_0)$,

$$Pv = \frac{z^2e^2}{3l}\left[\frac{2}{5}\frac{3^{5/3}}{x_0^2} - \frac{9}{10x_0} - \frac{3^{5/3}}{63}x_0 \times\left(\frac{7}{8}\ln x_0 + \frac{1}{4}\ln 4 - \frac{7}{12}\ln 3 - \frac{9}{4}\ln 2 + \frac{109858}{72000}\right)\right]. \quad (1.23)$$

Finally, we note that a similar expansion procedure can also be used to solve the Thomas-Fermi-Dirac[9] equation, which includes the electron exchange contribution.

2. THOMAS-FERMI SCREENING

The Thomas-Fermi approximation has also been applied[10,11,12] to obtain the screening of the Coulomb field of an ion by an electron gas in a neutral plasma. The effect of a uniform positive background of charge with density n_0 modifies Eq. (1.1) into the form

$$\frac{d^2\psi}{dx^2} = \frac{\psi^{3/2}}{x^{1/2}} - \eta x, \quad (2.1)$$

where $\eta = 4\pi l^3 n_0/Z$. The dominant term of the solution of Eq. (2.1) at high density is $\eta^{2/3}x$, and it can readily be seen that the remainder falls exponentially for large values of x. Hence, we assume the form

$$\psi(x) = \eta^{2/3}x + \xi(x)e^{-\alpha x}, \quad (2.2)$$

and find $\alpha = \sqrt{\frac{3}{2}}\,\eta^{1/6}$, while ξ satisfies the equation

$$\frac{d^2\xi}{dx} - 2\alpha\frac{d\xi}{dx} = f(x,\xi), \quad (2.3)$$

where

$$f(x,\xi) = \left(\frac{(\eta^{2/3}x + \xi e^{-\alpha x})^{3/2}}{x^{1/2}} - \eta x\right)e^{\alpha x} - \alpha^2\xi. \quad (2.4)$$

The boundary conditions for Eq. (2.3) are

$$\xi(0) = 1 \quad \text{and} \quad \lim_{x\to\infty}\xi(x) = \xi(\infty), \quad (2.5)$$

where $\xi(\infty)$ is a function of η to be determined below, Eq. (2.10). These conditions can be satisfied by converting Eq. (2.3) into the nonlinear integral equation

$$\xi(x) = 1 + \int_0^\infty K(x, x') f[x', \xi(x')], \tag{2.6}$$

where the kernel K has the form

$$K(x, x') = \begin{cases} (e^{-2\alpha x'} - 1)/2\alpha, & x' < x, \\ [(1 - e^{2\alpha x})/2\alpha] e^{-2\alpha x'}, & x < x'. \end{cases} \tag{2.7}$$

For $\alpha x \gg 1$, the function f has the asymptotic form

$$f(x, \xi) \cong \frac{3}{8} \frac{\xi^2}{\eta^{1/3}} \frac{e^{-\alpha x}}{x}. \tag{2.8}$$

Hence, we can solve Eq. (2.6) by successive iterations. A useful starting function is $\xi_0(x) = 1$, which we substitute in the integrand of Eq. (2.6). For $\eta \gg 1$, this first iteration leads to a high-density approximation for $\xi(x)$. For $x \lesssim 1/\eta^{2/3}$, we obtain

$$\begin{aligned} \xi(x) \cong 1 &- (x/4\eta^{1/3}) \left[\tfrac{3}{4}\ln\left(\tfrac{2}{3}\eta\right) - \tfrac{33}{4} + 7\sqrt{2} + 3\ln(\sqrt{2}+1)\right] \\ &+ (2/3\eta)\left\{(z+1)^{1/2} z^{1/2} \left(\tfrac{1}{4}z^2 + z - \tfrac{3}{16}\right) + \tfrac{3}{8}(3z \right. \\ &\left. + \tfrac{1}{2}\ln[(z+1)^{1/2} + z^{1/2}]\right\} - \tfrac{3}{4}\eta^{1/3}x^2 - \tfrac{1}{6}\eta x^3, \end{aligned} \tag{2.9}$$

where $z = \eta^{2/3}x$, while for $x \gg 1/\eta^{2/3}$

$$\xi(x) \cong \xi(\infty) \cong 1 - (1/8\eta)\left[\tfrac{3}{2}\ln\left(\tfrac{2}{3}\eta\right) - \tfrac{55}{6} + \tfrac{25}{3}\sqrt{2} + \ln(\sqrt{2}+1)\right]. \tag{2.10}$$

These equations give the high-density nonlinear corrections to the Debye–Hückel approximation, which corresponds to setting $\xi(x) = 1$. However, it should be remembered that the Thomas–Fermi equation cannot be used to obtain asymptotic screening for a degenerate electron gas, as was first pointed out by Friedel.[13]

[1] L. H. Thomas, Proc. Camb. Phil. Soc. 23, 542 (1927).
[2] E. Fermi, Z. Phys. 48, 73 (1928).
[3] For a detailed review, see N. H. March, Advan. Phys. 6, 1 (1957).
[4] N. H. March, Proc. Phys. Soc. A. 68, 726 (1955).
[5] J. J. Gilvarry, Phys. Rev. 96, 934 (1954).
[6] E. E. Salpeter and H. S. Zapolsky, Phys. Rev. 158, 876 (1967).
[7] E. Wigner and E. Seitz, Phys. Rev. 43, 804 (1933); 46, 509 (1934).
[8] E. E. Salpeter, Astrophys. J. 134, 669 (1961).
[9] P. A. M. Dirac, Proc. Camb. Phil. Soc. 26, 376 (1930).
[10] N. F. Mott, Proc. Camb. Phil. Soc. B2, 281 (1936).
[11] H. Fujiwara, J. Phys. Soc. Japan 339, 727 (1955).
[12] L. C. R. Alfred and N. H. March, Phil. Mag. 46, 759 (1955); Phys. Rev. 103, 877 (1956).
[13] J. Friedel, Advan. Phys. 3, 446 (1954).

On Litvin's lemma

P. H. Butler

Department of Mathematics, University of Toronto, Toronto 181, Ontario, Canada
(Received 8 November 1972)

We produce a counterexample to a lemma given by D. B. Litvin on the decomposition of direct products of irreducible representations [J. Math. Phys. **13**, 1386 (1972)]. The error in his proof is noted.

In a recent paper Litvin[1] claims to have proven the following lemma.

Litvin: "The direct product of two type A (orthogonal) or two type B (symplectic) irreducible representations of an arbitrary decomposable unitary group does not contain type B representations; the direct product of a type A and a type B irreducible representation does not contain type A irreducible representations."

Litvin does not seem aware of Mal'cev's lemmas.[2]

Mal'cev: "The Kronecker product of two contragredient representations is orthogonal. The Kronecker product of two orthogonal or two symplectic representation is orthogonal, and the Kronecker product of an orthogonal and a symplectic representation is symplectic." One needs to use his results on reducible representations also. Dynkin[3] quotes these as, "In order that a reducible representation ϕ be symplectic (orthogonal) it is necessary and sufficient that its decomposition into irreducible components has the form

$$\phi = \phi_1 \dot{+} \cdots \dot{+} \phi_k + \psi_1 \dot{+} \hat{\psi}_1 \dot{+} \cdots \dot{+} \psi_l \dot{+} \hat{\psi}_l$$

where the representations $\phi_1, \phi_2, \cdots \phi_k$ are symplectic (or orthogonal, respectively) and the representations ψ_i and $\hat{\psi}_i$ $(i = 1, 2, \ldots, l)$ are contragredient to each other."

Note that these lemmas do not rule out the possibility of orthogonal (or symplectic) representations occurring, as long as they occur in pairs. (The proofs apply to finite groups equally.)

In a recent study[4] of the properties of jm and j symbols for an arbitrary (finite or compact) group, I obtain Mal'cev result using techniques which were very similar to Litvin's. In that work, the question arose as to whether Litvin's result could be proved and if so was there a stronger result concerning the occurrences of type C (complex) representations. The application of a little character theory of the unitary groups suggests that complex representations (i.e., representation with complex characters) can be separated into two classes, "quasiorthogonal" or "quasisymplectic," where these terms are added into the statement of Litvin. King and I used[5] explicit Kronecker product formulas (in terms of S-functions) to show that this much stronger lemma holds for all compact semisimple Lie groups. However a finite group of order 16 fails the stronger condition. Frame[6] then produced a group, ${}^2F_4(2)$ of order $2^{12}3^35^213$, which contradicts Litvin's lemma.

The error in Litvin's proof is the assumption [after his Eq. (6)] that a certain unitary freedom exists. In fact his "arbitrary" unitary matrix has a symmetry imposed upon it by the reality of the characters, and not conversely.[7]

Note: D. Litvin agrees with the conclusions of this note.

---

[1] D. B. Litvin, J. Math. Phys. **13**, 1386 (1972).
[2] A. J. Mal'cev, Am. Math. Soc. Transl. Ser. 1 **9**, 172 (1962), see Lemma 1, p. 194.
[3] E. B. Dynkin, Am. Math. Soc. Transl. Ser. 2 **6**, 245 (1957), see Theorem 0.25, p. 361.
[4] P. H. Butler, "Wigner coefficients and n-j symbols for chains of groups", preprint, 1972.
[5] P. H. Butler and R. C. King, Can J. Math. (to be published).
[6] J. S. Frame, private communication, for details see Ref. 4.
[7] J-R. Derome and W. T. Sharp, J. Math. Phys. **6**, 1584 (1965), see Eq. (4.7).

Spin zero couplings in the Veneziano model*

Peter N. Dobson Jr.

University of Hawaii, Honolulu, Hawaii
(Received 20 September 1972)

The problem of the positivity of the residues in the simple Veneziano model is investigated. Previous studies of this problem have concentrated on the residues of poles lying along specific daughter trajectories, and hence needed to work only with polynomials of fixed low degree. A technique is developed here which is applied to all the spin zero daughters in the model, and which therefore deals with polynomials of arbitrarily high degree. It is found that the spin zero residues are all positive if $\alpha_0 \geq 1/2$, where α_0 is the intercept of the leading trajectory in the model.

I. INTRODUCTION

A number of investigators, including the present author in collaboration with W. A. Simmons, have looked into the problem of the positivity of the residues in the Veneziano model.[1-5] This model[6] in its simplest form represents a scattering amplitude in terms of ratios of gamma functions in such a way that "partial fraction" expansions of the amplitude may be interpreted in terms of s- and t-channel (zero width) resonances. For such an interpretation to be tenable, the residues of the poles which appear in the partial fraction expansion should be all positive; otherwise, the poles would have to be regarded as arising from negative norm intermediate states (ghosts). Finding a region of model parameters for which all the residues are positive has proven to be a very challenging mathematical problem. The relevant parameters of the model are α_0, the intercept of the leading trajectory, and $\alpha'\mu^2$ where α' is the slope of the trajectory, and μ is the particle mass. In Ref. 5, a region of these parameters is derived which ensures positivity of the residues for poles lying on the first six trajectories. The actual region is somewhat complicated in shape, but it includes the simpler region

$$\tfrac{1}{2} \leq \alpha_0 \leq 1 - 4\alpha'\mu^2.$$

The poles in the Veneziano model lie along straight-line trajectories in the J–m^2 plane, as shown in Fig. 1. J represents the angular momentum of the partial wave in which the pole appears, and m is the invariant center-of-mass energy. If we number the poles along a trajectory by the integer $K = 1, 2, \cdots$, then the parent trajectory has $J = K$, the first daughter trajectory has $J = K - 1$, and so on. As shown in Ref. 5, the residue at each pole

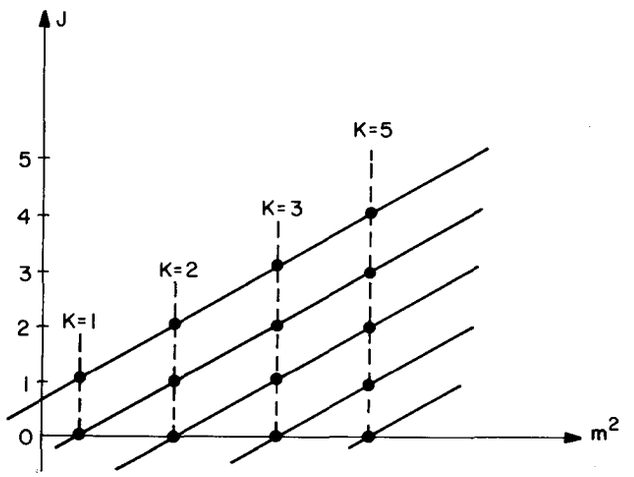


FIG. 1. Regge trajectories in the Veneziano model.

is expressed in terms of the model parameters by a polynomial which increases in degree as $K - J$ increases. In fact, except for easily handled common factors, the residues of the nth daughter trajectory involve a polynomial of degree n for n even and $n - 1$ for n odd. Thus, work done so far has treated all the residues only on the leading trajectories, where the polynomials involved are of reasonably low degree. In this paper, we will look at the residues along a line of constant J rather than constant $J - K$; as a result, we will have to come to grips with polynomials of arbitrarily high degree. The particular case dealt with here is $J = 0$, primarily because of simplifications in the forms of the polynomials that arise. However, studies of the high-K asymptotic limit by Nambu and Frampton[7] indicate that the $J = 0$ residues are among the most crucial ones in the sense of having the values closest to violating positivity for a given set of model parameters. Thus, the results presented here lend considerable additional support to the already widely accepted conjecture that there are some values of the parameters for which all the residues are positive.

In Sec. II, we will derive equations for the boundaries of the regions in which the spin zero residues are positive, and discuss some properties of the polynomials which define these boundaries. In Sec. III, we will see how the boundary curves are formed, and discuss some aspects of their shape and location. In Sec. IV, a bound is derived which limits the locations of the boundary curves sufficiently to give a region of positivity for all $J = 0$ residues. Numerical calculations for the finite number of cases not encompassed by the bound then show that the spin zero residues are positive for $\alpha_0 \geq \frac{1}{2}$.

II. EQUATIONS FOR THE BOUNDARY OF THE POSITIVITY REGION

Following the notation and definitions of Ref. 5, we introduce variables a_K and b_K defined by

$$a_K = a_1 + (K - 1)/2, \qquad b_K = b_1 - (K - 1)/2,$$

where a_1 and b_1 are related to the model parameters through the equations

$$a_1 = -\tfrac{1}{2}(\alpha_0 + 4\alpha'\mu^2 - 1),$$
$$b_1 = \tfrac{1}{2}(3\alpha_0 + 4\alpha'\mu^2 - 1).$$

We will also make frequent reference to the function

$$T_K(x) = x(x + 1)(x + 2)\cdots(x + K - 1).$$

As shown in Ref. 5, the residue of the Kth pole in the $J = 0$ partial wave is given by

$$C_0^K(a_K, b_K) = \sum_{m=0}^{[K/2]} \frac{a_K^{2m}}{2m + 1} B_{K-2m}^K(b_K), \tag{1}$$

where $[x]$ means the largest integer contained in x, and

$$B^K_{K-n}(x) = \frac{1}{n!}\frac{d^n}{dx^n}T_K(x). \tag{2}$$

To simplify the notation somewhat, we will henceforth drop the subscript K on the variables a_K and b_K. Then, combining Eqs. (1) and (2), we obtain

$$C^K_0(a,b) = \sum_{m=0}^{[K/2]} \frac{a^{2m}}{(2m+1)!}\frac{d^{2m}}{db^{2m}}T_K(b). \tag{3}$$

Let us now introduce the polynomial (of degree $K+1$)

$$I_K(x) = \int_0^x T_K(y)dy. \tag{4}$$

Consider a Taylor series expansion of $I_K(b+a)$ around the point $x = b$; since I_K is a polynomial of degree $K+1$, the Taylor series will be a finite sum:

$$I_K(b+a) = \sum_{n=0}^{K+1} a^n \frac{1}{n!}\frac{d^n}{db^n}I_K(b). \tag{5}$$

We can similarly expand $I_K(b-a)$; in fact, we can simply replace a by $-a$ in Eq. (5) to obtain:

$$I_K(b-a) = \sum_{n=0}^{K+1} (-1)^n a^n \frac{1}{n!}\frac{d^n}{db^n}I_K(b). \tag{6}$$

If we take the difference of Eqs. (5) and (6), all the even n terms in the two sums will cancel. The result will contain just the odd terms, and, letting $n = 2m+1$ for these, we have

$$I_K(b+a) - I_K(b-a) = \sum_{m=0}^{[K/2]} \frac{2}{(2m+1)!}a^{2m+1}\frac{d^{2m+1}}{db^{2m+1}}I_K(b). \tag{7}$$

But, from the definition of I_K given in Eq. (4), it is clear that

$$\frac{d^{2m+1}}{db^{2m+1}}I_K(b) = \frac{d^{2m}}{db^{2m}}T_K(b).$$

By taking account of this relation, a comparison of Eqs. (3) and (7) gives

$$C^K_0(a,b) = (1/2a)[I_K(b+a) - I_K(b-a)]. \tag{8}$$

The boundaries separating positive residues from negative residues are determined by $C^K_0(a,b) = 0$. From Eq. (8), these boundaries are given by

$$I_K(b+a) = I_K(b-a), \tag{9}$$

except possibly at $a = 0$. Letting a approach zero in Eq. (8), we see that

$$C^K_0(0,b) = T_K(b) = b(b+1)(b+2)\cdots(b+K-1).$$

Thus the boundaries of regions of positivity intersect the line $a = 0$ at the points $b = 0, -1, -2, \ldots,$ and $-K+1$. In terms of the variables a_1 and b_1, this means that the boundaries for a given K intersect the line $a_1 = -(K-1)/2$ at the points $b_1 = -(K-1)/2, -(K-2)/2, \ldots, (K-2)/2, (K-1)/2$. This result follows, of course, also from setting $a = 0$ in Eq. (1).

To make full use of Eq. (9), we will need to investigate in some detail the properties of the polynomials $I_K(x)$. We will first derive some symmetry properties. Consider

$$\begin{aligned} T_K(-x-K+1) &= (-x-K+1)(-x-K+2)\cdots(-x-1)(-x) \\ &= (-1)^K x(x+1)\cdots(x+K-1) = (-1)^K T_K(x). \end{aligned}$$

If we now investigate

$$I_K(-x-K+1) = \int_0^{-x-K+1} T_K(y)dy,$$

the change of variables $z = -y-K+1$ gives

$$\begin{aligned} I_K(-x-K+1) &= \int_x^{-K+1} T_K(-z-K+1)dz \\ &= (-1)^K \int_x^{-K+1} T_K(z)dz \\ &= (-1)^K[I_K(-K+1) - I_K(x)]. \end{aligned}$$

Thus, multiplying both sides by $(-1)^K$, we find

$$I_K(x) + (-1)^K I_K(-x-K+1) = I_K(-K+1). \tag{10}$$

Consider first the case where K is even, $K = 2k$. Then

$$I_{2k}(x) + I_{2k}(-x-2k+1) = I_{2k}(-2k+1),$$

and putting $x = -k+1/2$ gives

$$I_{2k}(-k+1/2) = 1/2 I_{2k}(-2k+1).$$

Our discussion of the properties of $I_K(x)$ for K even will be simplified if we work with a function shifted to make use of the symmetry inherent in the above relationships. So, let us define

$$A_k(x) = I_{2k}(x-k+1/2) - I_{2k}(-k+1/2). \tag{11}$$

Then the implication of Eq. (10) for the function $A_k(x)$ is simply that it is an odd function of x.

In the case that K is odd, $K = 2k+1$, Eq. (10) becomes

$$I_{2k+1}(x) - I_{2k+1}(-x-2k) = I_{2k+1}(-2k).$$

If we substitute $x = 0$ into this, we see that

$$I_{2k+1}(-2k) = 1/2 I_{2k+1}(0) = 0,$$

and so

$$I_{2k+1}(x) = I_{2k+1}(-x-2k).$$

This result may be summarized by noting that the function

$$B_k(x) = I_{2k+1}(x-k) - I_{2k+1}(-k) \tag{12}$$

is an even function of x.

The functions $A_k(x)$ and $B_k(x)$ should have maxima and minima at the zeroes of their derivatives. In particular, extreme points of $A_k(x)$ occur where

$$T_{2k}(x-k+1/2) = 0;$$

i.e., at

$$x = \pm 1/2, \pm 3/2, \ldots, \pm(2k-1)/2.$$

Similarly, $B_k(x)$ will have maxima and minima at

$$x = 0, \pm 1, \pm 2, \ldots, \pm k.$$

It will be important in our later discussion of the way the boundary curves are formed to have some estimates of the relative values of the functions $A_k(x)$ and $B_k(x)$ at

their maxima and minima. To make these estimates, we start with the following simple relation:

$$T_K(x-1) = (x-1)x(x+1)\cdots(x+K-2)$$
$$= [(x-1)/(x+K-1)]T_K(x).$$

We write this as

$$T_K(x) = -rT_K(x-1),$$

where

$$r = [K/(1-x)] - 1.$$

Note that for

$$1 - \tfrac{1}{2}K \leq x < 1,$$

we have $r \geq 1$. Thus, when we shift from a value x in the interval specified above to a value increased by 1, the absolute value of T_K increases (or at least does not decrease) while the sign changes. We can sharpen this statement as follows. Suppose we have

$$-n \leq x \leq -n+1,$$

where

$$0 \leq n \leq \tfrac{1}{2}K - 1.$$

Then we find

$$(K-n-1)/(n+1) \leq r \leq (K-n)/n.$$

Thus, if $T_K(x)$ is positive, we have the inequality

$$-[(K-n-1)/(n+1)]T_K(x-1) \leq T_K(x)$$
$$\leq -[(K-n)/n]T_K(x-1), \quad (13)$$

for x in the specified interval. If $T_K(x)$ is negative, the inequality is reversed.

Now we can apply this result to the function $A_K(x)$. From the definition of Eq. (11), we have

$$A_k(1/2) = \int_{-k+1/2}^{-k+1} T_{2k}(y)dy,$$

and

$$A_k(3/2) = \int_{-k+1/2}^{-k+2} T_{2k}(y)dy = A_k(1/2) + \int_{-k+1}^{-k+2} T_{2k}(y)dy.$$

Now T_{2k} may be either positive or negative for its argument between $-k$ and $-k+1$; for k odd it will be negative, and for k even, positive. For the sake of discussion, let us consider the case of k odd; as an example, see the graph of $A_3(x)$ given in Fig. 2. If T_{2k} is negative in the interval between $-k$ and $-k+1$, then clearly $A_k(1/2)$ will be negative. Furthermore, T_{2k} will be positive for its argument between $-k+1$ and $-k+2$, and so we may apply the inequality (13), taking n to be $k-1$, to the last integral in the expression for $A_k(3/2)$. We find

$$A_k(3/2) \geq A_k(1/2) - \int_{-k+1}^{-k+2} T_{2k}(y-1)dy$$
$$= A_k(1/2) - \int_{-k}^{-k+2} T_{2k}(z)dz,$$

where we have put $z = y-1$. Using the fact that $T_{2k}(x-k+1/2) = T_{2k}(-x-k+1/2)$, this last integral is

$$2\int_{-k+1/2}^{-k+1} T_{2k}(z)dz = 2A_k(1/2).$$

Thus

$$A_k(3/2) \geq -A_k(1/2). \quad (14)$$

We started from the assumption that k was odd, and hence that $A_k(1/2)$ was negative. We have shown that $A_k(3/2)$ is positive, and exceeds $|A_k(1/2)|$. We can also apply the other half of the inequality (13) to show that

$$A_k(3/2) \leq -[(k+1)/(k-1)]A_k(1/2). \quad (15)$$

The inequalities (14) and (15) are both just reversed if k is even and $A_k(1/2)$ is positive. We now wish to show that

$$-r_1(k,p)A_k((2p-1)/2) \leq A_k((2p+1)/2)$$
$$\leq -r_2(k,p)A_k((2p-1)/2) \quad (16)$$

is generally true [provided p is an integer such that the arguments of A_k in Eq. (16) are at maxima or minima of the function] and that

$$r_1(k,p) \geq 1, \quad r_2(k,p) \leq (k+3p)/(k-p), \quad \text{if } A_k((2p+1)/2) \geq 0$$
$$r_2(k,p) \geq 1, \quad r_1(k,p) \leq (k+3p)/(k-p), \quad \text{if } A_k((2p+1)/2) \leq 0. \quad (17)$$

We will prove this by induction, first noting that Eqs. (14) and (15) are compatible with Eqs. (16) and (17), with $p = 1$. Now we assume that Eqs. (16) and (17) are true for a given value of p, and will show that they hold for p replaced by $p+1$. We have that

$$A_k\left(\frac{2p+3}{2}\right) = \int_{-k+1/2}^{-k+p+2} T_{2k}(y)dy$$
$$= A_k\left(\frac{2p+1}{2}\right) + \int_{-k+p+1}^{-k+p+2} T_{2k}(y)dy.$$

We will assume that $A_k((2p+1)/2)$ is positive. Applying Eq. (13), with the inequality reversed, we obtain

$$A_k\left(\frac{2p+3}{2}\right) \leq A_k\left(\frac{2p+1}{2}\right) - \left(\frac{k+p}{k-p}\right)\int_{-k+p+1}^{-k+p+2} T_{2k}(y-1)dy$$
$$= -\left(\frac{2p}{k-p}\right)A_k\left(\frac{2p+1}{2}\right) + \left(\frac{k+p}{k-p}\right)A_k\left(\frac{2p-1}{2}\right). \quad (18)$$

Now note that Eq. (16) implies

$$A_k\left(\frac{2p-1}{2}\right) \leq -A_k\left(\frac{2p+1}{2}\right)/r_2(k,p).$$

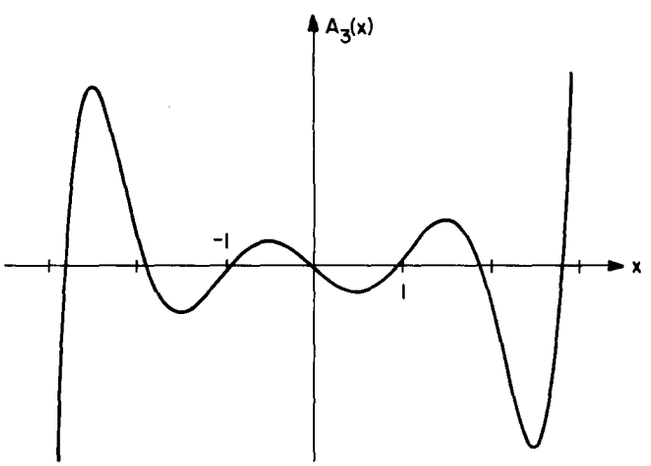


FIG. 2. The function $A_3(x)$.

Inserting this in Eq. (18), we see that

$$A_k\left(\frac{2p+3}{2}\right) \leq -\left[\left(\frac{2p}{k-p}\right) + \left(\frac{k+p}{k-p}\right)\frac{1}{r_2(k,p)}\right]A_k\left(\frac{2p+1}{2}\right). \quad (19)$$

On comparison with Eq. (16), we can define

$$r_2(k,p+1) = \left(\frac{2p}{k-p}\right) + \left(\frac{k+p}{k-p}\right)\frac{1}{r_2(k,p)}.$$

Now, by hypothesis

$$r_2(k,p) \leq (k+3p)/(k-p),$$

so

$$r_2(k,p+1) \geq \left(\frac{2p}{k-p}\right) + \left(\frac{k+p}{k+3p}\right) > \left(\frac{2p}{k+3p}\right) + \left(\frac{k+p}{k+3p}\right) = 1,$$

verifying the first inequality in Eq. (17) for p replaced by $p+1$, and choosing the case for A_k negative, as follows from Eq. (19).

If we apply the other inequality in Eq. (13), we also have

$$A_k\left(\frac{2p+3}{2}\right) \geq A_k\left(\frac{2p+1}{2}\right) - \left(\frac{k-p-1}{k+p+1}\right)\int_{-k+p+1}^{-k+p+2} T_{2k}(y-1)dy$$

$$= -2\left(\frac{p+1}{k-p-1}\right)A_k\left(\frac{2p+1}{2}\right) + \left(\frac{k+p+1}{k-p-1}\right)A_k\left(\frac{2p-1}{2}\right). \quad (20)$$

From Eq. (16),

$$A_k\left(\frac{2p-1}{2}\right) \geq -A_k\left(\frac{2p+1}{2}\right)/r_1(k,p),$$

which, inserted into Eq. (20), leads us to define

$$r_1(k,p+1) = 2\left(\frac{p+1}{k-p-1}\right) + \left(\frac{k+p+1}{k-p-1}\right)\frac{1}{r_1(k,p)}.$$

Using the assumed inequality,

$$r_1(k,p) \geq 1,$$

we see

$$r_1(k,p+1) \leq \frac{k+3(p+1)}{k-(p+1)},$$

verifying the second inequality in Eq. (17). Had we assumed $A_k((2p+1)/2)$ negative, the obvious modifications of the above argument verify the inequalities in Eq. (17) for the case of $A_k((2p+3)/2)$ positive.

The implications of these inequalities for the shape of the curves $A_k(x)$ is now fairly simple to state. As x increases from 0, alternate maxima and minima are encountered at half-odd-integer values of x up to $\frac{1}{2}(2k-1)$. $A_k(x)$ is positive at the maxima and negative at the minima, and the relative sizes of these are governed by

$$|A_k((2p+1)/2)| > |A_k((2p-1)/2)|, \quad (21)$$

which follows from Eqs. (16) and (17). Thus, as we go farther from $x=0$ we find the peaks and valleys become higher and deeper. Because of the fact that $A_k(x)$ is an odd function of x, the same thing holds true if we move away from $x=0$ in the negative direction. This property will prove to be crucial for the determination of the shapes of the boundary curves worked out in Sec. III.

The arguments we have developed here for the functions $A_k(x)$, relevant when K is even, may be applied with minor modifications to the functions $B_k(x)$. As defined, $B_k(0)=0$ is a maximum or minimum itself. As we move away from $x=0$ in either the positive or negative direction, we encounter successive positive maxima and negative minima at integer values of x from $-k$ to k. As is the case with the functions $A_k(x)$, the peaks and valleys become higher and deeper as we get farther from $x=0$.

III. PROPERTIES OF THE BOUNDARY CURVES

According to the discussion of the last section, the curves separating regions of positive residue from regions of negative residue are determined by Eq. (9). In terms of the variables a_1 and b_1, this can be rewritten

$$I_K(b_1+a_1) = I_K(b_1-a_1-K+1).$$

Let us consider the case of even $K=2k$. Then we can express the above as

$$A_k(b_1+a_1+k-1/2) = A_k(b_1-a_1-k+1/2).$$

Let us define $x = b_1+a_1+k-1/2$; then $b_1-a_1-k+1/2 = x-c$, where $c = 2a_1+2k-1$. The equation is now simply

$$A_k(x) = A_k(x-c). \quad (22)$$

If we plot $A_k(x)$ and $A_k(x-c)$ on the same graph as functions of x, then the intersections of these two curves will determine the solutions of Eq. (22). Thus, a fixed value of a_1 fixes the amount c by which the second curve is shifted, and the values of x at the intersections of the two curves determine values of b_1 which lie on the boundary curves. These values correspond to intersections of the line $a_1 = \text{const.}$ with the set of boundary curves.

When $c=0$, Eq. (22) is trivially satisfied for all x. But $c=0$ implies $a_1 = -k+1/2$, or $a=0$, and so this simply reflects the fact that we have found it convenient to deal with the product of a and the residue functions $C_0^K(a,b)$ instead of the latter directly. If we consider c small, it is clear that we will get intersections near the maxima and minima of $A_k(x)$; see, as an illustration, Fig. 3. For, let x_0 be a point at which $A_k(x)$ is a maximum or minimum; then for $x-x_0$ small enough, we can write

$$A_k(x) = A_k(x_0) + \tfrac{1}{2}A''(x_0)(x-x_0)^2,$$

ignoring terms of higher order in the small quantity.

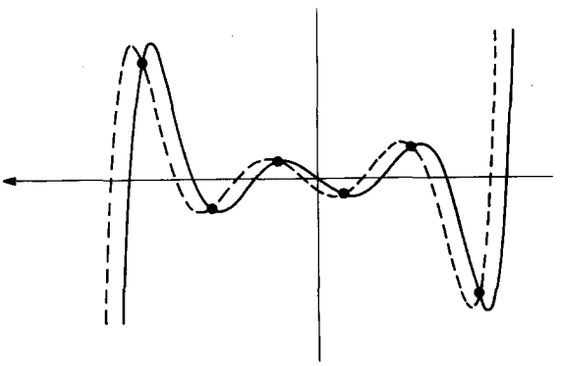


FIG. 3. Intersections of $y = A_3(x)$ (solid curve) and $y = A_3(x-c)$ (dashed curve) for small c.

But then $x = x_0 + \frac{1}{2}c$ will give a solution of Eq. (22) to the same order. As c approaches zero, the boundary curves approach the points $b_1 = \pm\frac{1}{2}, \ldots, \pm\frac{1}{2}(2k-1)$ on the line $a_1 = -k + 1/2$. These intersections were already discussed in Sec. II.

Now we must investigate the behavior of the solutions of Eq. (22) for arbitrary c. We will start by proving that all intersections between $y = A_k(x)$ and the shifted curve $y = A_k(x-c)$ originate in the fashion discussed in the last paragraph; i.e., any x that solves Eq. (22) will move smoothly to one of the extreme points of A_k as c approaches zero. First of all, note that if x is a solution of Eq. (22), then so is $c - x$. For,

$$A_k(c-x) = -A_k(x-c) = -A_k(x) = A_k(-x)$$
$$= A_k[(c-x)-c].$$

If we assume that c is positive, then either x or $c - x$ (or both) will be positive. Let us choose the largest of these two solutions; assume it lies between the pth and $(p+1)$th extreme point to the right of the origin, and call it x_p. The other solution we will call $x_{-p} = c - x_p$. So our assumed position for x_p is

$$\tfrac{1}{2}(2p-1) < x_p < \tfrac{1}{2}(2p+1).$$

Also, since $x_{-p} < x_p$, we must have $c < 2x_p$. The shift c given by this limiting value will move the dashed curve in Fig. 3 so that the point on it corresponding to $-x_p$ will have the same horizontal position as the point at x_p on the unshifted curve. The point at $-x_p$ has only the extreme points beginning with $-\frac{1}{2}(2p-1)$ to its right. Since, in the general case, $c < 2x_p$, we may have shifted fewer extreme points than this to the right of the point x_p, but we cannot have shifted more. Now consider what happens as we let c decrease so that the shifted curve moves to the left. The point at x_p will slide up and down on the slope between $\frac{1}{2}(2p-1)$ and $\frac{1}{2}(2p+1)$ as the peaks and valleys on the shifted curve move past, but it will remain "trapped" on this slope until the approach of the extreme point at $\frac{1}{2}(2p-1)$ on the shifted curve as c approaches zero. This follows because none of the intervening peaks or valleys are high enough or deep enough to carry x_p up or down to the extreme points on either side of it. The only extreme point for which A_k has even a large enough magnitude is that at $-\frac{1}{2}(2p-1)$, and here A_k has the wrong sign. As c approaches zero, we come into a configuration like that in Fig. 3, and eventually x_p goes to $\frac{1}{2}(2p-1)$. It is not hard to see that in this same limit x_{-p} becomes $-\frac{1}{2}(2p-1)$.

Next we prove that for sufficiently large c the points x_p and x_{-p} coincide. Reversing some of the arguments above, we can see that as c is increased from zero, the intersection at x_p will move onto the slope between the extreme points at $\frac{1}{2}(2p-1)$ and $\frac{1}{2}(2p+1)$ and will be "trapped" there as the smaller peaks and valleys on the shifted curve move past. Eventually c will become large enough so that the intersection at x_p on the unshifted curve will be with the extreme point on the shifted curve at $-\frac{1}{2}(2p-1)$. That is, $x_p - c = -\frac{1}{2}(2p-1)$. But then $x_{-p} = \frac{1}{2}(2p-1)$, so the intersection at x_{-p} will have moved to the extreme point just to the left of the slope where x_p has been "trapped". The situation will now look like one of the cases in Fig. 4, depending on whether the pth extreme point was a maximum or minimum. Clearly, in either case a small additional increase in c, moving the dashed curve to the right, will cause the points x_p and x_{-p} to coincide. Since $A_k(x_p) = -A_k(x_{-p})$, the points can only coincide at a zero of A_k. Let us call the root of $A_k(z)$, which lies between the pth and $(p+1)$th extreme point, z_p. Then the points x_p and x_{-p} both become z_p when $c = 2z_p$. Let us show now that two intersections labeled by x_p and x_q cannot coincide unless $q = -p$. Consider first p and m, both positive, with $p \neq m$, and $q = -m$. If for some value of c we have $x_p = x_{-m}$, then

$$x_{-p} = c - x_p = c - x_{-m} = x_m.$$

Suppose $x_m > x_p$. Then $x_{-p} > x_p$ as well; but this cannot be true, since x_{-p} lies to the left of x_p until the two coincide, and after coincidence neither intersection exists. If we have $x_p > x_m$, we get the same sort of contradiction. If we take p and q, both positive, with $p \neq q$, it is clear that we cannot have $x_p = x_q$ for any value of c, since these points are trapped on slopes between different pairs of extreme points. But neither can we have $x_{-p} = x_{-q}$, since this condition implies $x_p = x_q$.

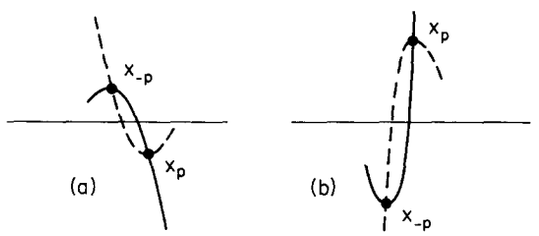


FIG. 4. Positions of intersections at x_p and x_{-p} just before they coincide. (a) x_p originates at a maximum, (b) x_p originates at a minimum.

We are now in a position to describe the system of curves determined by Eq. (22). Recall that

$$a_1 = \tfrac{1}{2}c - k + \tfrac{1}{2}$$
$$b_1 = x - \tfrac{1}{2}c$$

The values of x determined by Eq. (22) for a fixed value of c determine the values of b_1 at which a line of constant a_1 intersects the system of curves in the a_1–b_1 plane. When $c = 0$, the line is $a_1 = -k + 1/2$, and the intersections are at $b_1 = \frac{1}{2}, \ldots, \pm\frac{1}{2}(2k-1)$. As we move the line $a_1 = \text{const.}$ to the right of this, we increase c, and eventually, when $a_1 = z_1 - k + 1/2$, the curves that are connected with the points $b_1 = \pm 1/2$ on the line $a_1 = -k + 1/2$ join together. This occurs when $c = 2z_1$, $x = z_1$, and so $b_1 = 0$. The general pattern is clear: As the line moves through the positions at $a_1 = z_p - k + 1/2$, we find a joining of the curves which started from $b_1 = \pm\frac{1}{2}(2p-1)$ at $b_1 = 0$. Finally, when $a_1 = z_k - k + 1/2$, where z_k is the largest root of $A_k(z) = 0$, the last pair of curves joins, and this line is tangent to the outermost curve of the system. Using the fact that the system of curves is symmetric about the line $a_1 = -k + 1/2$, we find that they form a set of k nested closed curves, intersecting the line $b_1 = 0$ at $a_1 + k - 1/2 = \pm z_1, \ldots, \pm z_k$, and lying completely in the region $a_1 \leq z_k - k + 1/2$. An example of the system of boundary curves, for $K = 6$, is shown in Fig. 5.

We have treated the case of even K at some length, and much the same methods can be applied for odd $K = 2k + 1$. The boundary curves again form a system of k nested closed curves intersecting the symmetry line $a_1 = -k$ at $b_1 = \pm 1, \ldots, \pm k$. For this case, however, we can determine explicitly the intersections with the line $b_1 = 0$ as well. From Ref. 5, we see that $C_0^K(a,b)$ is proportional to b_1 when K is odd. Hence it is appropriate to replace Eq. (8) by

$$(1/b_1)C_0^K(a,b) = (1/2ab_1)$$
$$\times [I_K(b_1 + a_1) - I_K(b_1 - a_1 - K + 1)]. \quad (23)$$

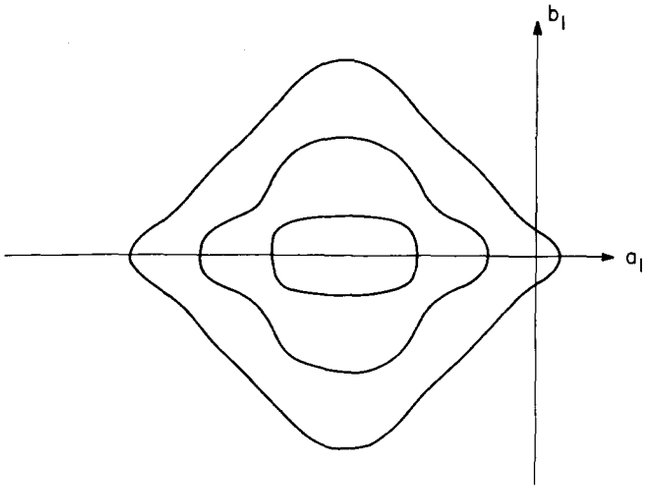


FIG. 5. The boundary curves for regions of positivity when $K = 6$.

Letting b_1 approach zero, the right-hand side of Eq. (23) becomes

$$(1/2a)[T_K(a_1) - T_K(-a_1 - K + 1)] = (1/a)T_K(a_1).$$

Thus the boundary curves intersect the line $b_1 = 0$ at $a_1 = 0, -1, \ldots, -K + 1$. The system of curves lies entirely in the region $a_1 \leq 0$, and since positive a_1 is already required to give positive residues on the leading trajectory in the model, the spin zero residues with K odd provide no new restriction.

IV. REGION OF POSITIVITY FOR ALL $J = 0$ COUPLINGS

The results of the last section imply that the residue of the Kth pole in the $J = 0$ partial wave is positive if a_1 exceeds some value. For K odd, this value is zero, and for even $K = 2k$, it is $a_k^{\min} = z_k - k + 1/2$. Since z_k lies to the right of the minimum of A_k which occurs at $k - 1/2$, we see that $a_k^{\min} > 0$. If we could determine a bound for z_k, such as

$$z_k \leq Z, \quad \text{all } k,$$

we would have a region of positivity valid for all K. It would not be a very useful region, however, because of the way a_1 is related to the model parameters α_0 and $\alpha'\mu^2$. A more useful result can be obtained by reconsidering (only briefly!) the formation of the boundary curves determined by Eq. (22). Let us write the latter in the form

$$A_k(x) = A_k(y). \tag{24}$$

The function $A_k(x)$ is large and negative for x sufficiently small; it passes through zero at $x = -z_k < -k + 1/2$, and so the extreme point at $x = -k + 1/2$ must be a maximum, and from the results of Sec. II it is the largest maximum of A_k. Now consider solving Eq. (24) as follows. Choose a large value of x. Construct a vertical line from this value on the x axis until it intersects the curve $y = A_k(x)$, and then draw a horizontal line from this intersection. If x is large enough that $A_k(x) > A_k(-k + 1/2)$, this horizontal line will have no other intersection with the curve $y = A_k(x)$. Thus for this value of x, the only solution of Eq. (24) is the trivial one $y = x$. As we reduce x, the horizontal line constructed as above will move down, until finally at some $x = x_0$, it will touch the peak at $-k + 1/2$. This means the line $x = x_0$ is tangent to the outermost boundary curve determined by Eq. (24) at $y = -k + 1/2$. The usefulness of this result becomes evident when it is noted that lines of constant x correspond to the lines $a_1 + b_1 = \text{const.}$ in the a_1–b_1 plane, and that $a_1 + b_1 = \alpha_0$. Also, $y = -k + 1/2$ implies $a_1 = b_1$. So we find that the boundary curves all lie in the region $\alpha_0 \leq 2a_1$, where $a_1 = b_1$ is the point on the outermost boundary where the tangent line discussed above intersects it. Furthermore, since any a_1 on the system of boundary curves must be less than $z_k - k + 1/2$, we have

$$\alpha_0 \geq 2(z_k - k + 1/2) \tag{25}$$

as a condition for positivity of the residue at the $2k$th pole.

We now want to prove that $\alpha_0 \geq 1/2$ is sufficient to guarantee positivity for all K. We proceed first to show that

$$z_k - k + 1/2 \leq Z_k,$$

where Z_k goes to zero as k goes to infinity. Then we can determine a value n such that $Z_k \leq 1/4$ for all $k > n$. We will then explicitly calculate the residue at the point $a_1 = b_1 = 1/4$ for the cases $k = 1, 2, \ldots, n$ and show that it is not negative.

The polynomial $A_k(x)$ has zeroes at $x = 0, \pm z_1, \ldots, \pm z_k$. For all of these except $\pm z_k$, we have some limits on their positions, since they lie between extreme points of $A_k(x)$. In particular,

$$\tfrac{1}{2}(2p - 1) \leq z_p \leq \tfrac{1}{2}(2p + 1), \quad p = 1, 2, \ldots, k.$$

Since we are interested in $z_k - k + 1/2$, it will be somewhat more convenient to work with the function

$$F(x) = A_k(x + k + 1/2).$$

Let us call its zeroes $x_0, -x_1, -x_2, \ldots, -x_{2k}$, where $x_0 = z_k - k + 1/2$, $x_1 = -z_{k-1} + k - 1/2, \ldots, x_{2k} = z_k + k - 1/2$.

We have the bounds

$$p - 1 \leq x_p \leq p, \quad \text{for } p = 1, 2, \ldots, 2k - 1. \tag{26}$$

Furthermore, all the x_p as defined are positive.

Since $F(x)$ is a polynomial, we can write it in the form

$$F(x) = A(x - x_0)(x + x_1) \cdots (x + x_{2k}).$$

The natural logarithm of this is

$$\ln F(x) = \ln A + \ln(x - x_0) + \sum_{p=1}^{2k} \ln(x + x_p).$$

Taking the derivative of this expression,

$$\frac{F'(x)}{F(x)} = (x - x_0)^{-1} + \sum_{p=1}^{2k} (x + x_p)^{-1}. \tag{27}$$

But $F'(x) = T_{2k}(x)$, and so $F'(0) = 0$. Putting $x = 0$ in Eq. (27), we have

$$\frac{1}{x_0} = \sum_{p=1}^{2k} \frac{1}{x_p} \geq \sum_{p=1}^{2k-1} \frac{1}{x_p}. \tag{28}$$

But the bounds in Eq. (26) imply $1/x_p \geq 1/p$, so we have the further inequality

$$\frac{1}{x_0} \geq \sum_{p=1}^{2k-1} \frac{1}{p}.$$

Thus we have the desired bound

$$x_0 = z_k - k + 1/2 \leq Z_k = \left(\sum_{p=1}^{2k-1} \frac{1}{p}\right)^{-1}.$$

We would like to know how large a value of k is required before $Z_k < 1/4$. We can find this rather easily from the bound

$$\sum_{p=m}^{2k-1} \frac{1}{p} > \int_m^{2k} \frac{dx}{x} = \ln\left(\frac{2k}{m}\right).$$

For example,

$$\sum_{p=1}^{2k-1} \frac{1}{p} > \left(\frac{25}{12}\right) + \ln\left(\frac{k}{2}\right),$$

and the right-hand side is greater than 4 for $k = 14$ or greater.

Thus, if we can show that C_0^K is positive at $a_1 = b_1 = 1/4$ for $K = 2, 4, \ldots, 28$, we will have established the positivity of all spin zero residues for

$$1/2 \leq \alpha_0 \leq 1 - 4\alpha'\mu^2.$$

Physically, this point corresponds to $\alpha_0 = 1/2$, and $\alpha'\mu^2 = 0$, and it is well known that residue for $K = 2$ vanishes there. Hence the lower limit cannot be improved, except by some function of $\alpha'\mu^2$. Wagner[2] reports computer calculations that indicate positivity of all residues at the point in question for K up to 50. Our own calculation of the relevant spin zero residues is given in Table I. As expected, all are positive except the $K = 2$ residue which vanishes.

TABLE I. The residues for even K from 2 to 28. Value given is C_0^K evaluated at $\alpha_0 = 1/2$, $\alpha'\mu^2 = 0$, divided by $\Gamma(K)$.

K	Residue
2	0
4	0.010 47
6	0.011 22
8	0.010 82
10	0.010 24
12	0.009 67
14	0.009 14
16	0.008 68
18	0.008 26
20	0.007 89
22	0.007 56
24	0.007 26
26	0.006 99
28	0.006 74

V. DISCUSSION AND CONCLUSIONS

The final result we have obtained is simple enough that it hardly warrants any extended discussion, but the argument used to obtain it was sufficiently involved that it might be worthwhile to summarize it here. The discussion began with a result set down in Ref. 5; namely, that any given residue is positive in the first quadrant of the a_1–b_1 plane outside a region determined by the vanishing of a polynomial in a_1 and b_1. In the case $J = 0$, we found that this polynomial in two variables was proportional to the difference of the values of a single polynomial in one variable evaluated at two different points. We investigated the properties of this latter polynomial and then showed that those properties implied a rather simple nature for the boundary curves; namely, that they consist of a system of k nested closed curves, where $k = [K/2]$, symmetric about the lines $b_1 = 0$ and $a_1 = -(K-1)/2$. For odd K, the outermost of these curves does not enter the first quadrant of the a_1–b_1 plane; for even K, it does so but is limited to a region determined by a vertical tangent line through its intersection with the horizontal (a_1) axis, and a tangent line of the form $a_1 + b_1 = \alpha_0 = \text{const.}$, which touches the curve where $a_1 = b_1$. We were able to show that for K greater than 28, the region so determined lies entirely outside the region $\alpha_0 \geq 1/2$, and treated the remaining cases by direct numerical evaluation of the residues.

It would be desirable to extend the sort of argument used in this study to the case of general J, but there are special simplifications for $J = 0$ which make it rather easier to treat. In the following Appendix, some formulas are given for the residue functions C_J^K which indicate a possible direction of approach to the general case. They unfortunately also indicate that the problem is not just a simple extension of the result obtained here.

APPENDIX

From Ref. 5, the formula for the residue function for arbitrary K and J is

$$C_J^K(a, b) = \sum_{m=1}^{[K/2]} \gamma_m^J a^{J+2m} B_{K-J-2m}^K(b), \tag{A1}$$

where the coefficients

$$\gamma_m^J = (J + 2m)!/2^m m!(2J + 2m + 1)!!.$$

From Eq. (A1), making use of the relation in Eq. (2), we can show that

$$\frac{\partial}{\partial a}(a^{J+2} C_{J+1}^K) = a^{2J+2} \frac{\partial}{\partial b}(a^{-J} C_J^K) \tag{A2}$$

and

$$\frac{\partial}{\partial b}(a^{J+2} C_{J+1}^K) = a^{2J+2} \frac{\partial}{\partial a}(a^{-J} C_J^K). \tag{A3}$$

A solution of these equations can be found in terms of the function $T_K(x)$ and repeated integrals of this function. If we define the polynomials T_N^K so that

$$\frac{d}{dx} T_N^K(x) = T_{N-1}^K(x),$$

with $T_0^K(x) = T_K(x) = x(x+1)\ldots(x+K-1)$, and such that the lowest power of x in T_N^K is x^{N+1}, then

$$C_J^K(a, b) = \sum_{n=0}^{J} \frac{(-1)^n (J+n)!}{n!(J-n)!} (2a)^{-n-1} \times [T_{n+1}^K(b + a) - (-1)^{J-n} T_{n+1}^K(b - a)]. \tag{A4}$$

This may be verified by direct substitution of Eq. (A4) into Eqs. (A2) and (A3).

*Work supported in part by the U. S. Atomic Energy Commission under Contract AT(04-3)-511.

[1] J. Yellin, UCRL Preprints Nos. UCRL 18637 and UCRL 18665 (unpublished).
[2] F. Wagner, Nuovo Cimento A **63**, 393 (1969).
[3] R. Oehme, Nuovo Cimento Lett. **1**, 420 (1969).
[4] J. Shapiro, Phys. Rev. **179**, 1345 (1969).
[5] P. N. Dobson and W. A. Simmons, J. Math. Phys. **12**, 778 (1971).
[6] G. Veneziano, Nuovo Cimento A **57**, 190 (1968).
[7] Y. Nambu and P. Frampton, "Asymptotic Behavior of Partial Widths in the Veneziano Model," in *Quanta*, edited by P. G. O. Freund, C. J. Goebel, and Y. Nambu (University of Chicago Press, Chicago, 1970).

Erratum: On the decomposition $SO(p,1) \supset SO(p-1,1)$ for most degenerate representations [J. Math Phys. 12, 2070(1971)]

Charles P. Boyer and Farhad Ardalan

Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802
(Received 25 September 1972)

1. The last two equations in Eq. (3.1) read η^{p-2} and η^{p-1} instead of η^{p-1} and η^{p} respectively.
2. In the first equation in Eq. (3.8) read

   $$\Gamma\left(n_j + \frac{j}{2}\right)2^{n_j+(j-1)/2} \quad \text{instead of} \quad \Gamma\left(n_j + \frac{j}{2}\, 2^{n_j+(j-1)/2}\right).$$

   and in the second equation of (3.8) the bracket [should be in front of $|\Gamma$ instead of $N_{\nu N}$.
3. The second equation in (3.9) should read $\bar{f}_{n,N}(\Omega'_p)$ instead of $f^{*}_{n,N}(\Omega'_p)$.
4. In Eq. (3.11) insert $N_\nu N_n$ before integral sign.
5. In Eqs. (4.4), (4.5), and (4.6) read

   $$\Gamma\left(n + \frac{p-2}{2} - j\right) \quad \text{instead of} \quad \Gamma(n + 1 - j).$$

6. In Eq. (4.7) read

   $$\Gamma\left(n + \frac{p-2}{2}\right) \quad \text{instead of} \quad \Gamma(n + 1)$$

   and the last Γ function in the second line should have $+i\nu$ in its arguments instead of $-i\nu$. In the ${}_4F_3$ function the first entry in the second row should be $-n-(p-4)/2$ instead of $-n$.
7. The superscript on the sum $\sum$ in Eq. (4.9) should be $n-1$ instead of $n=1$.
8. In line 11 p. 2075 read

   $|\nu|^{l+(p-2)/2}$ instead of $|\nu|^{l+(p-3)/2}$.

   In line 13 p. 2075 read

   $|\nu|^{n+(p-3)/2}$ instead of $|\nu|^{n+p-3}$.